Dieter Maurer

Acoustics of the Vowel

Indices

Peter Lang

Acoustics of the Vowel
Indices

In a first treatise on vowel acoustics, entitled Preliminaries, intellectual, methodological and empirical reasoning was exposed that gives rise to scepticism about the prevailing acoustic theory of the vowel, the formant theory.

In this second treatise, pursuing the quest for the acoustic representation of vowel quality, the variability of the vowel spectrum and its dependence on fundamental frequency is revisited on the new empirical basis of the Zurich Corpus of Vowel and Voice Quality, with the aim of formulating knowledge-based statements concerning the spectral representation of vowel quality in general and of questioning the cause of the observable relation of the vowel spectrum to fundamental frequency. As a central result, three statements are presented that serve as primary indices for a future acoustic theory: The vowel sound is a kind of perceptual and acoustic foreground–background phenomenon, spectral representation of vowel quality is nonuniform and, most importantly, the recognition and spectral representation of the vowel does not relate to fundamental frequency but to pitch (or to a comparable perceptual reference).

The treatise concludes with a reflection on the prerequisites and challenges of building a future acoustic theory of the vowel.

Dieter Maurer is a senior researcher and professor emeritus at the Zurich University of the Arts. His main scientific interest is focused on the question of the fundamental syntactic form character of vocal and graphic expressions. He is the author of numerous articles and books on this matter and is the leading author of the associated, extensive sound and picture archives, which are accessible online.

Acoustics of the Vowel

Dieter Maurer

Acoustics of the Vowel

Indices

PETER LANG
Lausanne · Berlin · Bruxelles · Chennai · New York · Oxford

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at http://dnb.dnb.de.

Library of Congress Cataloging-in-Publication Data
A CIP catalog record for this book has been applied for at the Library of Congress

Published with the support of the Swiss National Science Foundation (OA publication, grant 10BP12_224905/1), the Ernst Göhner Stiftung, Switzerland, and the Media and Information Centre of the Zurich University of the Arts.

Published as volume 30 of the series subTexte, edited by Anton Rey, Institute for the Performing Arts and Film, Zurich University of the Arts. https://subtexte.ch

Layout and cover design: Jacques Borel, Zurich

ISBN 978-3-0343-4668-9 pb.
ISBN 978-3-0343-5039-6 (ePDF)
DOI 10.3726/b22003

2024 Peter Lang Group AG, Lausanne
Published by Peter Lang Group AG, Lausanne, Switzerland
info@peterlang.com — www.peterlang.com

This publication has been peer reviewed.

To my parents

For Danju

Acknowledgements

Authorship

This treatise is written by the undersigned author, who takes full responsibility for the reasoning and presentation. However, the research carried out, the results obtained and the conclusions drawn would not have been achieved without the intellectual involvement of and discourse with Christian d'Heureuse, his general technical approach and the actual implementation of software tools for the sound database that served as the basis of our phenomenological and empirical studies – the Zurich Corpus of Vowel and Voice Quality (in short the Zurich Corpus) –, including tools for sound analysis, sound (re-)synthesis, sound manipulation and sound filtering, and his critical comments concerning empirical designs and the interpretation of results. Thus, his involvement was of crucial importance and must be taken into account accordingly. In particular, his contribution was essential for the most important assertion made here, namely that the perception and acoustics of the vowel are related to the pitch of a sound (or to a comparable perceptual referencing to a sound pattern repetition over time), and he therefore shares authorship of the vowel-pitch relation thesis with the undersigned author.

Creation of the Zurich Corpus of Vowel and Voice Quality

The creation of the comprehensive Zurich Corpus has taken years of intense work and effort, with countless hours of conceptual discussions, software development, testing, organisation of speakers, sound recording and editing, and administration. On the one hand, the creation of the corpus was made possible by Christian d'Heureuse, who designed, realised and maintained the database and its user interface; by Heidy Suter, who was in charge of finding speakers and of organising and conducting the recordings over a long period of time; by Daniel Friedrichs and Thayabaran Kathiresan, who were involved in the sound editing procedures; and by Volker Dellwo, with whom we continuously discussed conceptual aspects of the corpus. On the other hand, the creation of the sound corpus would not have been possible without the many children, women and men – untrained speakers and professional singers, actresses and actors – who took part in our studies as speakers and lent us their voices to aid our understanding of what we were scrutinising. The contribution of all those mentioned was essential and must in its turn be taken into account when appraising the corpus.

Text creation

The creation of the present text was supervised by Heidy Suter with attentive, persistent, detailed, and critical editing. She has never lowered her standards, providing a highly readable text with a fluent style, despite the scope and the long writing process (we worked on this text for four years), despite the many aspects that are new and difficult to work out in the academic discipline addressed, and despite the author's often "German" style of argumentation and writing. In addition, Hannah Jones has further improved the text through a thorough proofreading. When it comes to the appreciation of the text, her contributions were indispensable and must also be considered accordingly.

Affiliation and institutional support

The creation of the Zurich Corpus, the research carried out and the writing of this treatise were part of a broad field of research that was and is being conducted at the Institute of the Performing Arts and Film of the Zurich University of the Arts. Anton Rey, the head of the institute, supported all our works without hesitation and was undeterred by the risk of our studies leaving the prevailing theoretical ground, the associated scientific controversy that arose, and the reduced number of publications during the studies due to the workload of our phenomenological approach, the building up of the comprehensive Zurich Corpus, and the elaboration of many new experimental designs. Furthermore, as editor, he has integrated this second treatise on the acoustics of the vowel into his institute's publication series subTexte, as was the case with the first. In doing so, he provided us with a scientific environment, institutional embedding and long-term support that not many researchers achieve. The new evidence on vowel acoustics and also on vowel perception presented here are thanks to his rare attitude of institutionally integrating and supporting new and long-term research work.

Main cooperation partner, main partner for the empirical studies

Most of the work presented here relates to two research projects funded by the Swiss National Science Foundation, which we have carried out in cooperation with Volker Dellwo, Professor of Phonetics and Phonology at the Zurich Center for Linguistics of the University of Zurich. Many parts of the treatise reflect the numerous long discussions we shared, and he brought his broad phonetic knowledge into the context of our work, which was of great value in embedding our studies and their results in the prevailing scientific context. Furthermore, and most importantly, he was involved in a part of the reported empirical studies, together with Daniel Friedrichs and Thayabaran Kathiresan, who were doctoral students at his institute at the time. Some of these studies had previously been published as articles in specialist journals or as papers or abstracts in conference proceedings. Thus, the work presented here is in part the result of a long cooperation, mutual attention and in-depth discussions.

To all of you, too

My thanks also go to:

As said, to all the children, women, men, both the untrained speakers and the professionally active actresses, actors and singers, who participated in our studies and who lent us their voices, aiding our understanding of what we are questioning. I have them to keep anonymous.

Martina Bovet, Angelo Canonico, Lukas Hobi, Rahel Hadorn, Amaya Keller, Oliver Mannel, Susanne Peterson, Gianmarco Rostetter and Yael de Vries, who took part in our studies as expert listeners and taught us the manifold aspects of vowel and pitch recognition.

Franziska Feucht, who took part in the evaluation of the recording setting.

The teams of the Information Technology Centre, the Technical and Events Services, the Research Affairs Office, and the University Office of the Zurich University of the Arts for their constant technical and administrative support.

Jacques Borel, who brought the entire text into a readable graphic form despite the extensive size of the book and the many and time-consuming corrections that had to be made during the editing process.

I express my profound gratitude to everyone who has contributed to this work! It has been a long journey, and it has been not only exciting but crucial for me to work with you all. Without you, this book would not have been possible, and the question of the vowel sound – and with it the sound characteristic of the human voice – would not have been reopened.

Thank you!

Funding by the Swiss National Science Foundation SNSF

This publication relates to two research projects on vowel and voice qualities supported by two grants from the Swiss National Science Foundation SNSF (grant no. 100016_143943 and no. 100016_159350).

The subTexte Series

This book is published as volume 30 of the series subTexte, edited by Anton Rey, Institute for the Performing Arts and Film IPF, Zurich University of the Arts ZHdK. The subTexte series is dedicated to presenting original research within two fields of inquiry: Performative Practice and Film. The series offers a platform for the publication of texts, images, or digital media emerging from research on, for, or through the performative arts or film. The series contributes to promoting practice-based art research beyond the ephemeral event and the isolated monograph, to reporting intermediate research findings, and to opening up comparative perspectives. From conference proceedings to collections of materials, subTexte gathers a diverse and manifold reflection on, and approaches to, the performative arts and film. – All available titles can be obtained directly from the ZHdK publication platform:
https://subtexte.ch

Summary

> "An Index is a sign which refers to the Object that it denotes by virtue of being really affected by that Object. [...] In so far as the Index is affected by the Object, it necessarily has some Quality in common with the Object, and it is in respect to these that it refers to the Object."
> (Charles Sanders Peirce, CP 2.248)

It seems as if the fundamentals of how we produce vowel sounds and how they are acoustically represented have been clarified: We phonate and articulate. Using our vocal cords, we produce a sound, which is then shaped into a specific vowel sound by the resonances of the vocal tract. Accordingly, the prevailing acoustic description of vowel sounds relates to vowel quality-specific patterns of relative energy maxima in the sound spectra, known as patterns of formants.

In a first treatise, entitled Preliminaries, however, we have put forth intellectual and empirical reasoning that gives rise to scepticism with respect to such an understanding. Among the many critical arguments and observations, including the striking and unpredicted variability of the vowel spectrum and the measurement limitation of its envelope demonstrated, one phenomenon directly opposes the formant thesis: The spectrum of natural voiced vowel sounds is related to fundamental frequency and, therefore, formant patterns (if measurable) are ambiguous in that they represent different vowel qualities for sounds with marked differences in fundamental frequencies. The same holds true for the entire spectral envelope. Thus, formant pattern and spectral shape ambiguity disprove the thesis that formant patterns or spectral shapes are vowel-specific. As a consequence, the question of the acoustics of the vowel – and with it, the question of the acoustics of the voice itself, essential for human speech – proves to be an unresolved fundamental problem.

In this second treatise, entitled Indices, the variability of the vowel spectrum and its dependence on fundamental frequency is revisited on the new empirical basis of the Zurich Corpus of Vowel and Voice Quality, driven by a twofold motivation: To formulate knowledge-based statements with regard to the spectral representation of vowel quality *in general* and to question the cause of the observable relation of the vowel spectrum to fundamental frequency. As a main result, three statements that serve as primary indices for a future acoustic theory are presented: Firstly, the vowel sound does not relate to fundamental frequency but to pitch (or to a comparable perceptual referencing to a sound pattern repetition over time). Secondly, the vowel sound is a kind of perceptual and acoustic foreground–background phenomenon. Thirdly, the spectral representation of vowel quality is nonuniform.

These primary indices are derived from the examination of natural, (re-)synthesised and filtered vowel sounds within eight fields of investigation: Natural vowel sounds, vowel spectrum and fundamental frequency; formant pattern and spectral shape ambiguity; vowel spectrum and age and gender of the speakers; vowel spectrum, phonation type and vocal effort; vowel sound, vowel spectrum and pitch; spectral variation of vowel sounds and its nonuniform character; vowel recognition of filtered vowel sounds; resonance characteristics of vowel sound production and their detectability in the acoustic analysis of radiated sound. Additional reflections and speculations concern the difference between fundamental frequency and pitch, vowel quality and its representation in the harmonic spectrum, and vowel quality as its own sound dimension (not subsumed under sound timbre) and, at the same time, as a phenomenon of a produced form of expression.

The treatise concludes with a reflection on the prerequisites and challenges of building a future acoustic theory of the vowel.

Contents

Part III Commentary

Materials

Note: No further details are given for Chapter 1 in the Materials. They are presented in the Handbook of the Zurich Corpus.

Introduction

Topic

If a vowel quality is recognised when listening to a sound, what acoustic characteristics relate to this quality *in general?*

A text of transition

This treatise is a transitional text. It begins with contents exposed in the middle of a previous treatise entitled “Acoustics of the Vowel – Preliminaries” (Maurer, 2016; hereafter referred to as Preliminaries), it replicates, renews and extends the corresponding experimentation and documentation of the relation between recognised vowel quality and acoustic sound characteristics, and it further explores said relation based on new experimental approaches, including sound (re-)synthesis and sound manipulation, the entire work being motivated to provide phenomenological and knowledge-based indications for the acoustic representation of vowel quality.

In the Preliminaries, the line of argument and the related documentation of natural vowel sounds and speech extracts is focused on falsifying the hypotheses of the prevailing acoustic theory that either formant patterns or spectral shapes are vowel quality-related. However, the re-examined contents of the Preliminaries and the newly conducted experiments and their results presented in this treatise serve a different purpose: They go beyond criticising an existing theory and take an essential part in formulating indications for future theory-building.

Preliminaries

On its part, the Preliminaries (p. 1) start with a summary of the prevailing acoustic theory of the vowel: “The vocal cords – when oscillating and modulating air expelled from the lungs – produce a sound (a source sound), which is transformed by the resonances of the pharyngeal, oral and nasal cavities: Depending on the position of the larynx, velum, tongue, lips and jaw, different shapes of these cavities are formed thus creating different resonance characteristics, allowing different vocal sounds (phones) to be produced and perceived accordingly. If a vocal sound is perceived to belong to a particular linguistic unit (more precisely, a basic linguistic unit, a phoneme), and if the cavity formed by the pharynx and the mouth remains open, then the sound produced is referred to as a vowel sound and its linguistic identity as a vowel quality or simply as a vowel.

“The prevailing theory of vowel acoustics begins with such formulations or similar ones. According to this theory, with respect to human utterances, the vocal cords produce a general sound, which is transformed into a specific vowel sound by the resonances of the (supralaryngeal) vocal tract: As human beings, we phonate and articulate.

“Because of this, vowel sounds, as sounds, are expected to exhibit relative spectral energy maxima in those frequency ranges that correspond to the resonances of the vocal tract during speech production. These spectral energy maxima are known as formants.

“Such a perspective gives rise to the prevailing psychophysical principle of the vowel: Vowel sounds that are perceived as having the same vowel quality have similar formant patterns, that is, similarly patterned relative spectral energy maxima. By contrast, vowel sounds that are perceived as different vowel qualities have dissimilar formant patterns.”

First and foremost, as is emphasised in the Preliminaries, everyday experience, statistical investigations of the acoustic characteristics of natural vowel sounds and vowel synthesis seem to confirm such an understanding: When we speak, our vocal cords vibrate, and we move our articulators (larynx, velum, tongue, lips and jaw) to form different vocal sounds. Moreover, we can often “lip read” speech, an ability highly developed by deaf people. In these terms, the prevailing theory seems self-evident. Furthermore, statistical investigations of the spectral characteristics of natural vowel sounds generally report vowel-related average patterns of relative spectral maxima or formants, at least for sounds of a given speaker group of children, women and men and their respective relaxed speech. In these terms, the prevailing theory seems empirically grounded. Finally, by transforming artificial source sounds with the help of filters, sound synthesis can produce recognisable vowel sounds in terms of a synthetic (re-)production, at least for lower fundamental frequencies that are not markeldy varied.

In light of this, referring to Kent et al. (2019), “… it may be tempting to regard the topic of vowel acoustics as basically settled and closed in contemporary science and practise.” Hence, existing questions regarding the analysis and determination of the acoustic characteristics of vowel sounds rarely address the basics of formant theory. (Exceptions concern the debate on formant patterns versus spectral shapes as the acoustic representation of vowel quality.) Rather, they are related to specific aspects of the complexity and dynamics in speech production and perception. To give some examples: (i) speaker- and speaker group-specific acoustic characteristics and their normalisation; (ii) different types of phonation and their impact on vowel acoustics (whispered, creaky and breathy phonation, voice register changes, etc.); (iii) different types of phonation and articulation and their effects on vowel acoustics (speech with different vocal effort levels, speech related to different emotional states, infant-directed speech, everyday speech versus speech in the field of the performing arts, including speech versus singing and the differentiation of different singing styles, etc.); (iv) static versus dynamic characteristics of speech (including the role of transitions); (v) speech disorders and their acoustic correlates. In addition, the method of formant pattern and spectral shape estimation is also an ongoing issue. (This list is far from complete, and no indications are given for the attempts to relate acoustic characteristics to the simulation of articulation, the auditory process, and therapeutic aids.)

However, notwithstanding, the text of the Preliminaries returns to the assertion of the prevailing acoustic theory of vowel-related formants, and it presents a critical reading – indeed a falsification – of this assertion. Referring to intellectual reasoning, followed by experimental investigation, it raises and reopens for discussion an unresolved fundamental problem of the voice and voiced speech sounds. Concerning intellectual reconsideration and validation, the Preliminaries elaborate on four reflections that oppose the understanding of prevailing theory: Vowels and numbers of formants; vowels and fundamental frequency; formant patterns and speaker groups; terms of reference and methods of formant estimation. Observations and experiences then follow these reflections within three main fields of investigation: Unsystematic, lacking or ambiguous correspondence between vowels and patterns of relative spectral energy maxima or estimated formant patterns for sounds of single speakers or speakers of a particular speaker group; lacking correspondence between patterns of relative spectral energy maxima or estimated formant patterns and different speaker groups or vocal tract sizes; lacking correlation between methodological limitations of formant estimation and limitations of vowel recognition.

In view of these reflections, observations and experiences, it is concluded that they represent a falsification of the prevailing theory (see Preliminaries, Part IV). Among the many critical arguments, the central problem can be formulated as follows: If the estimation of formant patterns or spectral envelopes of vowel sounds *is* methodologically substantiated, in the majority of cases, these patterns and envelopes are ambiguous in that they represent different vowel qualities for sounds with marked differences in fundamental frequencies; if the estimation of formant patterns or spectral envelopes is *not* methodologically substantiated – e.g. for middle and higher fundamental frequencies – vowel sounds remain recognisable despite this lack of measurable sound characteristics, assumed to be vowel quality-related. Furthermore, in this context, it also has to be taken into account that the existing documentation of vowel sounds hitherto published is fragmentary.

The Preliminaries end with the following appraisal: “Our vocal cords produce sound. The resonances of the pharyngeal, oral and nasal cavities could form its characteristics into a formant pattern that always and uniquely represents a vowel physically and thus allows the listener to perceive it accordingly. Empirical investigation reveals, however, that the spectral characteristics of vowel sounds systematically deviate from such an option. This observation leads to the conclusion that, at present, we are but in the preliminary stages of understanding the physical representation of the vowel and, thus, its materialised form.” (p. 93)

Status

Any present and future introduction to the acoustics of vowel sounds should start with the statement that contrary to a traditional understanding, patterns of spectral energy maxima or estimated formant patterns or spectral envelopes are *not* vowel quality-specific *per se, in general terms.* (If expressions such as “per se” or “in general terms” are used, they are often written in italics to emphasise that a given statement does not concern the characteristics of certain vowel sounds but of all of them.)

Up to now, the lack of an alternative theory might have been taken as a reason to maintain the traditional formant theory. The same holds true for its alternative, the spectral shape theory. But the lack of a new theory is not a reason to maintain any existing theory that can be empirically falsified, even if the latter is useful for modelling a limited set of vowel sounds and their acoustic characteristics.

However, building a new theory addressing the acoustics of the vowel is a highly challenging, laborious and time-consuming project, and it needs to be undertaken step by step. Thereby, the vowel sound may prove to be a phenomenon that is incomparable to any other sound characteristic or sound quality, and even the prevailing perspective and type of its acoustic representation to look at – the spectral perspective – may be called into question.

Towards a new theory – the need for phenomenological references and formulating indices

Taking into account more than 70 years of modern research on the matter, from the first studies using a spectrographic examination of vowel sounds up to now, we have such a great number of single studies at our disposal that they are barely surveyable, and no attempt at bringing them to a *general* statement in terms of a robust reference predicting the acoustic characteristics of vowel quality is able to withstand intellectual counterarguments and contradicting experimental results. Most probably, the disparity in the specialist literature, that is, the diverse approaches taken and the heterogeneous results and interpretations given, is a consequence of a lack of phenomenological references, which many of the different studies would mutually rely on when concluding on *general* vowel quality-related acoustic characteristics. Concerning acoustic characteristics, many existing studies have reported sound analysis results that are based on a set of vowel qualities whose production is limited to medium vocal effort at levels of fundamental frequency (hereafter f_o; abbreviation according to Titze et al., 2015) in the lower vocal range of the speakers, e.g. citation-form utterances produced with relaxed speech in a quiet room. Some studies have compared sounds related to two or three production parameter variations, e.g. voiced and whispered and/or creaky phonation, or V and CVC context (vowel sounds investigated as isolated sounds or sound nuclei versus sounds embedded in a consonantal context), still limiting the production of the voiced sounds to lower f_o levels of relaxed speech in a quiet room. Other studies have compared voiced sounds with varying vocal effort levels, in part combined with f_o variation (e.g. studies investigating different emotional characteristics of speech or studies investigating shouting). Most studies which included an extended f_o variation were concerned with singing, above all, singing in the European classical singing style. Similarly, many studies of vowel synthesis were related to a very limited variation of synthesis parameters.

However, when interpreting the results, many scholars assume that the spectral characteristics of vowel sounds are uniform or systematic and that single findings related to a specific set of vowels and sound production parameters allow for a generalisation more or less independent of that set. Yet, the variation of spectral characteristics of vowel sounds is related to specific vowel qualities in a nonuniform or unsystematic manner, and what may be true for a specific set of vowel qualities and sound production parameters may not be true for another. (For details, see the Preliminaries and extended demonstration in this treatise.)

Consequently, as said, future attempts to assess the *general* acoustic characteristics of vowel quality – more precisely, the differences of acoustic characteristics representing the differences of recognised vowel qualities – should relate to phenomenological references, that is, to systematic compilations of vowel sounds of a given language, including all long vowels of that language at a minimum in order to cover and differentiate the close–open, front–back and (if included in the vowel system of the language in question) unrounded–rounded dimensions, and also including the variation of basic production parameters relevant for vowel quality, such as age and gender differences of the speakers, phonation type, fundamental frequency, vocal effort and vowel context. (Note that long vowels fulfil the condition of quasi-static characteristics for their sound nuclei much better than short vowels, and the impact of sound duration and phoneme context on vowel production and recognition is likely to be much weaker for sounds of long than of short vowels; on this matter, see also Chapter 11.1.) In addition, different speaking and singing styles may also be integrated. Only on the basis of references of this kind is an evaluation and validation or rejection of an approach to defining the primary acoustic cues of vowel quality feasible.

However, a phenomenological approach of this kind will still have gaps, above all, because both manipulations of natural vowel sounds and vowel synthesis can produce sounds with recognisable vowel quality even though their acoustic characteristics deviate from any observed or observable natural sound. (Some types of such vowel sounds, from alien to natural, are presented and discussed in this treatise.)

The creation of phenomenological references as a first step towards building a new theory should then be associated with attempts to formulate knowledge-based statements that, *in general,* apply to the parallelism between differences in recognised vowel qualities and differences in acoustic characteristics of respective vowel sounds. Here, these kinds of statements are named primary indices: They are *general* indicators of the actual acoustic characteristics representing vowel quality, characteristics that are to be defined in a future theory.

Towards a new theory – the need for open-access publications of vowel sound corpora and related software tools

If, as a first step towards a new theory, phenomenological references are built up, they should be published open access, and they should include a user interface that allows for browsing, sound selection, sound playback, viewing sound spectra and sound export for further analysis. Such an open-access form and functionality that facilitates sound investigation is needed for a sound corpus in order to attain a reference status for verification or falsification attempts of future hypotheses.

However, the aforementioned disparity in the specialist literature is likely not only due to a lack of phenomenological references but also, at least in part, to a lack of open-access analysis and verification tools directly linked to the sound samples under investigation. As a consequence, up to now, the reconsideration, replication and validation of experiments and results published in many cases require a great effort, which poses a basic investigation problem because of the very large number of studies and disparate character of the literature. Some software tools used by existing studies are indeed available open access (e.g. Praat software, referenced below) and considerably facilitate the replication and verification of published results. But existing tools are rarely linked to the respective sounds and sound samples under investigation and rarely run online. (However, the technological progress made in the last years has to be accounted for when considering and appraising earlier studies, corpora and tools.)

A contribution – fields of investigation

Over the last years, we have worked on a contribution addressing the above demands within three main fields of documentation, investigation and reflection as well as the field of software tools:

- The creation of a new empirical basis and reference for natural sounds of all long Standard German vowels, including an extensive variation of basic production parameters, to be published in the form of a comprehensive open-access sound corpus with a corresponding user interface
- The compilation of extensive sound documentation related to specific aspects of natural vowel sounds and related acoustic characteristics thereof, and the creation and conduction of paradigmatic experimentation to further investigate vowel acoustics, including sound resynthesis, synthesis and filtering
- The formulation of knowledge-based, *general* rules concerning vowel quality recognition and related acoustic characteristics of vowel sounds
- Open-source web applications for spectral analysis, resynthesis, synthesis and sound filtering

In addition, reflections and speculations that exceed observational and experimental evidence were also made.

This treatise presents and discusses the research, the results, the conclusions and the development of software tools related to the above fields of investigation.

Sound corpus: The previous treatise, Preliminaries, was based on sounds of previous single studies, which were recorded under varying conditions and with varying sound qualities, and for which permission for an online sound playback could not be retrospectively obtained from all speakers. In order to study and provide sounds to the scientific community that are recorded under systematically controlled conditions, including an extended variation of production parameters and full permission for online audio playback, and also including a standard vowel recognition test, we have created a large new sound corpus termed the Zurich Corpus of Vowel and Voice Quality (hereafter referred to as the Zurich Corpus; Maurer, d'Heureuse et al., 2018 for version 1, 2024 for version 2).

Documentation, experiments, and formulation of knowledge-based general rules concerning vowel quality recognition and related acoustic characteristics of vowel sounds, termed primary indices: In parallel to the creation of the Zurich Corpus, extensive documentation of sound series was compiled, resynthesis, synthesis and filtering experiments were conducted, and vowel recognition tests were performed. This documentation and experimentation was based on several motivations: (i) To reconfirm and to extend the documentation already given in the Preliminaries on the new basis of the Zurich Corpus; (ii) to deepen the knowledge about the relation of the vowel spectrum to f_o and the resulting formant pattern and spectral shape ambiguity of vowel sounds; (iii) to replicate experiments already described in the literature but thereby further vary experimental parameters; (iv) to question the relation and relevance of lower and higher spectral frequency regions for vowel quality recognition; (v) in doing so, to contribute to formulating what can be said *in general* about the relation between vowel recognition and sound acoustics, that is, to work out knowledge-based *general* rules concerning this relation, termed primary indices here, and paving the way to a future theory of vowel acoustics.

Browser-based tools: In parallel to the above documentation and experimentation, browser-based web applications were created to allow for an acoustic analysis, resynthesis, synthesis and filtering of vowel sounds, including parameter variations, and at the same time to link the corresponding procedures to the sounds in the Zurich Corpus.

Reflections and speculations: Finally, reflections and speculations that exceed observational and experimental evidence also emerged. Here, they are exposed separately from the main text in the form of three excursuses concerning (i) the need to distinguish between fundamental frequency and pitch with regard to vowel quality recognition, (ii) a hypothesis predicting vowel quality-specific differences in the harmonic spectrum, and (iii) an attempt to understand vowel quality not as being an aspect of sound timbre but as being a sound dimension on its own.

It will all come down to three notions

It will all come down to three notions: The vowel sound relates to pitch (or to a comparable perceptual referencing to a sound pattern repetition over time), it is a kind of perceptual and acoustic foreground–background phenomenon, and the spectral representation of vowel quality is nonuniform.

Content and structure of the present treatise

The presentation and discussion of the research conducted and results obtained consist of a main body and a materials section (hereafter Materials).

The main body is further subdivided into three parts, followed by an afterword: Part I describes the Zurich Corpus in terms of the phenomenological sound corpus used for the present treatise. It further presents the software tools developed, integrated into the corpus user interface and used for the documentation and experimentation below. Part II outlines in short terms the documentation and experimentation conducted and summarises and discusses corresponding results, including exemplary sound series and graphic illustrations. Part III focuses on primary and secondary indices derived from the studies presented and reflects on a future theory of the acoustics of the vowel. In addition to these three parts, reflections and speculations exceeding observational and experimental evidence are exposed in the form of the three excursuses mentioned, integrated into the text of Parts II and III. Finally, a general valuation of the treatise is made in the Afterword.

Within the present scientific context, the text of Part II of the main body may seem unusual: Most of the presented studies are only outlined briefly, and only summarising tables and figures and a limited number of sound examples are given in terms of exemplary illustrations. With few exceptions, the text thus largely dispenses with extended background information and references to previous studies published in the literature (including our own) as well as details of methods and results in order to present the main argument without any detailed review and referencing of individual aspects, facilitating the reading and understanding thereof. This condensed presentation of the entire investigation is the result of a reaction to the many sounds and sound samples investigated, the high number of studies involved, the complexity of some of the results and their embedding in the existing phonetic knowledge so as to structure and organise the content of this treatise.

The Materials, however, present a more detailed and extended version of Part II of the main body: In order to meet the scientific standard, the documentation and experimentation conducted and the related discussions presented only in short in Part II are replicated in detail in the Materials section with extended background information and references, details of experimental design, method and results (combined with detailed documentation in the form of tables with sound links to the Zurich Corpus) and in-depth discussions. (Exceptions are the excursuses in the main body of the text, covering the entire content and its background, including references.)

This dual form of presentation offers different ways of reading, prioritising a main line of exposition and argument in the main body while providing all necessary details for a scientific appropriation and replication in the Materials. However, this parallelism of presentation also means various sections of parallel text in the main body and the Materials. Readers are asked to accept this redundancy as a consequence of the chosen form (see also the Afterword).

Figures, tables and additional links to sound series are always given at the end of a chapter (main body) or in the chapter-related appendices (Materials section).

If a text of another publication is directly cited in quotes in the Materials and includes references to other studies published, these references in the cited texts are not given in the References sections of this treatise. For corresponding details, please consult the cited publications.

Additions

Additional documentation of sounds and indications of experimental ideas will be progressively given in a side document of the Zurich Corpus (see Chapter 1.1).

How to read

Against the background of the content and the structure of the treatise, three options regarding reading and comprehension may be considered: (i) The reading of Parts I to III and the Afterword allows for an understanding of the main line of argument and taking note of corresponding main observational and experimental findings and their interpretation, including exemplary illustrations. This way of reading demands a limited effort. (ii) Alternatively, reading Part I, the Materials and the first two excursuses, Part III and the Afterword (in this order) gives a complete overview of background information, references, details of methods, a detailed presentation of results, including links to the investigated sound samples (accessible in the Zurich Corpus), and extended discussions. Such reading is time-consuming and requires a great effort. (iii) Finally, a somewhat intermediate way of reading may involve reading Parts I to III and the Afterword, and consulting the Materials only if needed. However, text redundancy will occur due to the aforementioned dual form of presentation.

Perspective adopted

As was the case for the Preliminaries (see pp. 4–6), the present treatise adopts a perspective akin to a psychophysical perspective, focusing on the relation between recognised sound quality and acoustic sound characteristics. It involves a possible reversal, as testing vowel recognition may precede acoustic analysis in the examination process.

In the first chapters of this treatise, only general reference is made to the production and perception of sounds: Sound production is referred to with regard to the variation of basic production parameters and the utterances of different speakers investigated. (However, note that within the framework of source–filter theory, formants as a result of acoustic analysis, such as LPC analysis, also refer to vocal tract resonances of sound production.) Sound perception is referred to because the considerations and reflections presuppose that the vowel sounds discussed can be attributed to (perceptually identified as belonging to) specific vowel qualities.

However, in the course of the experiments, their results and the related reflections, it turns out that the acoustic characteristics of recognised vowel quality do not relate to the f_o of the vowel sound (as an acoustic characteristic) but to pitch (as a perceptual characteristic). This outcome makes for a particular psychophysical perspective in a broad and language- and speech-specific sense. Also, the question of the direct relation between resonances of vowel sound production and acoustically identifiable resonances of radiated sound, as well as the related question of resonance patterns of sound production that differ from measured formant patterns and harmonic spectral envelopes of radiated sounds, arise and are addressed.

Further aspects of production and perception are not discussed. This does not mean that they are less important for the acoustic description of vowels than the aspects investigated here. It merely serves to focus on the psychophysical question of the vowel: "Given that an utterance – or its reproduction, manipulated or not, or a resynthesis or synthesis – is recognised as a specific vowel quality, which describable physical characteristic or which ensemble of physical characteristics may be said to represent that quality?" (Preliminaries, p. 4) In line with this, the present argument focuses on steady-state natural vowel sounds as monophthongs, produced in isolation or extracted from syntactic and semantic context as sound nuclei, and on their (re-) synthesis or manipulation.

Restricting the main consideration to vowel sounds that are isolated from syntactic and semantic contexts and exhibit quasi-static spectral features by no means implies that such static spectral features are an absolute prerequisite for vowel recognition. Thus, the restriction made here does not contradict the phenomena described in the literature concerning the relation between vowel recognition and dynamic sound properties. However, as was the case for the Preliminaries, this treatise again refutes the conclusion partly drawn in the literature that the recognition of isolated, steady-state vowel sounds with quasi-static spectral characteristics is impaired or even insufficient when compared to vowel sounds in a syntactic and semantic context that manifest pronounced dynamic acoustic characteristics, including transitions.

In other words: Perception has various complex strategies for recognising speech and individual speech sounds. In particular, concerning utterances, different strategies are developed for different speech and sound contexts, and concerning listeners, the strategies involve learning and training, attention level, intention and selective focusing, individual abilities and habits, etc. However, *structurally speaking,* the differentiation of vowel qualities is a general phenomenon underlying actual speech perception and related recognition of sound qualities. It is assumed here that this differentiation not only belongs to the core of human speech (see the Preliminaries, p. 6, and pp. 90–92) but at the same time refers to acoustic characteristics without which recognisable sound transmission (acoustic transmission) would not be possible:

Isolated vowel sounds are recognisable and, therefore, a vowel sound carries acoustic characteristics *directly* related to vowel quality. More precisely, a vowel sound carries acoustic characteristics allowing vowel quality differentiation. An acoustic theory of vowel quality should first address this principle of structural differentiation and its acoustic representation. Thus, psychophysics of the vowel in terms of formulating the relation between vowel quality differentiation and differentiation of acoustic characteristics of vowel sounds is possible.

As stated in the Preliminaries (p. 5), "… there is good reason to understand and pursue the psychophysics of voiced speech sounds as a phenomenology: That is, for research not to start from a model and to conduct single experiments based on it but instead from an open-ended and continually expanding collection and compilation of vocal utterances, together with a simultaneously evolving description of their acoustic characteristics related to recognised vowel qualities. Experimentations and theses then should emerge from that description." Note in this context that taking a phenomenological perspective and focusing on isolated steady-state vowel sounds implies particular attention to artistic expressions and related knowledge of vocal expression in the performing arts (see also the Afterword). Hence, this treatise is published by an institute affiliated with an arts university.

Terms and notation

To facilitate reading, the key terms, notation style and abbreviations adopted in this text are explained below in a content-related order. (Please note that, for this paragraph, parts of the terms and notation given in the Preliminaries, pp. 7–11, are re-presented here, with minor adaptations. Since this concerns the section terms and notation only, quotation marks were omitted for improved readability.)

Sound, noise: The distinction between sound ("Klang", a quasi-periodic sound with a harmonic spectrum) and noise ("Geräusch", a sound generally considered to be aperiodic) is made only when it matters for the argument. In all other cases, the term sound is used as a generic term.

Vocal tract: The term vocal tract is used as a short form referring to the supralaryngeal (or supraglottal) tract in terms of the pharyngeal, oral and nasal cavities.

Vowel sound, vowel quality, vowel notation: The term vowel sound refers to a single concrete vocal sound possessing linguistic value, that is, a phone. It is termed a vowel sound – in distinction from other phones – because it is perceived to have vowel quality (see below). According to the literature, vowel sounds are quoted in square brackets, for instance, [a]. In some cases, additional suprasegmental characteristics are also given, for instance, with regard to the distinction between [a:] in the German word 'Kahn' (long vowel sound) and [a] as in 'Kamm' (short vowel sound).

The term vowel quality denotes a class of vowel sounds for an individual language, that is, a phoneme. Thus, concrete single vowel sounds as phones are attributed to abstract classes of vowel qualities as phonemes. In the literature, vowel qualities are quoted between two slashes, such as /a/. Here, quotations accord to the symbols of the International Phonetic Alphabet (revised to 2005).

If the context allows, the terminological distinction between vowel sounds and vowel qualities is shortened to the distinction between sounds and vowels.

In general, most reflections, observations and experiences presented refer to the long vowels of Standard German /i, y, e, ø, ɛ, a, o, u/. The vowel /ɑ/ is included here because it can be encountered as a long vowel in some regions of Germany and Austria and also in some singing styles. Therefore, the indication /a/ in this text refers to the vowel area /a–ɑ/, that is, including all allophones of /a/ or /ɑ/.

In some experiments and/or listening tests, sustained sounds of the vowels /ə, ɔ/ were also included.

In this text, the vowels investigated are often subsumed into different quality subgroups: close /i, y, u/, close-mid /e, ø, o/, open-mid /ɛ, ɔ/ and open /a–ɑ/ area; front /i, y, e, ø, ɛ/, back /o, u/ and /a–ɑ/ area; front unrounded /i, e, ɛ/ and front rounded /y, ø/. (Because the Standard German vowels /ɔ, o, u/ are all rounded, no further roundness differentiation is made for back vowels.) Subdivision and terminology are adopted from the literature (see the Handbook of the International Phonetic Association, 1999), but they have no further significance here. In particular, their attributed association with the actual articulation in sound production (above all with tongue height and backness and with lip articulation) is not intended. In these terms, the terminology is adopted here for reasons of tradition and readability only. (Indeed, as will be reflected on in the Afterword, the use of these terms in the present context is open to criticism.)

Note that, depending on the subject of demonstration or discussion, the vowel order given in the text sometimes deviates from a consistent order of close–open, front–back and unrounded–rounded.

The documentation and experimentation presented in this treatise focus on German vowels because most of the author’s experiences and observations concern the sounds of the German language. However, the corresponding general statements on vowel differentiation and related acoustic differences are understood to apply to other languages as well.

Vowel sounds and vowel context: In most cases, vowel sounds were investigated and are documented here as steady-state sounds produced with a monotone intonation in isolation (V context, in short V). However, the Zurich Corpus also includes sounds which were produced in consonant–vowel–consonant–vowel context (sVsV context, in short sVsV, s = [s] or [z]) or in the context of syllables (syl context, in short syl) or minimal pairs (mp context, in short mp) or in speech (text read, in short tr) or singing (text sung, in short ts). In Chapter 2 of the main text, some documentation relates to sounds produced in these additional contexts.

Periodicity, fundamental frequency, pitch: The term periodicity commonly refers to a single vibration pattern as a single period which appears in successive repetition in a sound wave. Fundamental frequency (f_o; abbreviation as mentioned according to Titze et al., 2015) is a term used for either a periodic source characteristic of sound production or an acoustic measure of the radiated sound. These denotations pertain to physiology and acoustics. Pitch is a term for sound quality recognition. This denotation pertains to perception.

The f_o measurement of radiated sound depends on an applied algorithm. The pitch measurement of a perceived sound depends on a recognition test involving listeners.

f_o is often understood as being directly related to a sound periodicity generally represented in the sound spectrum by the first harmonic $H1$ as well as by the HCF of the harmonics. Pitch is also often understood as being related to a sound periodicity, but it is well known that this relation is not imperative.

Although often made, the distinction between periodic and aperiodic sounds is not trivial. The same holds true for the assumption of a simple parallelism of periodicity, f_o and pitch. In the course of the treatise, the corresponding questions are directly exposed and discussed in Chapters 5 and 6 of the main text and the excursus on the matter. However, in the first four main chapters, the terms are used in a pragmatic way: If a natural sound (or sound nucleus) manifests a quasi-periodic vibration pattern with a very limited variation degree over its duration, as indicated by its harmonic spectrum, it is named periodic without further specification. The same holds true for a correspondingly resynthesised or synthesised sound with a periodic or quasi-periodic vibration pattern. If a natural vowel sound (or sound nucleus) manifests a periodic vibration pattern only in a rough approximation, as is the case, for example, for some sounds produced with creaky phonation, the sound is subsumed under the class of periodic sounds, but sometimes an additional comment is added. If the spectrum of a natural vowel sound does not manifest harmonics, the sound is named aperiodic. In all other cases, the question of periodicity is explicitly commented on or discussed. Furthermore, if the differentiation of f_o and pitch is not important for a specific aspect discussed, only f_o or both terms in parallel are used.

Note that, dependent on a given documentation or experimentation, two different types of f_o levels are given: either the intended f_o (and pitch) during sound production or the measured f_o of the radiated and perceived sound. Furthermore, f_o levels are given either as levels according to the musical C-major scale notes (in Hz or according to musical notation) or as the measured levels (in Hz). For the reference of the C-major scale notation and the corresponding frequency values, see Titze (2000, p. 293).

Spectrum, spectrogram: The term spectrum refers to the sound spectrum of a vowel sound, generally resulting from a Fourier analysis related to a given time segment. In some instances, the term can refer to a spectrogram, generally understood as the variation of the spectrum over time.

Harmonics, harmonic spectrum, highest common factor (HCF): The term harmonic spectrum refers to a series of harmonics in the sound spectrum, namely a series of quasi-sinusoidal or sinusoidal components of a complex tone whose frequencies are an integral multiple of the fundamental frequency. The harmonicity of this type of spectrum can be expressed by the highest common factor (HCF) of the frequencies of the harmonics.

However, even if this terminology is common, it is not unquestionable. Above all, vowel spectra may not always exhibit the first (or the first few lower) harmonics (consider, for example, high-pass filtered sounds), and the harmonic structure of a spectrum may be only vaguely manifest (consider, for example, breathy or creaky vowel sounds). As said, this matter is investigated and discussed in the course of this treatise. Here, harmonics are abbreviated as *H*1, *H*2, *H*3, …, harmonic frequencies are abbreviated as H_1, H_2, H_3, … and given in Hz, and harmonic levels are abbreviated as L_{H1}, L_{H2}, L_{H3}, … and given in dB. Patterns of harmonics are abbreviated as *H*-patterns. This form of abbreviation was adopted to allow for a notation of *H*-patterns with without *H*1. For this reason, the notation deviates from the proposition of Titze et al. (2015). Also, for design reasons, abbreviations in the figures and the tables (and in part also in the Zurich Corpus) deviate from the above definitions: They are given without font effects such as italics and subscripts. In these cases, the abbreviations used are explicated in the legends.

Note that the term *H*-pattern is ambiguous since it can denote either a pattern of harmonic frequencies or a pattern of harmonic numbers. However, the context in which the term is used generally clarifies the denotation.

In some experiments, incomplete harmonic spectra in terms of selected dominant harmonics were investigated. In these cases, the character *H* is replaced by *D* in the corresponding abbreviations. Also, in some experiments, incomplete harmonic spectra in terms of series of sinewaves limited in sinewave numbers were investigated. In these cases, the character *H* is replaced by *S* in the corresponding abbreviations.

Partials, partial spectrum: Here, the term partial spectrum refers to a series of quasi-sinusoidal or sinusoidal components of a complex tone whose frequencies are not an integral multiple of a fundamental frequency and do not have a common (quasi-equal) frequency distance that can be compared to a harmonic spectrum of a natural vowel sound. (Note, however, that there is no clearly defined difference between a harmonic spectrum and a partial spectrum, and transitional phenomena are to be expected for experiments on vowel sounds.)

In some experiments, partial spectra in terms of series of sinewaves with no HCF comparable to natural vowel sounds are investigated. In these cases, the partials are abbreviated according to the abbreviations of harmonics, replacing the character *H* with *S* and adopting the simplifications mentioned.

Spectral shape, spectral envelope, filter curve, harmonic envelope: The terms spectral shape and spectral envelope are used here as synonyms because the term spectral shape is commonly understood as the pitch-independent envelope of the spectrum derived from some kind of smoothing operation (see Hillenbrand and Houde, 2003). In speech analysis, it is common to apply LPC analysis to vowel sounds (see below). If LPC measurement is methodologically substantiated, the resulting LPC filter curve represents one type of spectral envelope. – In early studies, the term spectral envelope corresponded to an imaginary smooth line drawn to enclose an amplitude spectrum, above all, to enclose the amplitudes of the harmonics (harmonic envelope).

Relative spectral energy maximum, spectral envelope peak: The term relative spectral energy maximum refers to a narrowly delimited frequency range of a spectrum that exhibits a markedly increased energy level compared to the frequency ranges immediately preceding and immediately following. In the literature, such relative maxima are generally determined by estimating formant patterns (see below) or spectral envelope peaks in terms of peak frequencies and related bandwidths.

Here, spectral peaks are abbreviated according to the abbreviations of formants (see below), replacing the character *F* with *P* and adopting the simplifications mentioned.

Formant, formant pattern, formant statistics: The term formant is used in different ways in the literature. In particular, it can refer to a resonance as a physical property of the vocal tract (or a corresponding filter of vowel synthesis), to a spectral envelope peak as a physical characteristic of a radiated vowel sound, or to a filter as a part of a series of filters used for acoustic sound analysis and related to an analytical method of speech processing (e.g. LPC analysis). The term can also denote two or even all three of these aspects simultaneously.

Here, for natural sounds, a fundamental distinction is made between the resonances of the vocal tract and the formants of the vowel sound produced and radiated. This distinction corresponds to the perspective adopted, namely, not to discuss in detail the production process of a vowel sound but, instead, the radiated vowel sound itself, including the related recognition of the corresponding vowel quality. If formant patterns are given below, they correspond to values estimated for a radiated sound, the estimation based for the most part on LPC analysis (see below).

In the text, according to Titze et al. (2015), formant frequencies are referred to as F_1, F_2, F_3, …, and configurations, referred to as F_1–F_2 or F_1–F_2–F_3 …, are termed formant frequency patterns or *F*-patterns in short. Frequency levels are given in Hz. Formant bandwidths are referred to as B_{F1}, B_{F2}, B_{F3}, … and are also given in Hz. Formant levels are referred to as L_{F1}, L_{F2}, L_{F2}, … and are given in dB.

Where formants are considered without specification of frequencies, bandwidths and levels, they are referred to as *F1*, *F2*, *F3*, …, and configurations of *F*1–*F*2 or *F*1–*F*2–*F*3, … are termed formant patterns or *F*-patterns in the sense of a more generic term. Note that, like the term *H*-pattern, the term *F*-pattern is ambiguous since it can denote either a pattern of formant frequencies or a pattern of numbers of formants. However, the context in which the term is used generally clarifies the denotation.

For formants of resynthesised or synthesised vowel sounds, a single quotation mark is often added to the abbreviations in the literature. However, here, no such additional mark is used. (Note that, for synthesis experiments, the term formant is "inverted" because it denotes a characteristic of sound production and not a measured characteristic of the radiated sound. The arising terminological problem is left open here.)

If references are made to formant values with the phrase "as given in formant statistics" – or simply named "statistical *F*-patterns" –, corresponding investigations generally concern formant measurements for sounds produced in citation-form words with a medium or spontaneous vocal effort at related lower fundamental frequency levels, in a quiet room in front of a microphone. These values are often assumed to be representative of so-called "normal speech" or "conversational speech" or "relaxed speech", and the limitation of measurement in terms of not considering vowel sounds produced at very different fundamental frequencies is often ignored and remains unmentioned.

Vocal tract resonances are abbreviated according to the abbreviations of formants, replacing the character *F* with *R* and adopting the simplifications mentioned. (Note the difference between this use and the notation of resonance characteristics proposed by Titze et al., 2015.) For the ongoing debate on terminology and abbreviations, please refer to the Preliminaries, Chapter M6.

LPC: The abbreviation LPC stands for Linear Predictive Coding, a method used to analyse the acoustic characteristics of speech sounds.

Manipulation, resynthesis and synthesis of vowel sounds: The term manipulation denotes a manipulation of a natural vowel sound (e.g. sound filtering or amplifying or attenuating levels of single harmonics).

The term resynthesis (or vowel resynthesis) denotes a synthesis (e.g. Klatt or sinewave or harmonic synthesis) based on an estimated *F*-pattern or spectral envelope (including LPC filter curve) or *H*-pattern of a natural sound, independently of whether this type of synthesis is also based on the calculated f_o of that sound or whether f_o is varied.

Accordingly, the term synthesis (or vowel synthesis) denotes any other type of synthesis procedure.

Indications of frequency ranges and frequency limits: General frequency ranges and limits for observed aspects of f_o and spectral characteristics (including formant patterns and spectral shapes), as well as for methodological considerations, are given as rough approximations.

Speakers, speaker group: Both speakers and singers, whether professionally trained or not, are referred to here as speakers as a generic term. (For detailed speaker information, see the Zurich Corpus.)

The term speaker group is used as a short form for age- and gender-specific groups of speakers, that is, children and adults and women and men, as they are referred to in the literature. As explained in the text, the differentiation of these three speaker groups is motivated by three different average vocal tract sizes. (Note that some scholars differentiate further in terms of age, gender and size, which is ignored here.) If only aspects of gender are discussed, then gender is given as women (w) and men (m) or female speakers and male speakers. If gender relates to speakers of the Zurich Corpus, gender relates to their self-denomination.

In the literature, age- and gender-specific speaker groups are generally given in the order of men, women, and children. However, a systematic adherence to this order carries an age and gender bias and poses a corresponding problem. Moreover, it mirrors a tradition in phonetics to favour the analysis of men's voices (see Kent et al., 2002, pp. 189–190). In order to counterbalance this bias, as a temporary solution, different orders are given in this treatise. (Sometimes, the order is adapted to the content of the investigation or the literature cited.) For future investigations in phonetics, the standard for the listing order of these speaker groups should be discussed, and an adequate linguistic form should be established in terms of a reference.

Further differentiation of speaker subgroups, related to nonspecific or specific speaking and singing styles: According to the standard of the Zurich Corpus (see Chapter 1.1), speaker subgroups are further differentiated: untrained and nonprofessional speakers, actors and actresses of straight theatre and singers of contemporary singing style (and their substyles) and European classical singing. Accordingly, speaking and singing styles were differentiated: nonstyle (untrained speakers producing sounds, or professionally trained speakers producing sounds abandoning a specific style and favouring the intelligibility of vowel quality over sound timbre); straight theatre speaking style (ST, including actors and actresses in films and relating to utterances according to an artistic performance), contemporary singing style (CS, including contemporary musical theatre, pop and jazz, and relating to utterances according to a professional singer's artistic performance), and European classical singing style (EC, relating to utterances according to a professional singer's artistic performance). Some studies refer to other types of speakers and production styles, which will be explicitly discussed.

Vocalises: The term vocalises refers to a series of natural sounds of a given vowel produced by a speaker at stepwise increasing or decreasing f_0 with the f_0 variation generally according to a musical scale.

Low-pass, band-pass and high-pass filtering, cutoff frequencies: A low-pass or LP filter attenuates frequencies above a frequency limit set, termed cutoff frequency or CF. Accordingly, a high-pass or HP filter attenuates frequencies below a CF, and a band-pass or BP filter attenuates frequencies below and above two CFs.

Listening tests, vowel and pitch recognition, recognition rate

Standard listening test procedure for vowel recognition and expert listener panel: For the natural vowel sounds documented in the Zurich Corpus and in this treatise, a standard listening test according to a standard procedure was performed when creating the corpus. Due to the vast number of sounds in the corpus, it was necessary to restrict the number of listeners. However, to account for this numerical limitation, vocally trained speakers and singers were involved in the listening tests. For details of the standard procedure and the listeners, see Chapter 1.1 and the handbook of the corpus in Maurer et al. (2024).

Vowel intention, vowel recognition, vowel recognition rate: The term vowel intention denotes the vowel quality intended by the speaker producing a vowel sound.

The term vowel recognition is generally used as a term for a vowel quality or a boundary between two vowel qualities recognised by a majority of listeners in a listening test.

The term vowel recognition rate is used for the ratio or percentage of listeners of a listener group who labelled a vowel quality or quality boundary. In most cases, the rate is given as a percentage even if the rate relates to the recognition results of the five standard listeners of the Zurich Corpus.

Experiment-specific listening test procedures for vowel recognition and expert listener panel: In many experimental studies reported in this treatise, the above standard procedure was also applied for vowel recognition. In the present text, for these studies, this is flagged accordingly, and only additional experiment-specific aspects of the procedure are described in detail. Otherwise, the entire experiment-specific listening test procedures are described in detail.

For several reasons, the same restriction in the number of listeners was made for most of the experimental studies conducted and the related recognition tests (including pitch recognition, see below), which were performed by the above five standard listeners: Above all, (i) the experiments conducted were phenomenological and explorative in their character, in their turn including a high number of sounds; (ii) often, they were the result of several pre-studies in order to configure a particular experimental setting, including recognition pre-tests; (iii) some of the studies concerned sounds with very different sound timbre because of marked variation of the production parameters, and vowel and/or pitch recognition had to be held apart from differences in sound timbre; (iv) some of the studies concerned sounds with impaired sound quality because of (re-)synthesis or sound manipulation, and vowel and/or pitch recognition tasks were difficult to perform; (v) some studies analysed in detail the recognition strategies and consistencies of individual listeners; (vi) for all these reasons, and because both vowel quality and pitch level recognition were subjects of investigation, we considered a precondition for an exploration as presented in this treatise that the listeners had extensive vocal and musical training. Since the five standard listeners involved in the standard listening test when creating the Zurich Corpus were professionally trained singers, actresses and actors and were experienced in recognition tasks because they performed extensive vowel recognition tests of the natural sounds during the creation of the Zurich Corpus, we decided to involve them also in most of the experimental studies as an expert panel. In these terms, most of the recognition results presented in this treatise correspond to the assessment of an expert listening panel, indispensable for an extensive exploration as presented here (see also Chapter 1.1).

However, the question is posed as to whether the recognition results obtained on this basis allow for generalisation, for general indications of the relation between vowel recognition and vowel acoustics. Future research will address this question. Yet, in our view, the need for replication and verification may be highly limited: Against the background of the entire line of experimentation and argument presented in this treatise, and on the bases of the sound samples of the experiments which are made accessible, including the software tools used for investigation, only few experiments addressing a few basic questions may have to be replicated – adapted to trained and untrained listeners – in order to confirm or reject the main conclusions made in this treatise.

Listening test procedures for pitch recognition and expert listener panel: Pitch recognition tests were experiment-specific in the sound presentation (either presentation of single sounds or presentation of sound pairs) and the labelling tasks (details are described in the method sections of the experiments). If not explicitly specified, however, the listeners used the same test screen as was the case for vowel recognition; for this purpose, the screen included a separate section for pitch labelling (for details, see the handbook of the Zurich Corpus).

In several listening tests, a single test item contained two sounds (separated by a 0.5 or 1 sec. pause), and the listeners were asked to identify the pitch level difference between the first and the second sound as falling (second sound lower in pitch than the first one), flat (no pronounced pitch level difference between the sounds) or rising (second sound higher in pitch than the first one).

Sound normalisation: Unless stated otherwise, sounds presented in the listening tests were normalised to the same RMS level of 0.2 relative to the maximum (for the corresponding procedure, see d'Heureuse, 2014).

Acoustic analysis, sound selection, numerical indications, graphic illustration

Standard of acoustic analysis: For a natural vowel sound produced in V context, acoustic analysis was conducted on the middle 0.3 sec. of the sound for a frequency range of 0–5.5 kHz on f_o contour and average f_o frequency (whispered and creaky sounds excluded), average spectrum, spectrogram, average formant pattern (frequencies, bandwidths, levels) and formant tracks. In addition, the average spectrum was also calculated for a frequency range of 0–11 kHz. Concerning formant pattern estimation, LPC analysis (Burg algorithm, window length = 25 ms, time steps = 5 ms, pre-emphasis = 50 Hz) was conducted in parallel for three parameter settings according to three commonly used age- and gender-related standards of 12 poles (standard for men; abbreviation in the tables = P6), 10 poles (standard for women; abbreviation in the tables = P5) and 8 poles (standard for children; abbreviation in the tables = P4) for the frequency range of 0–5.5 kHz (6, 5 or 4 formants at a maximum for that frequency range). The same analysis was conducted on sVsV sounds for the middle 0.3 sec. of the first or the second vowel sound, depending on their duration (for details, see the handbook of the Zurich Corpus, Chapter 3.3). Likewise, the same analysis was conducted on syllables and minimal pairs for the middle 0.3 sec. of the vowel sound in question. For read texts, songs/arias and speech extracts, the sounds were analysed for f_o contour, spectrogram (0–5.5 kHz) and long-term average spectrum (LTAS; 0–5.5 and 0–11 kHz). The acoustic analysis was conducted with a script using the Praat functionalities (for Praat software, see below).

For manipulated, resynthesised and synthesised sounds, acoustic analysis either corresponded to this standard or details of the analysis are specified in the method section of a given experiment.

Visual spectral crosscheck of results of acoustic analysis, direct visual estimation of spectral characteristics, sound selection: For some experiments, calculated *F*-patterns, spectral peaks and LPC curves were visually crosschecked based on the respective sound spectrum, spectrogram and formant tracks. If needed, parameters for LPC analysis were changed in order to improve the correspondence of calculated *F*-patterns and the vowel spectrum, spectrogram and formant tracks (see Hillenbrand et al., 1995, for this interactive estimation procedure). Also, for some experiments, the general spectral similarity of vowel sounds or their relative spectral maxima and dominant harmonics were estimated based on a direct visual appraisal of the respective spectra. Crosschecks and direct visual estimations are detailed in the method sections of the experiments described in the Materials. The viewing and appraisal of sound spectra and the related sound selection were conducted by the author.

A note on the methodological limitation of *F*-pattern and spectral envelope estimation: Methodological substantiation of *F*-pattern and spectral envelope estimation of a sound is related to its f_o level: With rising f_o, the two spectral estimates become problematic because of spectral undersampling and interrelated distortions, the estimation problem being severe for $f_o \geq 300$ Hz (see e.g. de Cheveigné and Kawahara, 1999; Hillenbrand and Houde, 2003). In the text, this methodological condition and limitation of *F*-pattern and spectral envelope estimation is referred to as an aspect of the general methodological estimation problem, or, for sounds produced at middle or higher f_o, as a lack of methodological substantiation. However, for natural sounds, the results of LPC analysis are always given independent of f_o, and the significance of these results is discussed separately for each experiment. For a detailed discussion of the matter, please refer to the

Preliminaries, Chapters 6, M6 and M11. In the future, further examples of sounds of the Zurich Corpus demonstrating the methodological estimation problem may be presented in the Additions section of the corpus (see Chapter 1.1).

Standard information for natural sounds: Below, the standard information in the display of a sound record in the Zurich Corpus (Layouts S, M, L, Twin) is given. For details of the production parameters and related abbreviations, see Chapter 1.1.

For all sounds, the following indications are shown in the first line of the sound legend: ID number of the speaker, gender (w or m), age (adults = A, children = C), production style, and record number of the sound in the corpus.

For vowel sounds produced in V or sVsV or syllable or minimal pair or word context, the following indications are shown in the second line of the sound legend: Vowel quality intended by the speaker, speech (including single words, syllables and single vowels) intended by the speaker, measured f_o in Hz (if validated), first language of the speaker, phonation type of sound production (v, b, c, w), vowel context (V, sVsV, mp, syl), vocal effort (med, low, high) and the five individual vowel quality assignments of the five listeners (if tested). For these sounds, the formant frequencies of LPC analysis (if applied) are given in the third line of the sound legend for a standard parameter setting that accords to the age or gender of the speaker (P4 = standard for children, P5 = standard for women, P6 = standard for men). Below these indications, the links to the tools for (re-)synthesis and sound filtering are given in the fourth line. In lines 5–8, all three patterns of formant frequencies of LPC analysis (if measured) for all three parameter settings are listed. If details are displayed by opening the "i" information, details are listed for sound and file information, speaker information, results of the listening test (if conducted), and indications on the sound selection (ranking) and the analysed sound nucleus (sound probe selection; for details of the ranking and the determination of the sound nucleus analysed, see the handbook of the Zurich Corpus). Furthermore, all three formant patterns of LPC analysis for the three standard parameters applied are given in full (and linked to the above tools), including average formant levels and bandwidths.

For speech and singing in terms of texts read (tr) or texts sung (ts), the following indications are shown in the second line of the sound legend: the intention of sound production (//Text//), first language of the speaker, phonation type (v, b, c, w, m), phoneme context (text) and vocal effort (med, low, high, vocal effort variation = var). If details are displayed by opening the “i” information, details are listed for sound and file information, speaker information and indications on the sound selection (ranking). (Indications for the sound probe selection can be ignored.)

Exceptions to the above standard: Exceptions to the above standard occur for duplicates of natural sounds or extracted sound nuclei used for recognition tests whose results are not related or are not entirely related to a single sound record, for some of the resynthesised sounds and for the synthesised sounds. In these cases, some of the standard indications are omitted. Furthermore, for synthesised sounds, the ID number of the speaker is replaced by an artificial ID number. Finally, for sounds of a few experiments, the display of *F*-patterns and LPC curves is disabled to put forward a direct spectral perspective.

Estimated *F*-patterns as given in figures and tables and in the Zurich Corpus: In general, estimated *F*-patterns as given in figures and tables correspond to the patterns as given in the Zurich Corpus in terms of default patterns (age- and gender-related default parameters applied in LPC analysis). Estimations based on non-default parameters are mentioned in the method sections and the tables. In some cases, marginal differences between the indications in the text, figures, tables (values that correspond to their calculation at the time of investigation) and the online corpus occur due to sound editing (improvement of on- and offset time for the display in terms of adding introductory and ending silence) and to a recalculation of the patterns when updating the corpus. Unless otherwise specified, these differences are ≤ 5 Hz and are neglectable.

Standard of graphic illustration in the Zurich Corpus: In the Details L layout of the corpus, for natural vowel sounds produced in V or sVsV or syllable or minimal pair context, graphic representation includes the display of the entire sound wave, the sound nucleus analysed, the measured f_o contour (if validated), the spectrum (0–5.5 and 0–11 kHz), the spectrogram and the three formant tracks for the three parameter settings mentioned. In addition, three related LPC filter curves of the middle window of the analysed sound nucleus are overlaid on the spectrum to illustrate the correspondence between spectral peaks and calculated formants.

For read texts, songs/arias and speech extracts, their graphic representation includes the display of the entire sound wave and the respective f_o contour, spectrogram and LTAS (0–5.5 and 0–11 kHz).

For manipulated, resynthesised and synthesised sounds, graphic illustration either corresponds to this standard or the illustration is specified in the method section of a given experiment.

The sound pressure level is given in terms of relative levels in dB, adapted for the display of the spectra.

All other layouts of the corpus present an extract of the above full graphic illustration.

Figures, figure legends and related sound links

In the main body of this treatise, figure legends and figures are given at the end of a chapter or an excursus, and in the Materials, figure legends and figures are given in the appendix to each chapter. Figures have a code indicating the general text part, the chapter and the figure order they belong to, given in square brackets. To give two examples: The code [C-02-01-F02] indicates that the figure belongs to the main body and Chapter 2.1, and that it is the second figure shown for this chapter; the code [M-06-04-F01] indicates that the figure belongs to the Materials and Chapter M6.4, and that it is the first figure shown for this chapter.

For most figures, a link to the corresponding sounds in the Zurich Corpus is given. For activation, please refer to the link symbols below the figure legends. In order to adapt the display of the sound spectra or f_o contours in the corpus to a figure given in the book, please refer to the records per page menu and adjust the size of your browser window.

As a standard, in the figures, the following indications are given for vowel sounds: First line of the legend of a single sound = number of the graph in the figure, intended vowel quality, intended or calculated f_o, vowel context, vocal effort, ID number, age group and gender of the speaker, recognised vowel quality and/or recognised pitch level (labelling majority; vowel qualities in square brackets; pitch levels in slashes, with /l/ = lower level of comparison, /ia/ = intermediate level between a lower and a higher level, /comp/ = comparable level to the level of a second compared sound, /h/ = higher level; double-vowel recognition is given as [d], and double-pitch recognition is given as /d/; if no labelling majority was found, vowel qualities are assigned as [–] and pitch levels are assigned as /–/). Second line of the legend of a single sound = record number of the sound in the Zurich Corpus and, if included in the acoustic analysis and graphic illustration, calculated F-pattern applying default parameters for LPC analysis (in most cases F_1–F_2 for sounds of /u, o, ɔ, a/ and F_1–F_2–F_3 for sounds of /i, y, e, ø, ɛ/).

Note that automatically calculated *F*-patterns with default LPC parameters related to the age and gender of the speakers may either not match with the sound spectrum, the spectrogram and the course of the formant tracks or lack methodological substantiation (above all for $f_o > 300$ Hz). For a crosscheck, please refer to the Details L layout of a sound in the Zurich Corpus. Note in this context that, for some experiments, *F*-patterns were crosschecked based on the sound spectrum, spectrogram and formant tracks, and changes in the LPC parameters were sometimes applied. The results of these experiments were always related to the *F*-patterns given in the tables in the Materials section.

For speech extracts, the following indications are shown in the figures: First line of the legend of a single sound = the number of the graph in the figure, indication [Speech] and ID number, age group and gender of the speaker. Second line of the legend of a single sound = record number of the Zurich Corpus.

Exceptions occur that are related to experimental settings and recognition tests. Above all, for the resynthesised and some of the synthesised sounds related to natural reference sounds, the ID number of the speaker is extended and marked with “res” (resynthesis) or “syn” (synthesis). For the remaining synthesised sounds, the ID number of the speaker is replaced by an artificial ID number (see above) combined with “syn”. For vowel synthesis, also, indications of synthesis parameters such as f_o, *D*- or *H*- or *R*- or *S*-patterns, and HCF and level attenuation of harmonics are added in the second line or a third one. If further specifications are applied, they are mentioned explicitly.

Correspondingly, in addition to intended vowel quality, intended or calculated f_o, vowel context, vocal effort, ID number, age group and gender of the speaker and *D*- or *F*- or *H*- or *R*- or *S*-patterns, the following characters and abbreviations are used in the figure legends, above all in the legends of the figures in Chapter 6 of the main text:

- Indication in [ ] = recognised vowel quality, including [d] = double-vowel recognition, [–] no labelling majority in the vowel recognition test
- Indication in / / = recognised pitch level, including /l/ = lower level, /h/ = higher level, /comp/ = approximately equal level, /ia/ = intermediate level/, /d/ = double-pitch recognition, /–/ no labelling majority in the vowel recognition test
- res = resynthesis, syn = synthesis
- AH1 = attenuation of the level of *H*1, AH(i) = attenuation of harmonics not being integer multiples of *D*1 frequency (for details, see Chapter M6.10)

Line breaks for the first or the second line occur when the text exceeds the text field width below a graph.

Tables and included sound links

The table presentation accords with the presentation of figures. To give two examples for the code of the tables: The code [C-02-01-T01] indicates that the table belongs to the main body and Chapter 2.1 of the treatise and that it is the first table shown for this chapter; the code [M-03-01-T02] indicates that the table belongs to the Materials and Chapter M3.1 and that it is the second table shown for this chapter.

If a table is of limited size, it is shown in full. However, if its size exceeds the page dimensions of the present book, only the summarising part is shown, and details are shown in full online (see the corresponding link in the table title).

In most tables, links to the sounds in the Zurich Corpus are included. For activation, please refer to the link symbols in the tables.

A note on the form of the abbreviations given in the figures and the tables

As mentioned above, abbreviations in the figures and tables (and in part also in the Zurich Corpus) are given without font effects such as italics and subscripts.

A note on sound playback (open-access sound playback and login-dependent sound playback)

In the Zurich Corpus, open-access playback functionality is given for all sounds for which, in the figures and the tables, sound links are provided. For all other sounds, a login is required for the playback functionality (see Chapter 1.1 and the handbook of the corpus). If access to sound playback is provided, a player or a corresponding symbol is displayed. Sound playback relates to sounds that are normalised to the same RMS level 0.2 relative to the maximum. An exception is the Details S layout: In this layout, three playback options are included related to the normalised sound, its nucleus used for acoustic analysis and the original sound as it was recorded.

A note on headphones and loudspeakers for sound playback

It is imperative to use state-of-the-art headphones to listen to the sounds presented in the Zurich Corpus that feature a playback option;

otherwise, sound quality is often significantly impaired. Above all, using PC loudspeakers and other loudspeakers and headphones with non-linear frequency characteristics will often cause unclear perception and recognition of vowel sounds. In some cases, the recognised vowel quality may be affected as a direct effect of sound distortion. (Please note again that, here, the vowel recognition is tested by professionally trained speakers and singers; see the above comment on this condition of the vowel and pitch recognition tests and further details given in Chapter 1.1).

A note on consistency of sound duration, duration of silent intervals, and fade in/out

If unmanipulated natural vowel sounds are referred to or are presented, their duration accords to the actual recording as indicated in the Zurich Corpus. The duration of resynthesised or manipulated natural vowel sounds is experiment-specific (for details, see the respective method sections). The same holds for sound synthesis, for intervals of silence between two vowel sounds presented as single test items in a listening test, and for the application of fade in/out. Accordingly, the corresponding durations are not uniform among different experiments. Experiment-specific sound durations, intervals and values for fade in/out were the result of general experiment-specific experimental designs, sometimes completed with specific settings made by the investigators or by the author that, related to the matter of investigation, were based on their perceptual appraisal of sound characteristics and sound quality. (Note also that the experiments were done at different times and in different experimental contexts.)

Software tools

Acoustic analysis: If not explicitly specified, acoustic analysis of the sounds was conducted using the Praat software (Boersma and Weenink, 2020; versions used from 2014 onwards). The corresponding Praat script used was written or revised by Daniel Friedrichs, Thayabaran Kathiresan, Volker Dellwo and Christian d’Heureuse.

Klatt synthesiser: Resynthesis and synthesis related to *F*-patterns were conducted using the Klatt synthesiser (Klatt, 1980; Klatt and Klatt, 1990), either as implemented in the Praat software using a script written by Thayabaran Kathiresan, or as a software tool developed by d’Heureuse (2019a, KlattSyn software tool; see also Chapter 1.2). In the Materials chapters, corresponding references are given in the method sections.

Sinewave synthesis: So-called sinewave vowel sounds were produced with a sinewave synthesiser developed by Christian d'Heureuse (2018; see also Chapter 1.2). The same holds for sounds based on series of sinusoids of various numbers and configurations.

Harmonic analysis and (re-)synthesis: Harmonic analysis of vowel sounds and the related harmonic resynthesis and synthesis were conducted using the harmonic analyser and synthesiser tool HarmSyn developed by Christian d'Heureuse (2019b; see also Chapter 1.2). The tool allows for extracting the harmonics of a natural sound, including the dynamic frequency and amplitude variations over time, and for subsequently resynthesising sounds related to the course of the harmonics. In addition, the tool allows for amplification, attenuation, or deletion of the level of single harmonics for synthesis. (Note that the HarmSyn tool is displayed only for sounds for which playback functionality is provided.)

LP and HP sound filtering: The filtering of vowel sounds was conducted using the filter tool implemented in the Praat software, with a corresponding Praat script written by Thayabaran Kathiresan. In order to allow for a verification of the filtering results by the readers of this treatise and the users of the Zurich Corpus, the filter tool SpecFilt was created by Christian d'Heureuse (2022; see also Chapter 1.2; note that the SpecFilt tool is displayed in the corpus only for sounds for which playback functionality is provided.)

Text characteristics

Wherever possible, closed compound words are used; e.g. nonprofessional (instead of non-professional), nonstyle (instead of non-style) and so forth. In-text citations and references are in APA style, except for the in-text references, where the "&" is replaced with "and", and the references to articles in proceedings, which are given in a short form.

A note on the term foreground–background phenomenon

As indicated, the vowel sound is understood here as a kind of foreground–background phenomenon. The related experimental basis is presented in Chapter 8, and the foreground–background thesis, including the term used, is further discussed in the excursus on vowel quality and harmonic spectrum. Note that the term "… does not refer to foreground–background relations as in auditory scene analysis, where a foreground communication signal may be extracted from an auditory scene and is then set in contrast to an unattended background noise". (Personal communication, K. Siedenburg, November 25, 2023).

Part I Phenomenological and Experimental Basis, and Software Tools

The first part of this text describes the Zurich Corpus of Vowel and Voice Quality (second version) in terms of the phenomenological sound corpus used for the present treatise. It further presents the software tools that were developed and integrated into the corpus user interface and used for the below documentation and experimentation.

1 Sound Corpus and Software Tools

1.1 The Zurich Corpus of Vowel and Voice Quality

As argued earlier (Maurer, d'Heureuse et al., 2018) and again discussed in the Introduction, existing databases of vowel sounds generally document only sounds produced with a limited variation of basic production parameters. Consequently, concerning the extent of the variation of vowel and voice quality-related sound characteristics occurring in everyday utterances and the field of the performing arts, there is a lack of phenomenological and descriptive references that allow for a comprehensive understanding of vowel acoustics.

Our first phenomenological studies and conclusions concerning natural vowel sounds (in most cases steady-state sounds in V context), presented in the Preliminaries, were based on sounds of various single studies (in part published as journal papers) compiled in a large sample of some 40 000 recordings in total. These sounds were produced by nonprofessional speakers (children, women and men) and professional singers (women and men), including phonation type and f_o variations. As a side sample, recordings of speech and singing were also included in order to demonstrate ranges and contours of f_o for various types or modes of everyday speech as well as of speech and singing in the field of the performing arts, that is, for utterances of stage voices in musical and straight theatre (including film). However, these sounds were recorded under varying conditions and with varying sound qualities. Different listening tests were conducted for the different studies, and the rights for online playback could not be obtained retrospectively from all speakers. Thus, the sound database the Preliminaries was based on did not have a systematic structure, and the audio playback function was not enabled for a large portion of the sounds.

Against this background, we have pursued the project of creating a new sound corpus to provide a systematic, large-scale database of sounds of the long Standard German vowels produced with extensive variation of basic production parameters. This corpus should serve as an empirical reference for the verification or falsification of any thesis regarding the acoustic representation of vowel quality *in general.*

Concerning our own research, the corpus should serve as a basis for replicating core experiments and related documentation presented in the Preliminaries (including an audio playback function for all sounds investigated). Furthermore, and most importantly, it should also serve as a basis for new experiments and new documentation that could reveal new indications regarding the acoustics of the vowel. Based on such a vast and systematic corpus, all the experimental results obtained should be fully verifiable and replicable, with the sounds investigated accessible in related publications.

Hence, the Zurich Corpus of Vowel and Voice Quality (in short the Zurich Corpus) was created in the form of an extensive unpublished sound database (hereafter termed work version), with selected, smaller open-access versions published online. The database is still being extended continuously. The online version 1 was already published earlier (see below, Maurer, d'Heureuse et al., 2018). The online version 2 was published in the context of this treatise (see below, Maurer et al., 2024).

The entire sound database (work version) consists of five different parts:

Part 1 – Natural vowel sounds, produced by single speakers with a systematic variation of basic production parameters; in addition, for each of the speakers, a read reference text and one or several songs sung were also recorded

Part 2 – Extracts of speech and singing documented from everyday utterances and utterances in the field of the performing arts

Part 3 – Syllables and minimal pairs produced by single speakers at different f_o levels

Part 4 – Manipulated natural, resynthesised and synthesised sounds

Part 5 – Miscellaneous

The first part addresses the question of observable acoustic characteristics of vowel sounds. The second part addresses the question of the observable f_o ranges in intelligible speech and singing. The third part documents vowel sounds produced in the specific context of syllables and minimal pairs by single speakers at various f_o levels. The fourth part documents sounds investigated in the context of different experiments related to sound filtering, resynthesis and synthesis. The fifth part consists of miscellaneous sounds that were cast aside during the creation of the corpus.

Below, details of all five parts of the entire sound database (work version, unpublished, including all recordings made until 2022) are given, followed by a description of the first two versions published online.

Part 1: Natural vowel sounds, produced with a systematic variation of basic production parameters

Part 1 of the corpus has a double structure: The main body consists of sounds of a large-scale investigation and documentation of the long Standard German vowels /i, y, e, ø, ɛ, a, o, u/, the sounds produced with extensive variation of basic production parameters by 16 nonprofessional and 24 trained and professionally active speakers and singers (hereafter nonprofessionals and professionals). For all speakers, a read text, and for professional speakers and singers, one or several songs are also included. The side body consists of reference recordings of the same set of vowel sounds produced by 30 nonprofessionals, with no production parameter variations except f_o variation within an everyday speaking range. A read text is also included.

Details of speakers, production parameters, recordings and listening tests are as follows. (For further information, see the handbook of the Zurich Corpus.)

Main body – speakers of extensive investigation: Sounds of untrained nonprofessionals (children and adults, gender-balanced) and professionals in the field of the performing arts (adults, gender-balanced) were investigated, the latter representing three different artistic production styles: straight theatre (ST), contemporary singing (CS, with substyles of musical theatre, pop, and jazz) and European classical singing (EC).

Nonprofessionals were selected according to two criteria: a vocal range of 24 semitones at a minimum (2 octaves) for adults and 19 semitones at a minimum (c. 1.5 octaves) for children, with recognisable vowel sounds over a range of at least 15 semitones for both adults and children. Professionals were selected according to a vocal range of 24 semitones at a minimum, their professional status, their praxis of performing in Standard German, their willingness to participate in a scientific investigation and their geographic availability. Their professional status was assigned according to Bunch and Chapman (2000), with ranking levels 2 or 3 of this taxonomy.

All speakers were native speakers of German, with origins in Germany, Austria or the eastern Swiss midland. An exception to this was the inclusion of four professionals (all singers), who were not native German speakers but who performed on stage professionally in Standard German.

Main body – vowel sound production, text read, songs sung: The speakers were asked to produce sustained sounds of the eight long Standard German vowels with varying basic production parameters regarding phonation type (voiced, breathy, creaky, whispered), vocal effort (medium, low, high), shouting, vowel context (V and sVsV) and, for sounds produced with voiced phonation except shouted sounds, f_o level variation according to the musical C-major scale, covering most of a speaker's vocal range. The speakers made all utterances as nonstyle productions in terms of favouring the intelligibility of vowel quality over sound timbre. Consequently, and most importantly, the professionals had to attempt to partially or fully abandon their artistic training and style of speech or singing.

In addition to the nonstyle utterances, the professionals were also asked to produce the same set of voiced sounds in their respective artistic performance style and for an f_o range reflecting their style, with a corresponding variation of vocal effort, vowel context and f_o. Thus, the vowel sound production of the professionals was investigated with regard to both their attempt at producing clearly recognisable vowel sounds and their performance in their respective professional styles.

The production of vowel sounds in sVsV context was limited to voiced sounds produced with medium vocal effort within an upper f_o range (≥ 523 Hz for children and women, ≥ 330 Hz for tenors and high male voices, ≥ 262 Hz for baritones and middle or lower male voices) and to shouted and whispered sounds, since the consonantal context was investigated only in terms of crosschecking its role for vowel recognition concerning three kinds of possibly critical vowel sound production: high f_o range, very high vocal effort and whispering.

The nonprofessionals also read a reference text on the spot ("Nordwind und Sonne", see Handbook of the International Phonetic Association, 1999, pp. 88–89) and were asked to sing a song in German. The professionals read the same text on the spot in nonstyle and style modes and sang a prepared song in German in their respective singing styles. Additional songs in Italian, English or French were also recorded for some professionals. (For details regarding recording procedure and exceptions, see the handbook of the corpus.)

Side body – speakers as additional references: Vowel sounds produced by 30 native Swiss German nonprofessionals (10 children and 20 adults, gender-balanced) with very limited f_o variation were also investigated. The speakers were selected according to their dialect (only speakers of Swiss German dialects from the eastern Swiss midland were included), their command of speaking and pronouncing Standard German (the primary language used in schools in this part of Switzerland) and their ability to produce recognisable vowel sounds on a specific pitch over an f_o range of 15 semitones at a minimum.

Side body – vowel sound production, text read: The speakers produced sustained sounds of the eight long Standard German vowels in isolation (V context) with medium vocal effort on reference f_o levels of 220–262–440–523 Hz for children, 220–262–440 Hz for women and 131–220–262 Hz for men. f_o variation was included in order to, firstly, investigate sounds that mirror the minimal range of f_o contours of everyday speech that has to be considered when investigating vowel sounds, that is, a range which can be observed with no pronounced register change for 262–523 Hz for children, 220–440 Hz for women and 131–262 Hz for men, respectively, and secondly, to allow for a comparison of sounds of different and similar f_o for different age- and gender-related speaker groups. In addition, the speakers were asked to spontaneously read out loud the above reference text.

This sample of reference speakers and utterances was collected to show that, in comparison to these speakers, all speakers of extensive investigation (with less regional restriction of speaker origin and, in part, with professionally trained voices) generally show comparable vowel pronunciation both in terms of acoustic characteristics and vowel recognition, given nonstyle mode, corresponding f_o and vocal effort levels of sound production.

Standard Listening Test: Five phonetic expert listeners (professionally trained singers, actresses or actors, or voice teachers, including the author of this treatise) performed a standard listening test for all vowel sounds of this first part of the corpus, labelling vowel quality. All listeners are native speakers of German with origins in Germany or the eastern Swiss midland. All listeners had extensive training in correct Standard German pronunciation within their professional voice training.

The vowel recognition test was organised into speaker- and style-specific subtests (blocked-speaker condition, further separating nonstyle and style utterances), with the sounds presented in random order. Before running a subtest, an extract of 50 sounds (or, for smaller subtests, all sounds) was played to familiarise the listeners with the speaker-specific phonation, articulation and production parameter variation. In the actual subtest, the listeners were asked to assign the recognised vowel quality of a sound in terms of labelling a single specific Standard German vowel /i–y–e–ø–ɛ–a–ɔ–o–u/ or a boundary region of two adjacent vowels. Alternatively, they could indicate “no vowel recognised” or write a comment. Sound playback repetition was allowed during the test. The vowel /ɔ/ was included in the listening test because the perceptual distance /a–o/ is very large, not representing adjacent vowel qualities. The assignment of the vowel /a/ included all variants in the region of /ɑ–a/ because the production of this vowel varies strongly among German speakers. (For further details of listeners and test conditions, including hardware and test screen, see the handbook of the corpus.)

Additional notes on speakers, listeners and vowel qualities: The geographic origin of the native German speakers is broad, and four professional singers were not native German speakers. However, here, reference is not made to the origin of a speaker but to the vowel quality recognised by trained listeners. Thereby, the main focus lies on sounds for which a unanimous vowel quality is recognised among the listeners. Interestingly, independent of speaker origin, we experienced highly comparable vowel sound production for the long close and close-mid German vowels /i–y–e–ø–o–u/. In contrast, sounds intended as /a/ often related to the geographic origin of a speaker, and their qualities concerned /ɑ/ and /a/ and the /ɑ-a/ boundary. As a consequence, /a/ in this text refers to the vowel area /a–ɑ/, that is, including all allophones of /a/ or /ɑ/ or their boundary (see the Introduction). Sounds intended as /ɛ/ also varied in actual vowel quality to some extent. Even if these sounds were labelled as /ɛ/ in the listening test, according to the author’s estimate, the actual quality sometimes corresponded to the vowel boundaries of /ɛ/ and one of the adjacent vowels /æ/, /œ/ or /ə/. However, this variation was not related to the origin of the speakers but rather to the different parameters of sound production.

In this context, it is important to note that timbre variation of allophones – or shifts to vowel boundaries or even to an adjacent quality – due to an extensive variation of production parameters are often marked for sounds of a single speaker (see the corresponding indications given in the literature, as referenced in Chapters M2, M5 and M7 of the Materials; for further evidence, also refer to the sound samples of single speakers documented in the Zurich Corpus). Thus, intra-speaker variation regarding allophones or vowel quality shifts often markedly exceeds inter-speaker variation for sounds produced at a given f_o level (see statistical F-patterns reported in the literature). Again, the results of the recognition test represent the most reliable reference for the actual vowel quality of a sound.

Note again that all speakers termed here as native German speakers were asked to produce sounds of Standard German, referring to the Standard of German used in their respective school environment.

Due to the vast number of sounds subjected to the standard listening test of the entire corpus, the extensive variation of production parameters, the testing of vowel boundaries and the long period of data collection, it was necessary to restrict the number of listeners and to involve vocally trained speakers (actresses and actors) and singers in the listening tests. (Note in this context the differences in vowel recognition between trained and untrained listeners and the effects of training in listening tests; see e.g. Carpenter and Morton, 1962, and Hillenbrand et al., 2011; concerning corresponding differences in pitch level recognition, see e.g. Titze, 2000, p. 212, citing Murray, 1990, and Murray and Zwirner, 1991.) Furthermore, because of the long period of data collection, sound editing and continuous listening tests, only three of the five listeners were able to participate in all subtests, one listener was replaced once, and another listener was replaced three times in order to keep the number of listeners always constant; however, all listeners were professionally trained speakers or singers, as described above. Throughout the course of the investigation, the gender distribution among the listener panel was either three women and two men or two women and three men. The author was part of the listener group. All listeners were remunerated.

Part 2: Extracts of speech and singing

Part 2 of the corpus consists of speech extracts produced by speakers without formal vocal training, by politicians, journalists and TV hosts, and by professionally trained speakers from the field of the performing arts (actresses/actors and singers). The utterances were either recorded in person (live recordings conducted by the author, with consent for publication given by the speakers) or extracted from taped TV shows, Internet content or DVDs/CDs. The extracts aim to document and highlight the observable f_o range found for everyday speech and for speech and singing in the field of the performing arts.

Part 3: Syllables and minimal pairs

Part 3 of the corpus consists of syllables and minimal pairs produced by single speakers at various f_o levels. It includes sounds collected in the context of studies on the intelligibility of vowel sounds produced at middle and high f_o levels (see Maurer et al., 2014; Friedrichs, Maurer and Dellwo, 2015; Friedrichs, Maurer, Suter and Dellwo, 2015; Friedrichs et al., 2017) and sounds of selected professional speakers recorded in the general context of the building up of the corpus.

Part 4: Manipulated natural sounds, resynthesised and synthesised sounds

Part 4 of the corpus consists of manipulated natural sounds and of resynthesised and synthesised sounds investigated in the context of specific experiments, that is, LP- and HP- filtered sounds as well as resynthesised and synthesised sounds using a Klatt synthesiser, sinusoidal synthesiser or harmonic synthesiser (for the corresponding tools, see the next chapter).

Part 5: Miscellaneous

Part 5 of the corpus is comprised of additional natural sounds that do not belong to the sound sample of the previous parts. They consist of the following categories: (1) Sounds produced by speakers who were not able to satisfactorily produce vowel sounds of sufficient quality for all investigated production parameters during the recording sessions; (2) various sounds of the vowel /ɔ/ that, initially, were intended to be part of the sample of Part 1 but proved to be too difficult to produce for some speakers (above all for nonprofessionals; therefore, in the course of creating the Zurich Corpus, we decided not to pursue further recordings of this vowel while still keeping the sounds we had already recorded as part of the miscellaneous sounds of this fifth part); (3) sounds produced by some of the speakers in sVsV context at f_o levels in a lower frequency range not corresponding to the standard range of the sounds of Part 1; (4) duplicates (entire sounds or sound nuclei) of natural sounds used for specific experiments and related vowel and/or pitch recognition tests. In addition, for some speakers, a few glissandi and schwa sounds were also recorded.

A note on the redundancy of recorded sounds

In numerous cases, we kept duplicate recordings for one single production task (above all for the sounds in Parts 1, 3 and 5): If either the speaker or the investigator believed that a repeated recording of a specific utterance could improve sound and/or vowel quality, the recording was repeated one or several more times.

Acoustic analysis, numerical indications, graphic illustration

Details of the general acoustic analysis, sound-related numerical indications and graphic illustrations are given in the Introduction (see also the handbook of the published corpus).

The Zurich Corpus, Version 1

Based on this sample of the work database, the first published version of the Zurich Corpus (Maurer, d'Heureuse et al., 2018) presented selected sounds of Part 1. Since, in many cases, two or multiple recordings were made for a single speaker and a specific configuration of production parameters when creating the database, a systematic subset of sounds was compiled for publication: If only one sound was recorded for a specific configuration of production parameters, then this sound was selected; if two or more recordings for a specific configuration of production parameters were made, then the sound with the highest recognition rate, the longest duration and the smallest difference between the f_o intended by the speaker and the average f_o calculated for the radiated sound was selected (according to this order). For nonstyle productions and each level of vocal effort separately, the sound selection was further limited to an f_o range for which, for a single speaker, all vowels investigated were represented by a sound. (Note that not all speakers were able to produce a complete set of investigated vowels at the very lowest or highest limit of their vocal range. Therefore, f_o levels with incomplete vowel sounds for a specific production parameter setting were excluded.) For productions in style mode, the f_o range was generally set at the discretion and the style-specific range of the artist. (For further details of the sound selection, see the handbook of the published corpus.)

As a result, a systematic corpus with one sound per production task for the investigated f_o range of a speaker (if available for all eight long Standard German vowels) was created for the publication of version 1 of the corpus. Its main body presented some 33 710 recordings of sounds of all long Standard German vowels, read texts and songs/arias produced by 16 nonprofessionals (adults and children, gender-balanced) and 24 professionals from the fields of straight theatre, contemporary singing and European classical singing (gender-balanced), with extensive variation of basic production parameters as described above. Its side body presented 830 recordings of sounds of all long Standard German vowels (V context, medium vocal effort) and of read texts produced by 30 native German nonprofessional reference speakers (see above).

In these terms, the first published version of the Zurich Corpus encompassed some 34 540 recordings in total, with sound- and speaker-related information, graphic and numerical display of the acoustic analysis results and the standard listening test results. The corpus was endued with a graphic user interface and additional functionalities (playback, search, speaker and sound information export, and sound download). Also, a Klatt synthesiser as a web application was integrated into the corpus (KlattSyn tool, see next chapter). However, restrictions for the use of the corpus were applied since the use of the database is limited to scientific purposes only (for details, see the handbook of the corpus).

The Zurich Corpus, Version 2

The second published version of the Zurich Corpus, presented online in the context of the publication of this treatise (Maurer et al., 2024), consists of the sounds of the first version (with minor corrections), additional sounds of Part 1 of the work database, related to this treatise and the Additions section of the corpus, and the following selected sounds from Parts 2–5 of the work database: Speech extracts, syllables and minimal pairs, filtered sounds, resynthesised and synthesised sounds, sound samples of new speakers related to a variation of only a part of the investigated production parameters and sounds of /ɔ/ and of /ə/. These additional sounds relate to the tables and figures in this treatise or to the Additions section of the corpus. In total, at the publication date, c. 37 400 natural sounds and manipulated or artificially produced sounds are presented. (Note that this number will change with database updates as additional sounds will be added, above all sounds related to the Additions section of this treatise; see below.) Also, the graphic user interface of the corpus was further developed. In addition to the Klatt synthesiser (KlattSyn tool), it now features online tools for sinewave synthesis, harmonic analysis and (re-)synthesis, and sound filtering (SinSyn, HarmSyn and FiltSpec tools, see next chapter).

Furthermore, in this second published version, a new separate section is added and will be updated step by step (see Maurer et al., 2024, entry page, Additions). This section will present topic-specific sound examples selected from the entire sound database that are often related to this treatise in order to extend the exemplary sound documentation. Furthermore, this section will list and outline possible new experiments that allow for an empirical exploration of the phenomena discussed in this treatise.

Table 1 shows the general structure of this new version of the Zurich Corpus, and Tables 2 and 3 show the speakers and production parameters of systematic investigation as documented in the first part of the corpus. As mentioned, if a reference to the Zurich Corpus is made in this treatise or sound links for specific experiments are listed, the references and links relate to this second published version.

For first-time corpus users, please refer to the Help and Abbreviations menus: The submenu Assistant in the Help menu provides information on navigation, menu items, display of sound records, related sound information and graphic illustration as well as sound playback. The submenu How to Search in the Help menu gives detailed instructions on using the search form. The submenu Lightbox Help in the Help menu explains the functionality of the lightbox tool to manually create topic-specific sound compilations. The Abbreviations menu lists all abbreviations used for the sound information in the corpus and provides details concerning speakers and production styles, vowel qualities and vowel notation, production parameters and sound status. For all further details, please consult the handbook of the corpus. For links to the above menu items, see below.

Note again that login-free playback functionality is given only for the sounds for which sound links are provided in the figures and the tables. For all other sounds, a login is required (see the corresponding menu item). Without login, the sound player is disabled, and a link to the Terms of Use of the corpus is provided as a placeholder. This document lists the conditions for full open access to the playback functionality for all sounds and informs about the login request procedure (see also the title page of the Zurich Corpus, menu item Terms of Use). Furthermore, the document lists the conditions for downloading the entire database for scientific use.

Table 1. The Zurich Corpus, Version 2: General content. Column 1 = parts of the corpus. Column 2 = content of the parts. Columns 3 and 4 = links to the sounds and the speakers.
[C-01-01-T01]

Table 2. The Zurich Corpus, Version 2, Part 1: Speakers. Column 1 = style-related speaker groups (nonprofessionals and professionals; ST = Straight Theatre, CS = Contemporary Singing styles, EC = European Classical singing style). Columns 2–6 = number and age range of speakers (children and adults; f = female speakers, m = male speakers) of the main body (vowel sounds produced with extensive variation of production parameters) and the side body (voiced vowel sounds produced with limited f_o variation).
[C-01-01-T02]

Table 3. The Zurich Corpus, Version 2, Part 1: Production parameters. Column 1 = phonation type. Column 2 = vocal effort. Column 3 = vowel context (V = in isolation, sVsV = in /s/–V–/s/–V context). Column 4 = intended f_o level variation (musical scale = according to C-major scale; upper scale and reference f_o, see text). Column 5 = production style (N = nonstyle, S = style of ST, CS or EC).
[C-01-01-T03]

Links to the Zurich Corpus, Version 2
Title page: ↗
Assistant (Help menu): ↗
How to search (Help menu): ↗
Lightbox help (Help menu): ↗
Abbreviations: ↗
Handbook: ↗

Table 1. The Zurich Corpus, Version 2: General content. [C01-01-T01]

Part	Content	Sounds	Speakers
Part 1	Vowel sounds, text read, songs sung, systematic recordings	↗	↗
(thereof)	One sound per production parameter configuration (revision of version 1 of the Zurich Corpus; details see Tables 2 and 3)	↗	↗
(thereof)	Additional sounds related to this treatise and the Additions of the Zurich Corpus	↗	↗
Part 2	Speech extracts	↗	↗
Part 3	Minimal pairs	↗	↗
Part 4	LP/HP filtered, resynthesised and synthesised sounds	↗	↗
Part 5	Miscellaneous	↗	↗
Entire Corpus (Maurer et al., 2024)			↗

Table 2. The Zurich Corpus, Version 2, Part 1: Speakers. [C01-01-T02]

Speaker group	Children			Adults			Speakers
	f	m	age	f	m	age	(total)
Main body							
Nonprofessionals	4	4	7–10	4	4	23–40	16
ST actresses/actors	–	–	–	4	4	26–51	8
CS singers	–	–	–	4	4	26–50	8
EC singers	–	–	–	4	4	25–56	8
Side body							
Nonprofessionals	5	5	7–9	10	10	18–52	30
Total	9	9	–	26	26	–	70

Table 3. The Zurich Corpus, Version 2, Part 1: Production parameters. [C01-01-T03]

Phonation	Vocal effort	Vowel context	fo intended	Production style
voiced	medium	V	musical scale	N / S
voiced	low	V	musical scale	N / S
voiced	high	V	musical scale	N / S
voiced	medium	sVsV	upper scale	N / S
voiced	shouted	V	–	N
voiced	shouted	sVsV	–	N
breathy	low	V	–	N
creaky	medium	V	–	N
whispered	medium	V	–	N
whispered	medium	sVsV	–	N
voiced	medium	V	reference fo	N

1.2 Software Tools

As described in the Introduction, the standard acoustic analysis of the sounds of the corpus (average f_o, spectrum, spectrogram, formant tracks, average *F*-patterns, LTAS) was conducted with a script using the Praat sofware (Boersma and Weenink, 2020; earlier versions used from 2014 onwards). Also, a part of the source–filter resynthesis and synthesis experiments and all LP and HP filtering experiments presented in this treatise were based on the Klatt synthesiser and the filter options implemented in Praat. (For the other part of the source–filter resynthesis and synthesis experiments, see below, the KlattSyn tool.)

As further indicated in the Introduction, four software tools were developed by Christian d'Heureuse (2018, 2019a, 2019b, 2022) in the context of the experiments reported in this treatise and the creation of the user interface of the Zurich Corpus: a reviewed and adapted version of the Klatt synthesiser (KlattSyn), a sinusoidal synthesiser (SinSyn), a harmonic analyser and (re-)synthesiser (HarmSyn) and a sound filter (SpecFilt) tool. All tools are available as web-based applications, and they are also integrated into the functionality and user interface of the Zurich Corpus. The features of the tools are outlined in short below. For further details, source codes, demo versions and related comments, see the links in the References section. (Note that, in the References section, the publication year of the tools is given according to the Commits section of the referenced web pages.) For the functionality integrated into the corpus and its graphical user interface, refer to the Help menu in the user interface (the line of tools below a sound spectrum). For default parameters, see the parameter forms of the tools. Note also that each parameter field has a tooltip.

In this text and in the Zurich Corpus published online, the references to these software tools are given using the above abbreviations: Praat, KlattSyn, SinSyn, HarmSyn and SpecFilt.

Concerning the investigation presented in this treatise, as mentioned, a part of the Klatt resynthesis and synthesis experiments was based on the KlattSyn tool, the experiments related to synthesis based on static harmonic spectra of sinusoids and also on configurations of sinusoids without harmonicity were conducted using the SinSyn tool, and the experiments related to (re-)synthesis based on dynamic harmonic spectra of natural vowel sounds were conducted using the HarmSyn tool. The SpecFilt tool for sound filtering was developed for the publication of the second version of the Zurich Corpus.

KlattSyn

The KlattSyn tool is a redevelopment of the classic Klatt cascade-parallel formant synthesiser that allows for a source–filter sound synthesis. The concept of the synthesiser is elaborated in Klatt (1980) and Klatt and Klatt (1990), and the review and adaptations made for the newly developed KlattSyn tool are described in d’Heureuse (2019a).

In the Zurich Corpus, for every single sound, a link to the KlattSyn tool is listed. If activated, the parameter form for the synthesiser is displayed, in which, automatically, the results of acoustic analysis for a sound are inserted (average f_o and average frequencies, levels and bandwidths of the formants as a result of the standard analysis of the corpus, with age- and gender-related default parameters for LPC analysis applied). If needed, these parameters can be edited manually. Based on these parameters, the (re-)synthesised sound and its average spectrum and vocal transfer function are displayed as assessed by the Klatt synthesiser, and the sound can be played back and saved.

Thus, for a single vowel sound and its calculated average values for f_o and formants, the KlattSyn tool allows for a direct resynthesis to assess the perceptual relevance of an estimated *F*-pattern, that is, the perceptual correspondence of the original natural sound and the re-synthesised replica, the replica being produced based on the results of formant analysis. Further, the tool allows for an investigation of the perceptual significance of changes in the synthesis parameter setting, which is particularly important regarding the effect of source variation with a maintained filter pattern.

If the parameter form is reset, the Klatt synthesis can be configured for any set of parameters.

For a direct Klatt resynthesis (based on calculated f_o and formants) without display of the parameter form, a corresponding resynthesis play button and a related parameter field for f_o are given in the figure legend of a sound.

SinSyn

The SinSyn tool (d’Heureuse, 2018) is a tool for sound synthesis based on any series of sinewaves (frequencies, amplitudes and phases).

In the Zurich Corpus, for every single sound, a link to the SinSyn tool is listed. If activated, the parameter form of the synthesiser is displayed, in which, automatically, the results of LPC analysis for the first three formants (frequencies and levels as a result of the standard analysis of the corpus, with age- and gender-related default parameters for LPC analysis applied) are taken as *S*1–*S*2–*S*3 patterns according to the concept of so-called sinewave vowel sounds as discussed in the literature. Concerning the present investigation, the SinSyn tool allows for a three-sinewave synthesis that relates to an estimated F_1–F_2–F_3 pattern of a natural sound to assess the perceptual correspondence of the vowel quality of the two sounds.

Besides, sinewave synthesis allows for a resynthesis based on a series of sinusoids mirroring the harmonic spectrum of a natural reference sound. Further, sinewave synthesis also allows for investigating the vowel sound beyond the framework of source and filter, with any number and with or without a harmonic relation of the sinusoids. However, the quality of the synthesised sounds is very limited.

A corresponding synthesis play button is given in the figure legend of a sound for a direct sinewave synthesis (related to the calculated first three formants) without a display of the parameter form.

HarmSyn

The HarmSyn tool (d'Heureuse, 2019b) is based on an analysis and (re-)synthesis algorithm for quasi-periodic sounds. The sound analysis part allows for calculating the dynamic course of f_o and the harmonic spectrum (frequencies and amplitudes). The sound synthesis part allows for either a direct resynthesis in terms of calculating back a sound on the basis of acoustic analysis of the natural reference sound or a synthesis based on a manipulation of the analysed harmonic spectrum (deletion of single harmonics). In the command line mode, the synthesis based on a manipulation of the analysed harmonic spectrum includes the option of an alteration of harmonic amplitudes.

If the playback functionality is enabled for a sound in the Zurich Corpus, a link to the HarmSyn tool is listed. If activated, the parameter form of the tool is displayed, subdivided into the two parts of analysis and synthesis. For an acoustic analysis, the corresponding parameters can be selected. For a (re-)synthesis, the series of harmonics and their levels are inserted into the form after the analysis has been performed. If needed, the indications of the harmonic spectrum can be edited (enabling/disabling individual harmonics or, in the command line version, amplifying/attenuating their level). Subsequently, the (re-)synthesis can be performed, and the resulting sound can be played back and saved.

Thus, for a single vowel sound with quasi-periodic sound characteristics and with its calculated dynamic course of f_o and the harmonic spectrum, the HarmSyn tool allows for the resynthesis of the sound in order to assess the perceptual correspondence of the natural reference sound and its resynthesised replica. Further, the tool allows for an investigation of the perceptual correlate of selected harmonics and/or, in the command line version, of increasing or decreasing harmonic levels in sound synthesis, which is particularly important regarding the question of the spectral representation of vowel quality.

Notably, the HarmSyn tool allows for an investigation of the vowel sound beyond the framework of source and filter, and it is able to produce a very good sound quality of the (re-)synthesised sounds even for a highly reduced number of harmonics. With regard to sound quality, it far surpasses the Klatt and sinewave synthesisers.

SpecFilt

The SpecFilt tool (d'Heureuse, 2022) is a tool for LP, BP and HP sound filtering, including filtering based on a custom free-form filter curve. It allows for calculating a Fourier spectrum of a sound or part of it, for spectral filtering and subsequent inverse Fourier Transform and for final playback and saving of the filtered sound or sound part.

If playback functionality is enabled for a sound in the Zurich Corpus, for every single sound, a link to the SpecFilt tool is listed. If activated, the parameter form of the tool is displayed, in which the sound is inserted automatically. Filter parameters can be set, and the correspondingly filtered sound can be played back and downloaded. (Note that special attention should be given to the window function parameter.)

Concerning the matter of the present investigation, the SpecFilt tool allows for the verification of the LP and HP filter experiments and their reported results and for further exploration of the effect of sound filtering on recognised vowel quality, above all in the context of the foreground–background thesis put forward in this treatise. This kind of sound manipulation is again of particular importance regarding the question of the spectral representation of vowel quality.

For direct sound filtering without display of the parameter form, a corresponding resynthesis play button and related parameter fields for filter types and cutoff frequencies are given in the figure legend of a sound.

Part II Observations, Experiences and Experimentation

The second part of the main text outlines in short terms the documentation and experimentation conducted. It summarises and discusses the corresponding results on which the subsequent formulation and knowledge-based general rules and additional indications concerning the relation between vowel recognition and sound acoustics are based, including exemplary sound series and graphic illustrations. Also included are two excursuses.

For each chapter of this second part, in the Materials, references, extended background information, details of experimental design, method and results, an extended discussion and documentation of the sound sample and the results of the investigation (tables including sound links) are given.

2 Natural Vowel Sounds, Vowel Spectrum and f_o

2.1 Vocalises

In the literature, vowel recognition and spectral characteristics of natural vowel sounds produced in isolation or syllable context with extensive f_o variation are discussed mainly in the context of singing. Variation of f_o in speech – albeit in most cases for a more limited frequency range – is also discussed in the context of various specific aspects of speech, such as highly emotional, loud or infant-directed speech or with regard to speech within the field of the performing arts.

However, the question of whether or not vowel-specific spectral characteristics relate to f_o levels and the question of the actual f_o range for which vowel sounds are recognisable lie at the core of the understanding of vowel acoustics, independent of vowel context and independent of differentiations such as speech versus singing, relaxed versus loud speech, "normal" versus "emotional" utterances, indoor versus outdoor utterances and so forth. Hence, in this first chapter on natural vowel sounds, vowel spectrum and f_o, we address the question of the spectral characteristics of recognisable vowel sounds produced by single speakers with extensive variation of f_o (vocalises). The two subsequent chapters are concerned with investigating the upper-frequency limit of vowel recognition, and the fourth chapter addresses the question of f_o contours of intelligible speech, demonstrating the actual frequency extension of f_o contours observed in everyday speech and in the performing arts.

In one of our early studies on vowel acoustics, we investigated the spectral characteristics of natural vocalisations of Swiss German vowel sounds produced by untrained speakers as monophthongs at various f_o levels. As an observational result, the vowel spectrum indeed appeared to be related to f_o: Above all, for close and close-mid vowel sounds, spectral peaks and estimated formants below 1.5 kHz were found to shift upwards with substantially increasing f_o levels, a phenomenon we named the f_o-dependence of the vowel spectrum. We have further documented this phenomenon in the Preliminaries (pp. 158–169). However, the number of sounds per vowel was still limited, the sounds were recorded under varying conditions and with varying sound qualities, and the rights for online playback could not be obtained retrospectively from all speakers. Therefore, the documentation was renewed and extended based on the Zurich Corpus.

In the Zurich Corpus, recognisable vowel sounds for the eight long Standard German vowels are documented in a systematic way. The sounds were produced by single speakers at successive f_o levels (according to the musical C-major scale) and covering large ranges of f_o (see Chapter 1). For exemplary documentation in this treatise, the sounds of three single speakers (one child, one woman and one man) were selected from the corpus. For each speaker and each of the eight long Standard German vowels /i, y, e, ø, ɛ, a, o, u/, a sound series was compiled, the sounds produced in nonstyle mode with medium vocal effort in V context and with successive f_o variation within a range of f_o of 22 semitones for the child speaker and 31 and 34 semitones for the adult speakers (all according to C-major scale). The f_o ranges were 220–784 Hz for the 112 selected sounds of the child, 131–784 Hz for the 152 sounds of the woman and 110–784 Hz for the 168 sounds of the man. Almost all sounds were recognised in the standard listening test of the Zurich Corpus matching vowel intention, with an 80–100% recognition rate.

The documentation of the sound series produced by the three speakers confirms and exemplifies three main indications that have already been discussed in earlier studies on natural vocalises: (i) Vowel sounds are recognisable within an extensive range of f_o (for speakers with good vocal abilities, this range sometimes exceeds two octaves), (ii) vowel-specific spectral characteristics < 1.5–2 kHz relate to f_o, and (iii) this relation is nonuniform among different vowel qualities, frequency ranges and frequency shifts of f_o. These three indications are considered important references for vowel acoustics in general and are taken as a basis for the following arguments in particular.

As an example of the entire documentation created and presented in the Materials, Figures 1 to 3 show series of sounds and their spectra for the close back vowel /u/ (100% recognition rate for all sounds), the open vowel /a/ (100% recognition rate for all sounds) and the close-mid front vowel /ø/ (80–100% recognition rate for the sounds) produced by the woman over an f_o range of more than two octaves. The sound spectra exemplify the main indications mentioned: An extensive range of f_o of recognisable vowel sounds (all three series), a pronounced shift of the lowest spectral energy maximum for the sounds of close vowels (here /u/) for higher f_o levels above c. 250–300 Hz, conversely often nearly constant lower spectral energy maxima (here in the range of c. 1–1.5 kHz) for the sounds of the open vowel /a/ with rising f_o over an extensive frequency range (nonuniform character of the relation of vowel-specific sound characteristics to f_o), and a shift of the lowest spectral energy maximum for the sounds of /ø/ for f_o levels above c. 200 Hz, which is, however, difficult to assess in its full extent due to the frequency distance of prominent $H1$ and $H2$ for sounds at middle and higher f_o.

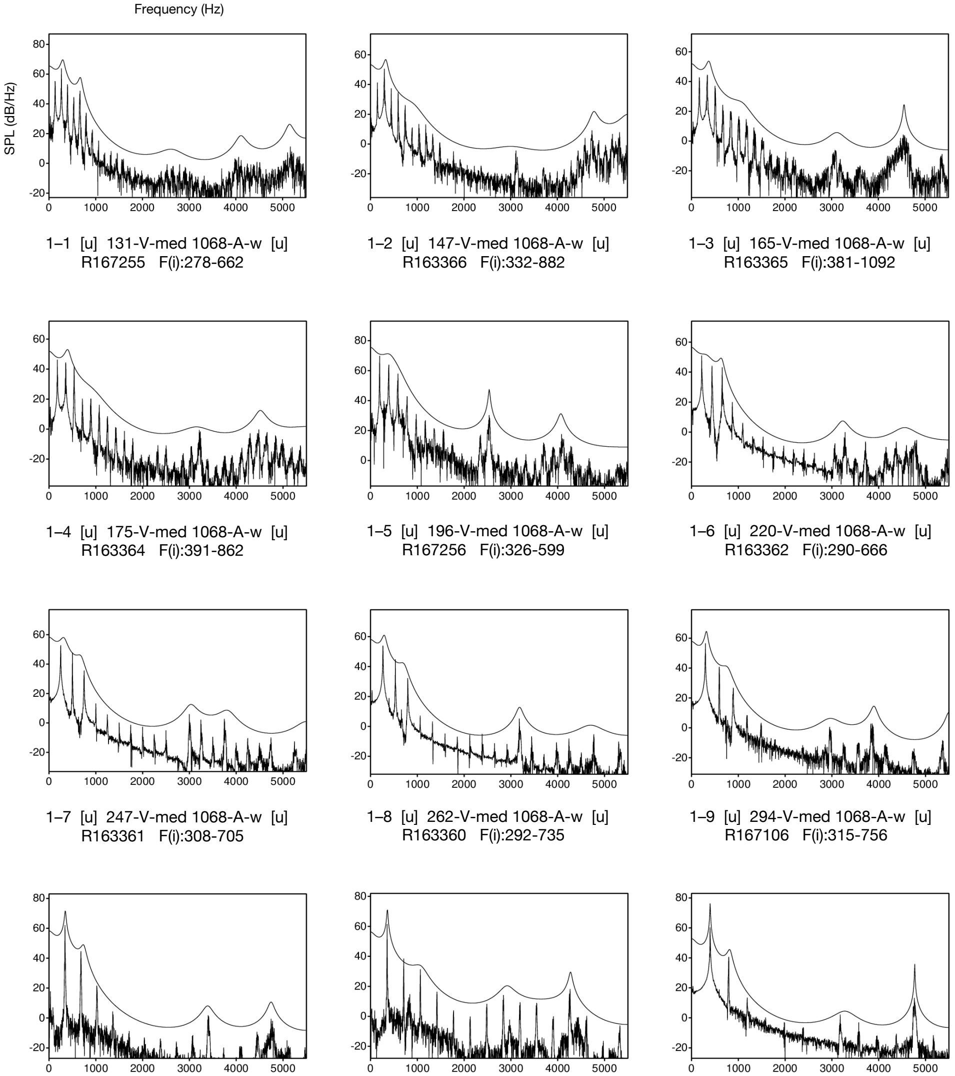
For references, extended background information, details of experimental design, method and results, an extended discussion and the complete documentation of the vocalises (tables including sound links), see the Materials, Chapter M2.1.

Figure 1. Voiced sounds of /u/ produced by a woman at different f_o levels. Extract of Chapter M2.1, Table 1 (see Series 11 in this table). Voiced sounds produced with medium vocal effort and stepwise f_o variation according to the musical C-major scale within a frequency range of intended f_o = 131–784 Hz are shown, for which a vowel recognition rate of 100% according to vowel intention was obtained. Intended f_o levels are given (for calculated levels, see the sounds in the Zurich Corpus). For the vocalises of this vowel, note that f_o/H_1 of the last sound of the series surpasses the second spectral peak of the first sound of the series.
[C-02-01-F01] ↗

Figure 2. Voiced sounds of /a/ produced by a woman at different f_o levels. Extract of Chapter M2.1, Table 1 (see Series 16 in this table). Voiced sounds produced with production parameters and recognised with a recognition rate as described for the sounds in Figure 1 are shown.
[C-02-01-F02] ↗

Figure 3. Voiced sounds of /ø/ produced by a woman at different f_o levels. Extract of Chapter M2.1, Table 1 (see Series 13 in this table). Voiced sounds produced with production parameters as described for the sounds in Figure 1 are shown. Vowel recognition rates obtained were 80–100%.
[C-02-01-F03] ↗

Figure 1. Voiced sounds of /u/ produced by a woman at different fo levels. [C-02-01-F01]



1–1 [u] 131-V-med 1068-A-w [u]
R167255 F(i):278-662

1–2 [u] 147-V-med 1068-A-w [u]
R163366 F(i):332-882

1–3 [u] 165-V-med 1068-A-w [u]
R163365 F(i):381-1092

1–4 [u] 175-V-med 1068-A-w [u]
R163364 F(i):391-862

1–5 [u] 196-V-med 1068-A-w [u]
R167256 F(i):326-599

1–6 [u] 220-V-med 1068-A-w [u]
R163362 F(i):290-666

1–7 [u] 247-V-med 1068-A-w [u]
R163361 F(i):308-705

1–8 [u] 262-V-med 1068-A-w [u]
R163360 F(i):292-735

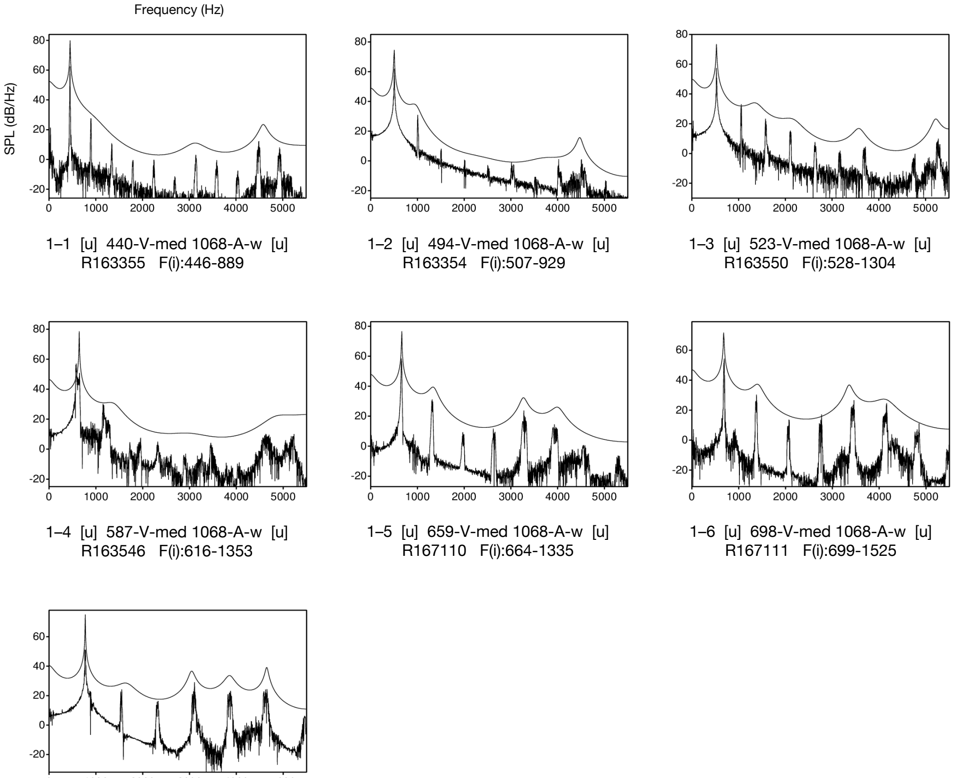
1–9 [u] 294-V-med 1068-A-w [u]
R167106 F(i):315-756

1–10 [u] 330-V-med 1068-A-w [u]
R163358 F(i):344-713

1–11 [u] 349-V-med 1068-A-w [u]
R167109 F(i):364-1002

1–12 [u] 392-V-med 1068-A-w [u]
R163356 F(i):393-804

Figure 1 (continuation). [C-02-01-F01]



1–1 [u] 440-V-med 1068-A-w [u]
R163355 F(i):446-889

1–2 [u] 494-V-med 1068-A-w [u]
R163354 F(i):507-929

1–3 [u] 523-V-med 1068-A-w [u]
R163550 F(i):528-1304

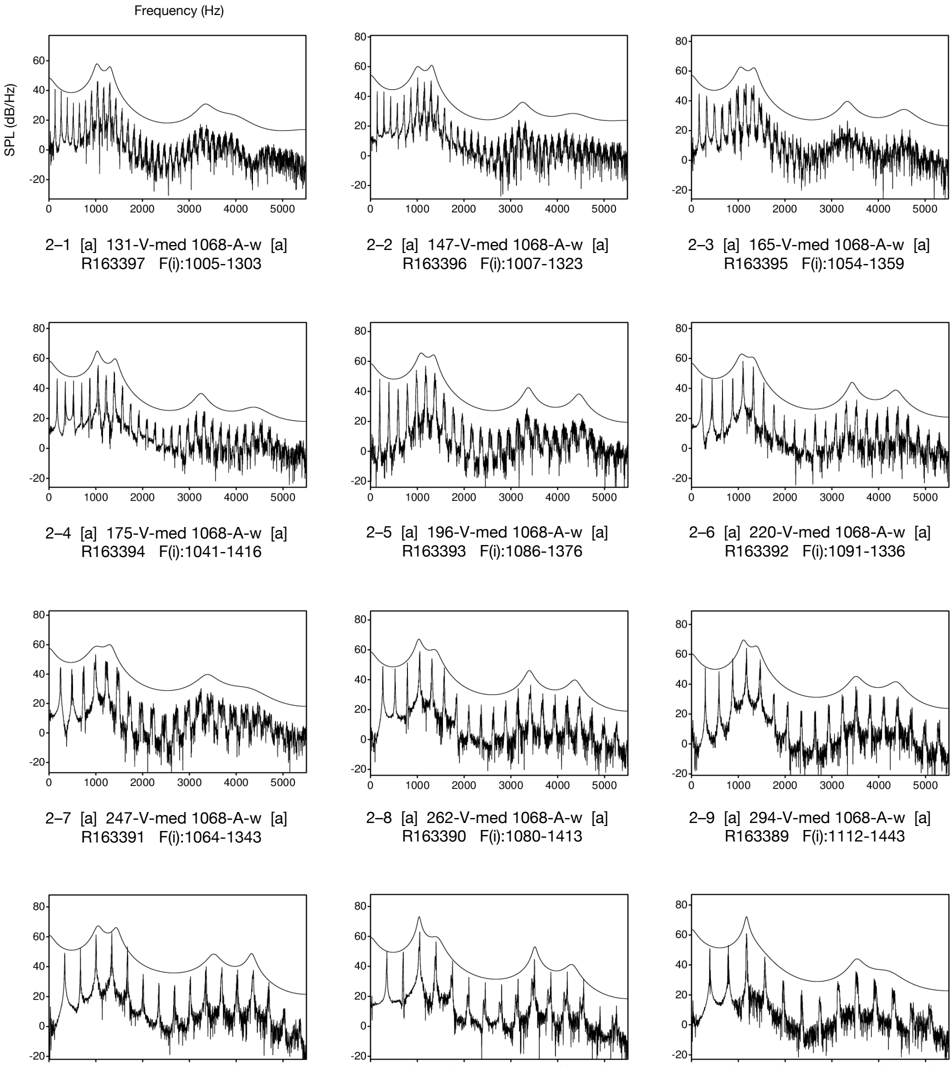
1–4 [u] 587-V-med 1068-A-w [u]
R163546 F(i):616-1353

1–5 [u] 659-V-med 1068-A-w [u]
R167110 F(i):664-1335

1–6 [u] 698-V-med 1068-A-w [u]
R167111 F(i):699-1525

1–7 [u] 784-V-med 1068-A-w [u]
R167112 F(i):767-1632

Figure 2. Voiced sounds of /a/ produced by a woman at different fo levels.
[C-02-01-F02]



2–1 [a] 131-V-med 1068-A-w [a]
R163397 F(i):1005-1303

2–2 [a] 147-V-med 1068-A-w [a]
R163396 F(i):1007-1323

2–3 [a] 165-V-med 1068-A-w [a]
R163395 F(i):1054-1359

2–4 [a] 175-V-med 1068-A-w [a]
R163394 F(i):1041-1416

2–5 [a] 196-V-med 1068-A-w [a]
R163393 F(i):1086-1376

2–6 [a] 220-V-med 1068-A-w [a]
R163392 F(i):1091-1336

2–7 [a] 247-V-med 1068-A-w [a]
R163391 F(i):1064-1343

2–8 [a] 262-V-med 1068-A-w [a]
R163390 F(i):1080-1413

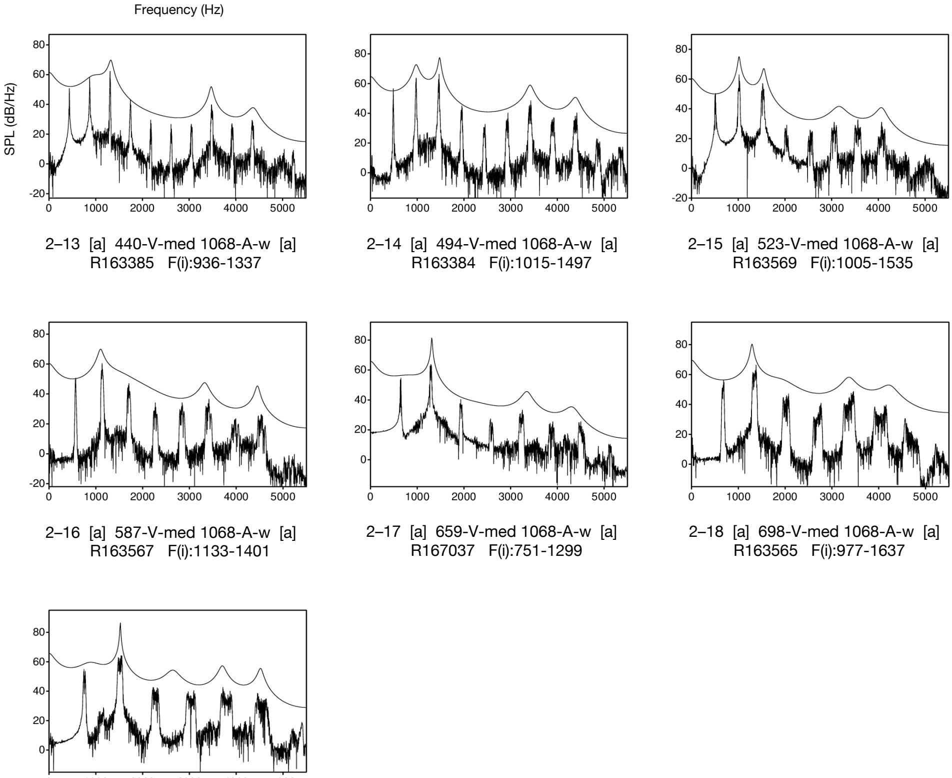
2–9 [a] 294-V-med 1068-A-w [a]
R163389 F(i):1112-1443

2–10 [a] 330-V-med 1068-A-w [a]
R163388 F(i):1117-1486

2–11 [a] 349-V-med 1068-A-w [a]
R163387 F(i):1052-1472

2–12 [a] 392-V-med 1068-A-w [a]
R163386 F(i):1179-1483

Figure 2 (continuation). [C-02-01-F02]



2–13 [a] 440-V-med 1068-A-w [a]
R163385 F(i):936-1337

2–14 [a] 494-V-med 1068-A-w [a]
R163384 F(i):1015-1497

2–15 [a] 523-V-med 1068-A-w [a]
R163569 F(i):1005-1535

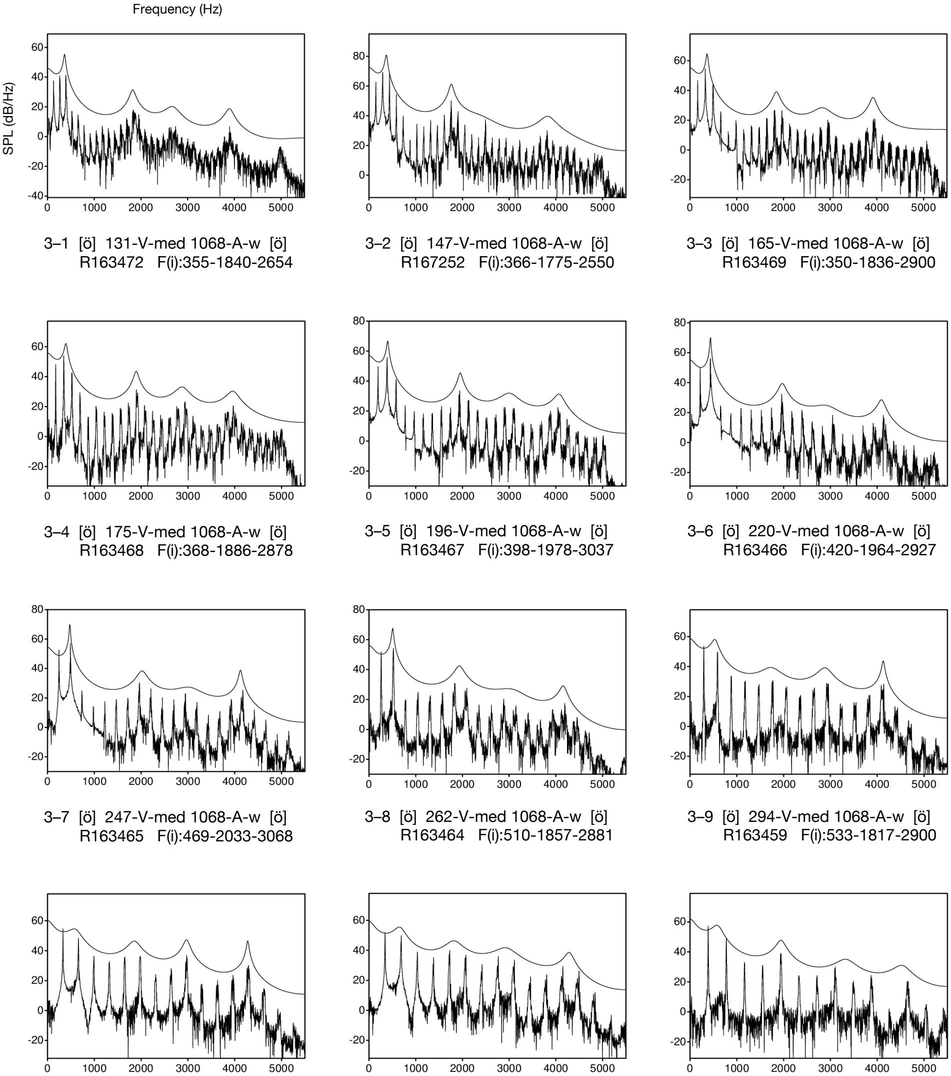
2–16 [a] 587-V-med 1068-A-w [a]
R163567 F(i):1133-1401

2–17 [a] 659-V-med 1068-A-w [a]
R167037 F(i):751-1299

2–18 [a] 698-V-med 1068-A-w [a]
R163565 F(i):977-1637

2–19 [a] 784-V-med 1068-A-w [a]
R163564 F(i):909-1528

Figure 3. Voiced sounds of /ø/ produced by a woman at different fo levels.
[C-02-01-F03]



3–1 [ö] 131-V-med 1068-A-w [ö]
R163472 F(i):355-1840-2654

3–2 [ö] 147-V-med 1068-A-w [ö]
R167252 F(i):366-1775-2550

3–3 [ö] 165-V-med 1068-A-w [ö]
R163469 F(i):350-1836-2900

3–4 [ö] 175-V-med 1068-A-w [ö]
R163468 F(i):368-1886-2878

3–5 [ö] 196-V-med 1068-A-w [ö]
R163467 F(i):398-1978-3037

3–6 [ö] 220-V-med 1068-A-w [ö]
R163466 F(i):420-1964-2927

3–7 [ö] 247-V-med 1068-A-w [ö]
R163465 F(i):469-2033-3068

3–8 [ö] 262-V-med 1068-A-w [ö]
R163464 F(i):510-1857-2881

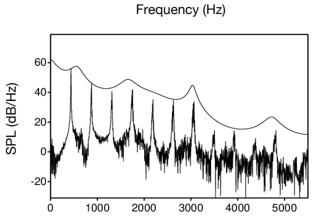
3–9 [ö] 294-V-med 1068-A-w [ö]
R163459 F(i):533-1817-2900

3–10 [ö] 330-V-med 1068-A-w [ö]
R163458 F(i):570-1864-2997

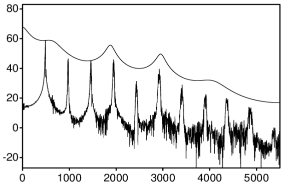
3–11 [ö] 349-V-med 1068-A-w [ö]
R163457 F(i):638-1868-2967

3–12 [ö] 392-V-med 1068-A-w [ö]
R163456 F(i):555-1925-2904

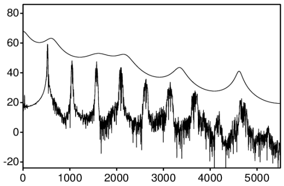
Figure 3 (continuation). [C-02-01-F03]



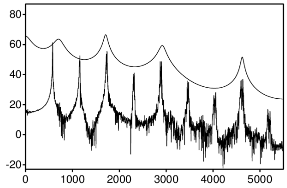
3–13 [ö] 440-V-med 1068-A-w [ö]
R163455 F(i):572-1752-2237



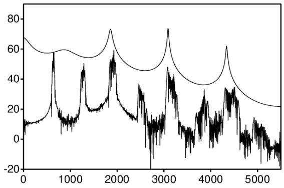
3–14 [ö] 494-V-med 1068-A-w [ö]
R163454 F(i):572-1827-2711



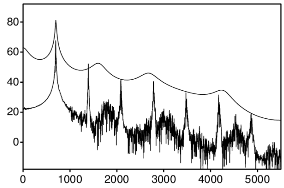
3–15 [ö] 523-V-med 1068-A-w [ö]
R163623 F(i):600-1688-2372



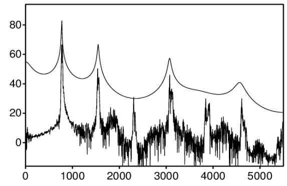
3–16 [ö] 587-V-med 1068-A-w [ö]
R163615 F(i):670-1739-2925



3–17 [ö] 659-V-med 1068-A-w [ö]
R163616 F(i):797-1615-2012



3–18 [ö] 698-V-med 1068-A-w [ö]
R167101 F(i):700-1613-2693



3–19 [ö] 784-V-med 1068-A-w [ö]
R167102 F(i):774-1558-3106

2.2 Isolated Vowel Sounds Produced at High Levels of f_o

In the literature, the upper-frequency limit of f_o for vowel recognition is a matter of debate. No consensus has been established as to whether or not, for recognition, vowel-specific statistical F_1, mode or style of sound production and vowel context play a substantial role for sounds at high levels of f_o. However, several studies reported possible vowel recognition for natural sounds (both in and out of consonantal context) produced within an f_o range of approximately 660 Hz to 1 kHz, depending on conditions of vowel production and related listening tests.

The documentation of the vocalises of the three speakers in the previous chapter has already confirmed possible vowel recognition up to an f_o level of 700–800 Hz for sounds of all long Standard German vowels produced in V context. Further examining the entire sample of the Zurich Corpus with regard to all eight long Standard German vowels, we found numerous sounds in the f_o range of 700–800 Hz, produced in V context by ten or more different speakers, that reached a recognition rate of 100% (5/5 listeners) according to vowel intention. The same holds true for the sounds of /i, y, a, u/ up to an f_o range of 950–1100 Hz. However, the listening test conducted when creating the Zurich Corpus was based on the entire sounds, and the sounds of single speakers were tested separately (speaker-blocked test condition), further separating nonstyle and style productions. (Sounds produced in V and sVsV conditions were tested together.) The sound series of the previous chapter have to be considered in this context.

In view of the foregoing, in two experiments, we addressed the question of whether or not successful vowel recognition of entire sounds produced at high f_o, as found for numerous sounds of the Zurich Corpus, could be confirmed if only the respective sound nuclei were investigated (excluding on- and offsets) and if sounds of different speakers were mixed in the listening tests conducted. For each of the eight long Standard German vowels and based on the sounds of the Zurich Corpus, in the first experiment, 20 sounds produced by ten or more speakers in V context at calculated f_o levels in the range of c. 700–800 Hz and with a 100% recognition rate matching vowel intention in the standard listening test when creating the corpus were selected by the author (best sound and vowel quality), resulting in a sample of 160 sounds in total. Production style and vocal effort of the sounds were ignored in this experiment. Depending on the duration of the sounds, middle sound nuclei with no on- and offsets and with a duration of 1 sec. at maximum were extracted, and a new listening test with the five standard listeners of the Zurich Corpus was performed. In the second experiment, the same procedure was applied for sounds of the vowels /i, y, a, u/ produced in V context at f_o in the range of c. 950–1100 Hz, resulting in a sample of 80 sounds in total.

According to the results of the first experiment, the vowel recognition rate for the sound nuclei of the close vowels /i, y, u/, as well as for the open vowel /a/, proved to be 100% in most cases, equal to the recognition rate of the original sounds including of on- and offsets. For the sounds of the open-mid vowel /ɛ/, in 18 of 20 cases, vowel recognition could either be maintained at 100% or dropped slightly to 80%. Conversely, for the sounds of the close-mid vowels, vowel recognition proved to be substantially impaired or even confused when comparing sound nuclei against original sounds, especially for sounds of /ø, o/: Only three sounds of /o/ were recognised with a rate of 80%, and only seven sounds of /ø/ were recognised with a rate of 100% or 80%. However, despite this impairment, the results indicate that speakers with excellent vocal abilities can produce sounds of all long Standard German vowels at f_o levels in the range of 700–800 Hz in a way that listeners with experience in vowel recognition tests can differentiate and recognise them based on isolated sound nuclei.

According to the results of the second experiment, the vowel recognition rate for the sound nuclei proved to be 80–100% for the majority of the close vowels, that is, near or equal to the recognition rate of the original sounds including on- and offsets. In contrast, for the open vowel /a/, the recognition rate for the sound nuclei dropped substantially. However, investigating the confusion matrix for these sounds in detail, most of the sounds were assigned to either /a/ or to an open-mid vowel (/ɛ/ or /ɔ/), and with one exception, no confusion with a close vowel occurred. In these terms, the results indicated that speakers with excellent vocal abilities can produce sounds of the long Standard German corner vowels and the intermediate vowel /y/ in between /i/ and /u/ at f_o levels of c. 1 kHz in a way that listeners with experience in vowel recognition tests can differentiate and recognise them based on sound nuclei only, above all in terms of differentiation of corner positions close versus open and front versus back, including rounded versus unrounded.

In conclusion, the two experiments and their results confirmed that vowel recognition for entire sounds produced in V context (including on- and offsets), as documented in the Zurich Corpus, can also be demonstrated for a substantial part of the respective sound nuclei (excluding on- and offsets) for all long Standard German vowels in the f_o range of c. 700–800 Hz and for the vowels /i, y, a, u/ at an f_o of approximately 1 kHz. Note again that the three vowels /i, a, u/ represent the corner vowels of the vowel triangle, and the vowel /y/ represents an intermediate peripheral vowel between /i/ and /u/, indicating that vowel recognition on very high levels of f_o may tend to identify vowel qualities forming strong oppositions (corner positions, open versus close and front versus back, the latter including unrounded versus rounded).

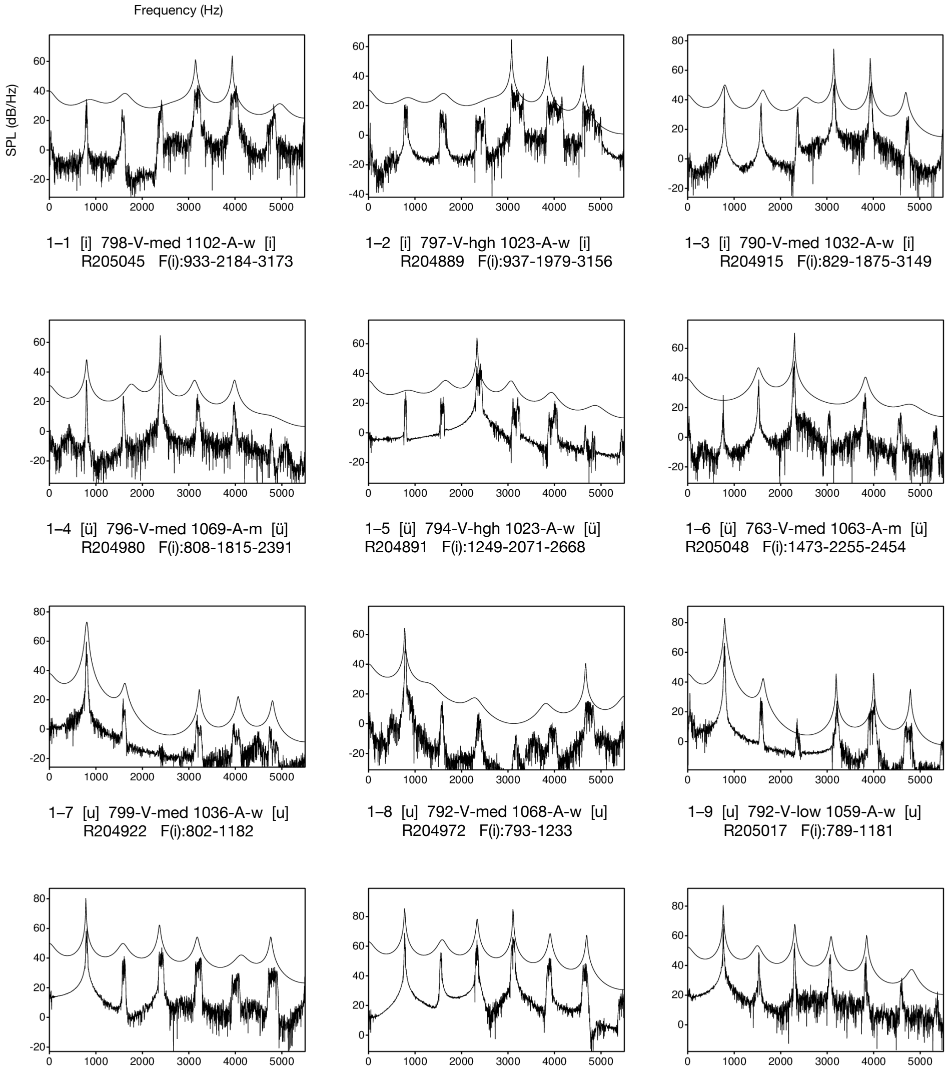
As an exemplary illustration, for each of the eight vowels investigated, Figure 1 shows three selected spectra of three nuclei of recognised sounds produced within a range of calculated f_o of c. 700–800 Hz. For the vowels /i, y, u, a/, Figure 2 shows three selected spectra of three nuclei of recognised sounds produced at an f_o of c. 1 kHz.

For references, extended background information, details of experimental design, method and results, an extended discussion and documentation of results (tables including sound links), see the Materials, Chapter M2.2.

Figure 1. Recognisable sound nuclei of natural sounds of the vowels /i, y, u/, /e, ø, o/ and /ɛ, a/ produced by children, women and men at calculated f_o in the frequency range of c. 700–800 Hz. Extract of Chapter M2.2, Table 2 (for a comparison of the original sounds and the sound nuclei, see this table). For each single vowel, the spectra of three sound nuclei are shown. Vowel recognition rates = 100% for the sounds of /i, y, e, ɛ, a, u/, 80–100% for the sounds of /ø/ and 80% for the sounds of /o/.
[C-02-02-F01] ↗

Figure 2: Recognisable sound nuclei of sounds of the vowels /i, y, u, a/ produced by women at calculated f_o of c. 1 kHz. Extract of Chapter M2.2, Table 5 (for a comparison of the original sounds and the sound nuclei, see this table). For each single vowel, the spectra of three sound nuclei are shown. Vowel recognition rates = 100% for the sounds of /i, y, u/ and 80–100% for the sounds of /a/.
[C-02-02-F02] ↗

Figure 1. Recognisable sound nuclei of natural sounds of the vowels /i, y, u/, /e, ø, o/ and /ɛ, a/ produced by children, women and men at calculated fo in the frequency range of c. 700–800 Hz. [C-02-02-F01]



1–1 [i] 798-V-med 1102-A-w [i]
R205045 F(i):933-2184-3173

1–2 [i] 797-V-hgh 1023-A-w [i]
R204889 F(i):937-1979-3156

1–3 [i] 790-V-med 1032-A-w [i]
R204915 F(i):829-1875-3149

1–4 [ü] 796-V-med 1069-A-m [ü]
R204980 F(i):808-1815-2391

1–5 [ü] 794-V-hgh 1023-A-w [ü]
R204891 F(i):1249-2071-2668

1–6 [ü] 763-V-med 1063-A-m [ü]
R205048 F(i):1473-2255-2454

1–7 [u] 799-V-med 1036-A-w [u]
R204922 F(i):802-1182

1–8 [u] 792-V-med 1068-A-w [u]
R204972 F(i):793-1233

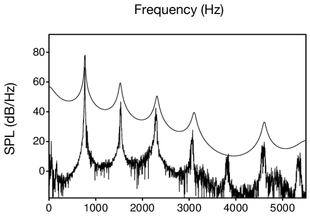
1–9 [u] 792-V-low 1059-A-w [u]
R205017 F(i):789-1181

1–10 [e] 801-V-hgh 1023-A-w [e]
R205015 F(i):825-2196-2885

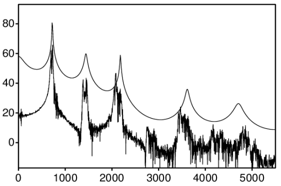
1–11 [e] 778-V-hgh 1036-A-w [e]
R204995 F(i):848-2392-3015

1–12 [e] 765-V-low 1001-A-w [e]
R205040 F(i):780-2232-2995

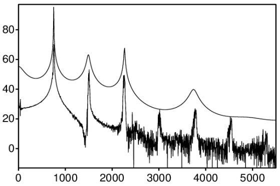
Figure 1 (continuation). [C-02-02-F01]



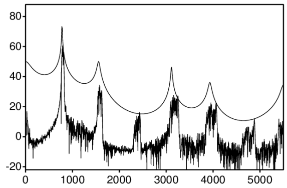
1–13 [ö] 766-V-med 1034-C-w [ö]
R205032 F(i):766-1817-2522



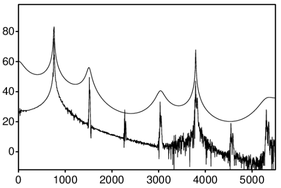
1–14 [ö] 715-V-low 1056-C-m [ö]
R204935 F(i):733-1945-3017



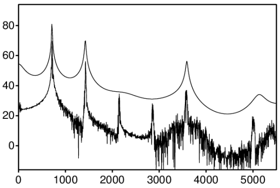
1–15 [ö] 753-V-med 1037-C-w [ö
] R204905 F(i):750-2036-2549



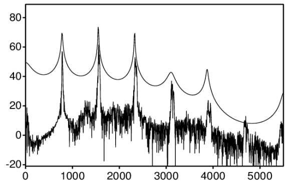
1–16 [o] 792-V-med 1070-A-m [o]
R205022 F(i):780-1203



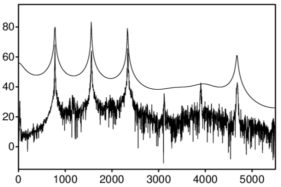
1–17 [o] 759-V-low 1046-A-w [o]
R204953 F(i):757-1442



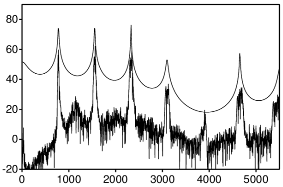
1–18 [o] 715-V-med 1069-A-m [o]
R205004 F(i):715-1428



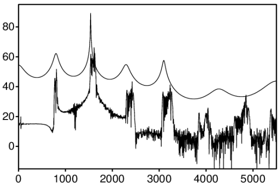
1–19 [ä] 781-V-hgh 1034-C-w [ä]
R204902 F(i):1367-2258-3331



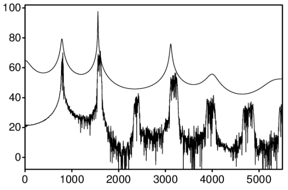
1–20 [ä] 780-V-med 1069-A-m [ä]
R205008 F(i):782-1564-2343



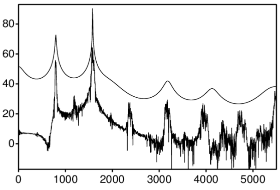
1–21 [ä] 777-V-hgh 1052-A-w [ä]
R204957 F(i):835-1547-2415



1–22 [a] 796-V-hgh 1023-A-w [a]
R204896 F(i):912-1590

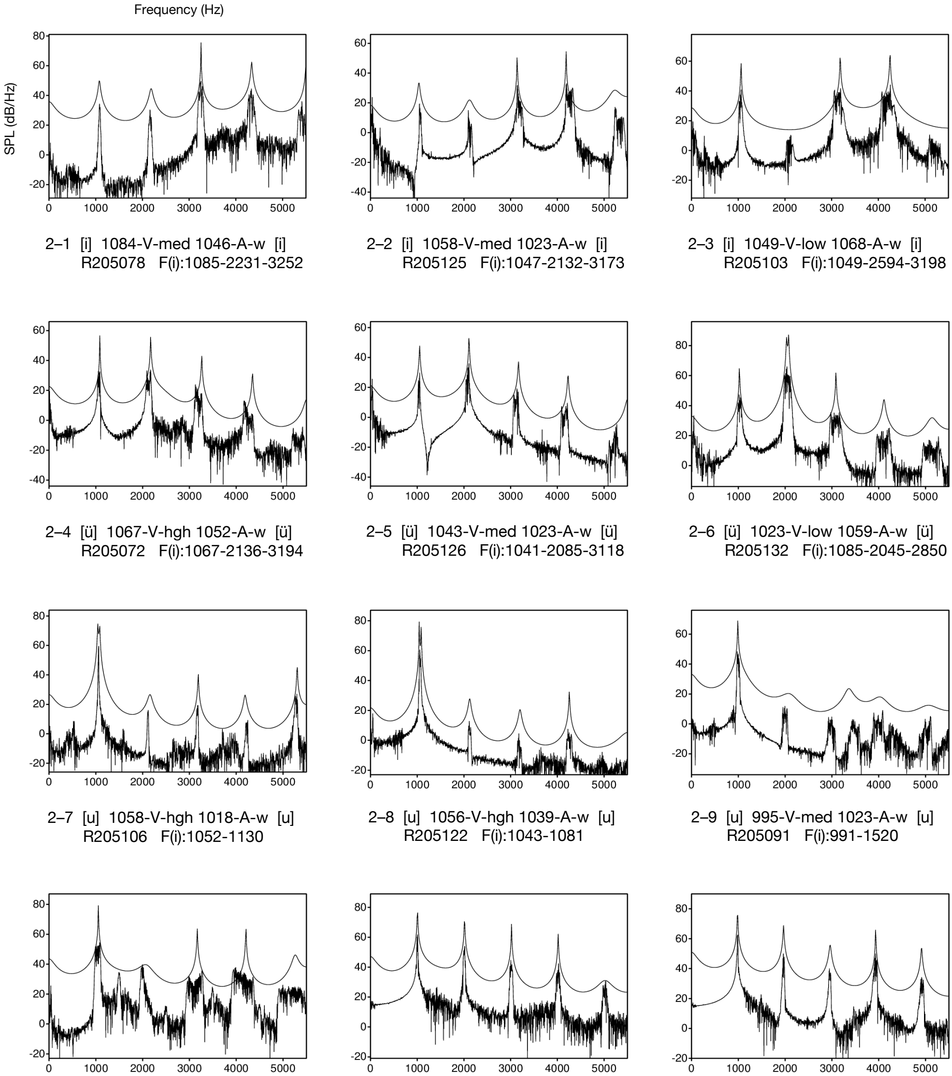


1–23 [a] 796-V-hgh 1050-A-m [a]
R204913 F(i):798-1591



1–24 [a] 791-V-hgh 1052-A-w [a]
R204926 F(i):823-1577

Figure 2. Recognisable sound nuclei of sounds of the vowels /i, y, u, a/ produced by women at calcutaed fo of c. 1 kHz. [C-02-02-F02]



2–1 [i] 1084-V-med 1046-A-w [i]
R205078 F(i):1085-2231-3252

2–2 [i] 1058-V-med 1023-A-w [i]
R205125 F(i):1047-2132-3173

2–3 [i] 1049-V-low 1068-A-w [i]
R205103 F(i):1049-2594-3198

2–4 [ü] 1067-V-hgh 1052-A-w [ü]
R205072 F(i):1067-2136-3194

2–5 [ü] 1043-V-med 1023-A-w [ü]
R205126 F(i):1041-2085-3118

2–6 [ü] 1023-V-low 1059-A-w [ü]
R205132 F(i):1085-2045-2850

2–7 [u] 1058-V-hgh 1018-A-w [u]
R205106 F(i):1052-1130

2–8 [u] 1056-V-hgh 1039-A-w [u]
R205122 F(i):1043-1081

2–9 [u] 995-V-med 1023-A-w [u]
R205091 F(i):991-1520

2–10 [a] 1035-V-med 1023-A-w [a]
R205127 F(i):1036-1743

2–11 [a] 1004-V-low 1046-A-w [a]
R205100 F(i):1001-2000

2–12 [a] 985-V-med 1023-A-w [a]
R205093 F(i):985-1954

2.3 Minimal Pairs Produced at High Levels of f_o

As indicated in Chapter 1.1, during the creation of the Zurich Corpus (see Part 3 of the corpus), minimal pairs produced at various levels of f_o by selected speakers with excellent vocal abilities were recorded according to three different experimental settings, and vowel recognition was tested in order to investigate the recognition of vowel quality in minimal pair context for sounds produced up to high f_o levels. These three experiments are described and discussed in this chapter, with a focus on the recognition results for utterances produced at intended f_o of 784, 880 and 1047 Hz.

In the first setting, 18 vowel contrasts for 18 minimal pairs (words; for details, see Chapter M2.3) were investigated, each minimal pair produced as a word pair in one utterance by a woman (professional musical theatre singer and actress) at nine different f_o levels of 220–440–587–659–698–740–784–831–880 Hz. Single words and vowel sound nuclei of 250 ms were extracted for vowel recognition. Two recognition tests were conducted with single test items: entire single words or single extracted vowel sound nuclei. Twenty listeners (students of the University of Zurich) per test were asked to listen to a word or a vowel nucleus, respectively, and to assign it to one of two words of a minimal pair displayed on a screen, the word or vowel nucleus presented being extracted from the production of that minimal pair.

According to the results of the listening tests of this first experiment, for the word condition, vowel contrast recognition for the minimal pairs was generally maintained up to an intended f_o of 880 Hz with a rate of ≥ 90%, except for the contrasts of /e/–/ø/ and /e/–/ɛ/, which had a recognition rate of 82–88%. For the sound nucleus condition, the vowel contrast recognition rate dropped for some of the minimal pairs; however, for the sounds at an f_o of 784 Hz, recognition was maintained at ≥ 80%, except for the contrasts /y/–/ø/ (78%) and /e/–/ø/ (72%), and the same held true for the sounds at an f_o of 880 Hz, except for the contrasts /e/–/ɛ/ (72%), /ø/–/ɛ/ (78%) and /ɛ/–/a/ (63%).

In the second setting, the vowel recognition of seven versions of German “lVgen” with all long Standard German vowels except /u/ was investigated, the words produced by the same woman as minimal pairs for all possible vowel contrasts (two words in one recording) of all front vowels and /a/ and for the contrast of /a/ and the back vowel /o/, at nine f_o levels of 220–440–587–659–698–740–784–831–880 Hz. For each vowel and each level of f_o, the word with the most recognisable vowel quality (as rated by two of the investigators) was selected, and vowel recognition was tested for the resulting sample: In a first recognition test involving 28 listeners (students of the University of Zurich), single words were presented in random order, and the listeners were asked to label one of the seven "lVgen" variants displayed on a screen. In a second recognition test involving the five standard listeners of the Zurich Corpus, single words were presented in random order, and the listeners were asked to label the recognised vowel quality according to the standard procedure of the corpus.

According to the results of the listening test for this second experiment involving 28 inexperienced listeners, for utterances produced at an intended f_o of 784 Hz, the vowel recognition rates were > 90% for /i, y, ø, a, o/, 57% for /ɛ/ and 53% for /e/. For utterances produced at an intended f_o of 880 Hz, the vowel recognition rates were > 95% for /i, y, a, o/, 86% for /ɛ/, 64% for /ø/ and 50% for /e/. According to the results of the listening test involving the five experienced standard listeners of the Zurich Corpus, vowel recognition rates of the sounds were 100% up to an intended f_o of 880 Hz for /i, y, e, ɛ, a, o/ and 80% for /ø/. In these terms, the results of the second experiment indicated that sounds of long vowels produced in minimal pair context could be recognised up to f_o of 880 Hz, with a higher recognition rate for experienced listeners compared with inexperienced listeners and, by tendency, with a more robust recognition of the close and open vowels compared with close-mid and open-mid vowels.

In an additional third experiment, vowel recognition of (i) other recordings of German "lVgen" words with all long Standard German vowels including /u/, (ii) three versions of "bVden" with the vowels /a–o–u/ and (iii) three versions of "schVf" with the corner vowels /i–a–u/ were investigated, the corresponding utterances produced as single words by two women (professional musical theatre singers and actresses) at various intended f_o (according to the musical C-major scale) within a frequency range of 523–1047 Hz. Vowel recognition of entire words was tested according to the standard procedure of the Zurich Corpus and involving the standard listeners of the corpus. Based on the recognition rates obtained, for each of the three sets of minimal pairs, each vowel and each of the two f_o levels of 880 Hz and 1047 Hz, the utterance with the highest recognition rate (vowel recognition matching vowel intention) in terms of "best" cases was selected. These cases demonstrate that a 100% vowel recognition rate could be maintained up to an intended f_o of 880 Hz. This also held true for at least one word per vowel and a recognition rate of 80–100% up to an intended f_o of 1047 Hz.

In conclusion, the three experiments showed that minimal pairs with long vowels could be successfully differentiated in the range of intended f_o of 784–880 Hz both in production and perception and if speakers are exceptional in their vocal abilities and listeners are very experienced in vowel recognition tasks, this may be possible up to an intended f_o level of 1047 Hz, not only for the corner vowels but also for other vowels or vowel contrasts. Thus, the phonological function of vowels in a minimal pair context can be maintained at f_o levels up to c. 1 kHz.

Depending on vowel qualities, the recognition rate for sound nuclei dropped somewhat in experiment 1 compared to the recognition rate obtained for the respective entire sounds. However, the results indicated that isolated sound nuclei of long vowels produced in the context of minimal pairs could be recognised above chance level at an f_o of 880 Hz.

For exemplary documentation, Figures 1 to 3 show the selected “best” cases of utterances investigated in the third experiment.

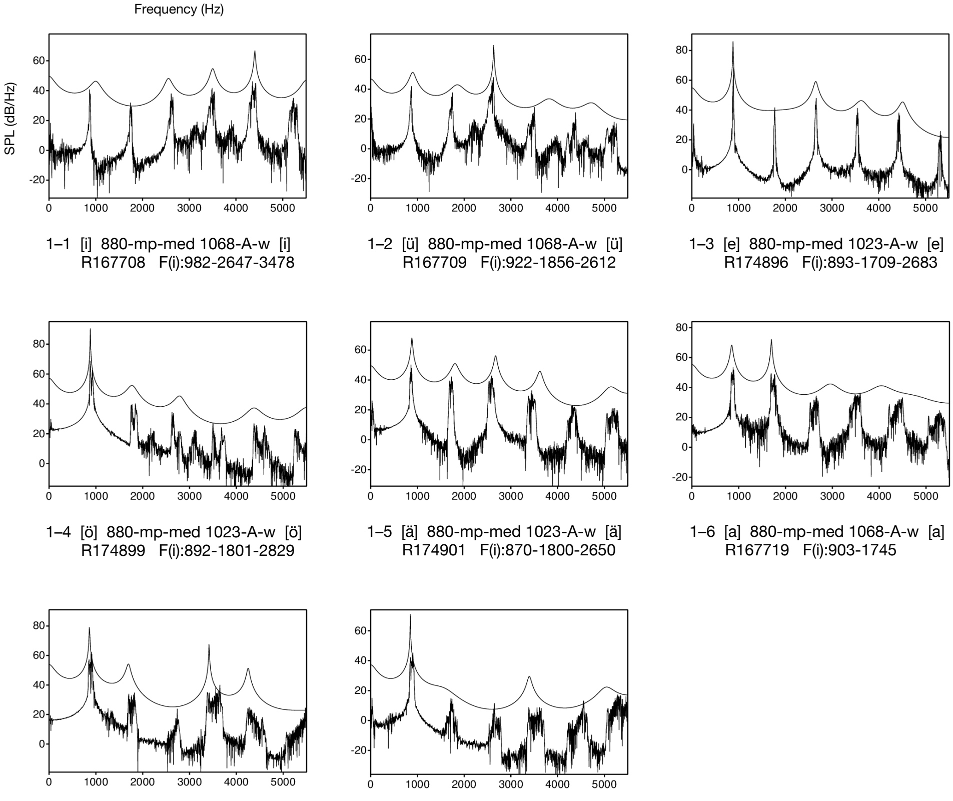
For references, extended background information, details of experimental design, method and results, an extended discussion and documentation of results (tables including sound links), see the Materials, Chapter M2.3.

Figure 1: Recognisable German “lVgen” minimal pairs produced by two women at an intended f_o of 880 Hz related to the vowels /i, y, e, ø, ɛ, a, o, u/ (selected “best” cases) and respective spectra of the vowel sound nuclei. Extract of Chapter M2.3, Table 3 (for a comparison of the original sounds and the sound nuclei, see this table). The vowel recognition rate obtained was 100% for all utterances shown (entire utterances, experienced listeners).
[C-02-03-F01]

Figure 2: Recognisable German “lVgen” minimal pairs produced by two women at an intended f_o of 1047 Hz related to the vowels /i, y, e, ø, ɛ, a, o, u/ (selected “best” cases) and respective spectra of the vowel sound nuclei. Extract of Chapter M2.3, Table 3 (for a comparison of the original sounds and the sound nuclei, see this table). Vowel recognition rates obtained were 100% for the utterances related to /i, y, ø, ɛ, o, u/, 80% for the utterance related to /e/ and 60% for the utterance related to /a/ (entire utterances, experienced listeners).
[C-02-03-F02]

Figure 3: Recognisable German “BVden” and “schVf” minimal pairs produced by two women at intended f_o of 880 and 1047 Hz related to the vowels /a, o, u/ and /i, a, u/ (selected “best” cases) and respective spectra of the vowel sound nuclei. Extract of Chapter M2.3, Table 3 (for a comparison of the original sounds and the sound nuclei, see this table). Sounds 1–6 show the utterances of “bVden”, and sounds 7–12 show the corresponding spectra of “schVf”. Vowel recognition rates obtained were 100% for all sounds related to /i, a, u/ and 100% and 80% for the two utterances related to /o/ (entire utterances, experienced listeners).
[C-02-03-F03]

Figure 1. Recognisable German “lVgen” minimal pairs produced by two women at an intended fo of 880 Hz related to the vowels /i, y, e, ø, ɛ, a, o, u/ (selected “best” cases) and respective spectra of the vowel sound nuclei. [C-02-03-F01]



1–1 [i] 880-mp-med 1068-A-w [i]
R167708 F(i):982-2647-3478

1–2 [ü] 880-mp-med 1068-A-w [ü]
R167709 F(i):922-1856-2612

1–3 [e] 880-mp-med 1023-A-w [e]
R174896 F(i):893-1709-2683

1–4 [ö] 880-mp-med 1023-A-w [ö]
R174899 F(i):892-1801-2829

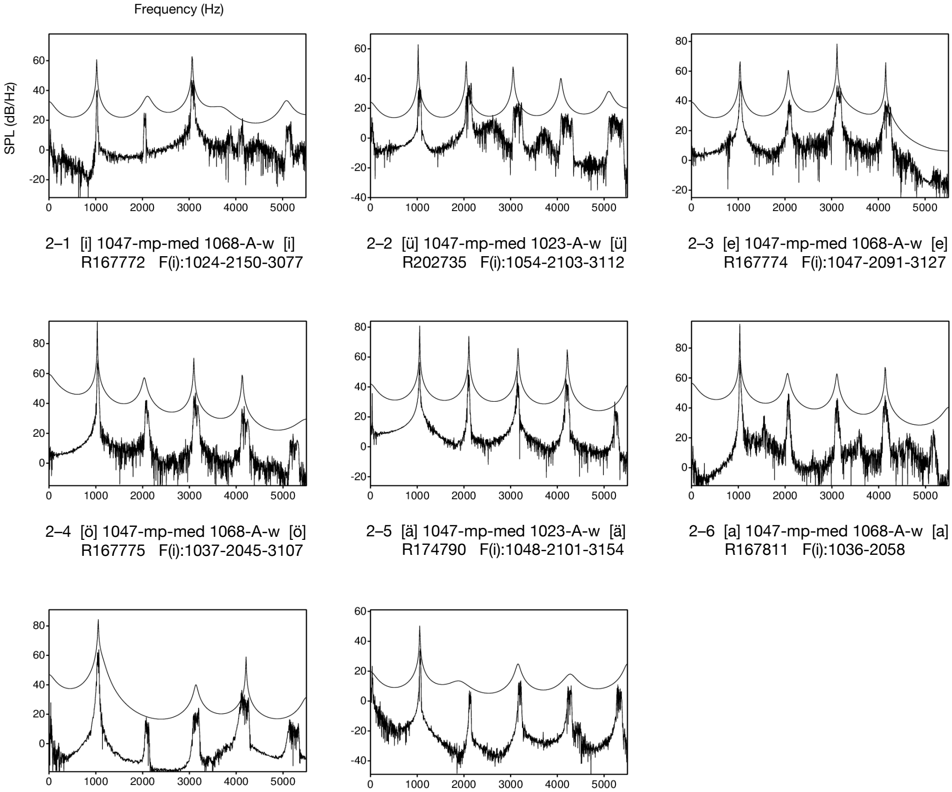
1–5 [ä] 880-mp-med 1023-A-w [ä]
R174901 F(i):870-1800-2650

1–6 [a] 880-mp-med 1068-A-w [a]
R167719 F(i):903-1745

1–7 [o] 880-mp-med 1023-A-w [o]
R111884 F(i):894-1759

1–8 [u] 880-mp-med 1023-A-w [u]
R113012 F(i):888-1606

Figure 2. Recognisable German “lVgen” minimal pairs produced by two women at an intended fo of 1047 Hz related to the vowels /i, y, e, ø, ɛ, a, o, u/ (selected “best” cases) and respective spectra of the vowel sound nuclei. [C-02-03-F02]



2–1 [i] 1047-mp-med 1068-A-w [i]
R167772 F(i):1024-2150-3077

2–2 [ü] 1047-mp-med 1023-A-w [ü]
R202735 F(i):1054-2103-3112

2–3 [e] 1047-mp-med 1068-A-w [e]
R167774 F(i):1047-2091-3127

2–4 [ö] 1047-mp-med 1068-A-w [ö]
R167775 F(i):1037-2045-3107

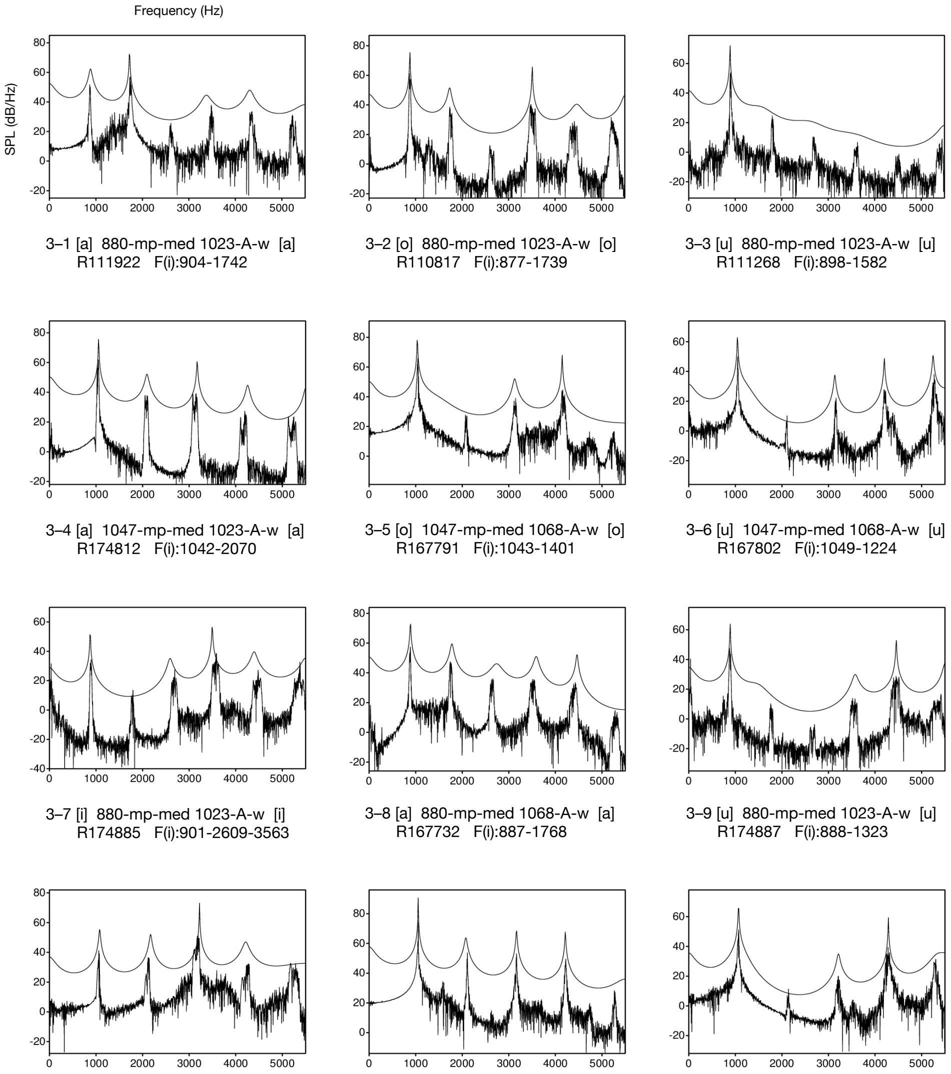
2–5 [ä] 1047-mp-med 1023-A-w [ä]
R174790 F(i):1048-2101-3154

2–6 [a] 1047-mp-med 1068-A-w [a]
R167811 F(i):1036-2058

2–7 [o] 1047-mp-med 1023-A-w [o]
R174796 F(i):1042-1093

2–8 [u] 1047-mp-med 1023-A-w [u]
R174798 F(i):1058-1893

Figure 3. Recognisable German “BVden” and “schVf” minimal pairs produced by two women at intended fo of 880 and 1047 Hz related to the vowels /a, o, u/ and /i, a, u/ (selected “best” cases) and respective spectra of the vowel sound nuclei. [C-02-03-F03]



3–1 [a] 880-mp-med 1023-A-w [a]
R111922 F(i):904-1742

3–2 [o] 880-mp-med 1023-A-w [o]
R110817 F(i):877-1739

3–3 [u] 880-mp-med 1023-A-w [u]
R111268 F(i):898-1582

3–4 [a] 1047-mp-med 1023-A-w [a]
R174812 F(i):1042-2070

3–5 [o] 1047-mp-med 1068-A-w [o]
R167791 F(i):1043-1401

3–6 [u] 1047-mp-med 1068-A-w [u]
R167802 F(i):1049-1224

3–7 [i] 880-mp-med 1023-A-w [i]
R174885 F(i):901-2609-3563

3–8 [a] 880-mp-med 1068-A-w [a]
R167732 F(i):887-1768

3–9 [u] 880-mp-med 1023-A-w [u]
R174887 F(i):888-1323

3–10 [i] 1047-mp-med 1068-A-w [i]
R167805 F(i):1066-2163-3192

3–11 [a] 1047-mp-med 1068-A-w [a]
R167808 F(i):1055-2084

3–12 [u] 1047-mp-med 1068-A-w [u]
R167807 F(i):1074-1200

2.4 Extensive Ranges of f_o Contours of Speech

In the literature, no frame of reference is given as a standard range regarding f_o variation for recognisable speech that has to be adhered to for speech acoustics in general (see also the Introduction). However, most frequency ranges reported concern measurements of citation-form words, read text, or so-called normal (sic!) or conversational speech. Thus, in most cases, the related frequency ranges are given for relaxed speech and the lower part of the actual vocal range of the speakers only. Besides, as mentioned above, speech with f_o levels in the middle part of the vocal range of a speaker is generally discussed as related to very specific speech characteristics, such as highly emotional, loud or infant-directed speech or speech within the field of the performing arts. Vowel sounds produced at higher f_o levels are understood as an aspect of singing. However, according to our estimate, there is no difference between the f_o ranges of recognisable speech and recognisable singing – above all not concerning the related upper f_o limit – , the distinction of normal or conversational versus highly specific speech is problematic, and the f_o ranges of speech as such, given in the literature, are in most cases too limited and too low.

To provide evidence for such an understanding and estimate, we have presented various examples of large f_o contours of intelligible speech in the Preliminaries (see pp. 170–182), which point to the actual frequency extension of the contours that can be observed for everyday speech and for the speech of performing artists. Yet, open-access playback for the sounds was not provided because of potential legal issues. Therefore, for the present treatise, the documentation was revised and further extended, including new examples of speech extracts with legal permit or consent for sound playback for all utterances: Based on the sounds and speakers documented in Part 2 of the Zurich Corpus (see Chapter 1.1), the sounds consisting of live recordings conducted by the author or of speech extracted from TV shows, online content or DVDs/CDs (different languages), one or several speech extracts per speaker were selected, primarily focussing on speech contours with upper f_o levels of 500 Hz and higher for women and 350 Hz and higher for men, respectively. However, some compilations of speech extracts of single speakers also include utterances produced at lower f_o levels so as to demonstrate both the upper f_o levels and the vocal range of the speaker in question. (For further details of the form of the extracts, see the Materials, Chapter M2.4.) Extracts of nonstyle speech were separated from extracts of speech produced in an artistic style. However, the main focus of the investigation concerned artists performing on stage, acting in film or doing voice-over work.

On this basis, for this treatise, 517 single speech extracts of 80 adult speakers (48 women and 32 men) illustrating f_o ranges for nonstyle speech and the speech of professional performing artists were compiled and are presented in the Materials (see Chapter M2.4). For nonstyle speech, the entire f_o range documented covers a frequency range of c. 125–1000 Hz for women and c. 100–600 Hz for men. For the speech of performing artists, the entire f_o range documented covers a frequency range of 110–1000 Hz for women and c. 90–850 Hz for men.

Figures 1 to 3 illustrate the phenomenon of such large f_o ranges observed for nonstyle speech and the speech of performing artists in terms of a few exemplary series of speech extracts. (As for the speech extracts presented in the figures, please note: The graphic illustration of the f_o contour does not always correspond to the actual f_o contour of the speech extract because of interfering noise or measurement problems. Please refer to the indications of perceived pitch ranges [author's estimate] as given in the Materials.) References (origins of the speech extracts) are provided online in the comment field of the sounds presented.

Against this background, we conclude that the range of f_o contours of intelligible speech corresponds to the frequency range of the f_o of recognisable vowel sounds, as discussed in the previous chapters, and that, therefore, an actual f_o range of speech of up to 800 Hz at a minimum has to be considered as a standard reference for an acoustic theory of speech sounds.

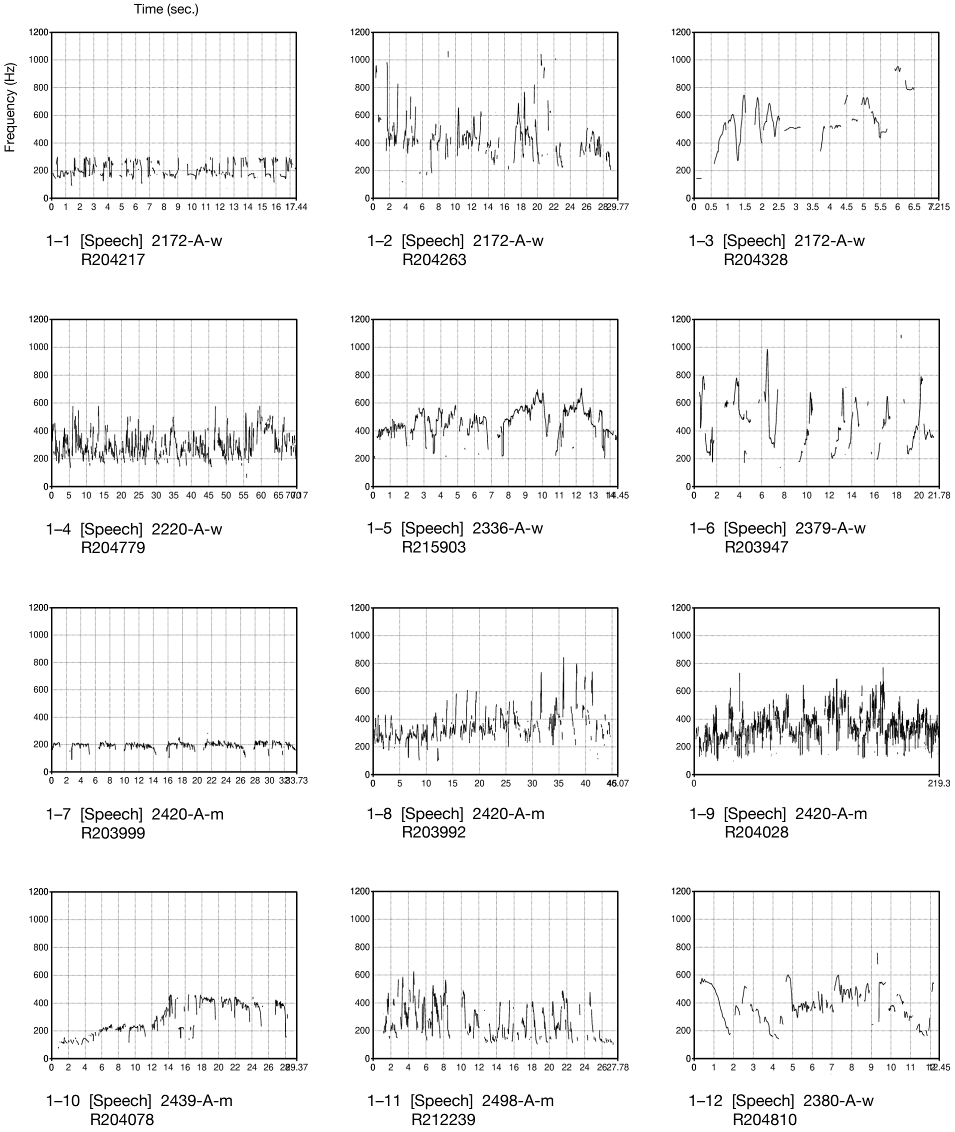
For references, extended background information, details of experimental design, method and results, an extended discussion and documentation of results (tables including sound links), see the Materials, Chapter M2.4.

Figure 1. Examples of intelligible speech produced by untrained speakers, illustrating extensive f_o contours in contexts of everyday life. Extract of Chapter M2.4, Table 1 (see Series 1, 2, 11, 16, 17, 23, 22 and 25 in this table). Sounds 1–3 = speech extracts of a woman selling grilled chicken at a market in Paris; f_o range documented = c. 220–700 Hz, with additional voiced sounds up to 1000 Hz. Sound 4 = speech extract of a woman (doctor) speaking on a TV show; f_o range documented = c. 200–500 Hz. Sound 5 = speech extract of a woman (well-known singer) speaking on a TV show; f_o range documented = c. 350–650 Hz. Sound 6 = speech extract of a woman demonstrating infant-directed speech; f_o range documented = c. 200–1000 Hz. Sounds 7–9 = speech extracts of a man (well-known Imam) giving a sermon; f_o range documented = c. 170–600 Hz. Sound 10 = speech extract of a man (sports reporter) reporting on a soccer game; f_o range documented = c. 130–440 Hz. Sound 11 = speech extract of a man (lawyer) speaking on a TV show; f_o range documented = c. 130–600 Hz. Sound 12 = speech extract of a man demonstrating infant-directed speech; f_o range documented = c. 150–600 Hz.
[C-02-04-F01]

Figure 2. Examples of intelligible speech produced by actresses (ST style) and female singers (CS and EC styles), illustrating extensive f_o contours in contexts of artistic performance. Extract of Chapter M2.4, Table 2 (see Series 31, 6, 1, 5, 9, 24, 2 and 3 in this table). Sounds 1–12 = speech extracts of a female narrator telling fairy tales; f_o range documented = c. 110–880 Hz. Sounds 13–20 = speech extracts of a female comedian performing on stage; f_o range documented = c. 200–850 Hz. Sounds 21–23, and 24 = three speech extracts of a female singer (CS style) reading the standard text of the Zurich Corpus and demonstrating relaxed speech and unexaggerated and exaggerated reading on stage (imitating two performance modes); f_o range documented = c. 170–650 Hz; as a fourth sound, a song is added in order to demonstrate that speech and singing do not differ in principle with regard to their f_o ranges. Sounds 25–30 = speech extracts of a second female comedian performing on stage; f_o range documented = c. 170–780 Hz. Sounds 31–33 = speech extracts of a third female comedian performing on stage; f_o range documented = c. 200–850 Hz. Sound 34 = speech extract of a voice-over actress performing in a film; f_o range documented = c. 220–800 Hz. Sounds 35 and 36 = speech of two female singers (EC and CS style) reading the standard text of the Zurich Corpus in stage mode and demonstrating text reading including high levels of f_o; f_o range documented = c. 250–800 Hz and 350–1000 Hz, respectively.
[C-02-04-F02] ↗

Figure 3. Examples of intelligible speech produced by actors (ST style) and male singers (CS style), illustrating extensive f_o contours in contexts of artistic performance. Extract of Chapter M2.4, Table 3 (see Series 4, 18, 6, 11, 15, 3, 5, 19, 8 and 1 in this table). Sounds 1–9 = speech extracts of a male comedian performing on stage; f_o range documented = c. 130–700 Hz. Sounds 10–12 = speech extracts of a voice-over actor performing in a film; f_o range documented = c. 100–650 Hz. Sounds 13–18 = speech extracts of a second male comedian performing on stage; f_o range documented = c. 200–780 Hz. Sounds 19–24 = speech extracts of a third male comedian performing on stage; f_o range documented = c. 200–800 Hz. Sounds 25–27 = speech extracts of a male impressionist performing in a radio broadcast; f_o range documented = c. 90–550 Hz. Sounds 28–30 = speech extracts of an actor recorded at his home and during a performance on an outdoor stage; f_o range documented = c. 100–580 Hz. Sounds 31–33 = speech extracts of a Kabuki actor performing on stage; f_o range documented = c. 250–700 Hz. Sound 34 = speech extract of a voice-over actor performing in a film; f_o range documented = c. 200–800 Hz. Sound 35 = speech extract of a fourth male comedian performing on a TV show; f_o range documented = c. 150–600 Hz. Sound 36 = speech of an actor reading the standard text of the Zurich Corpus and demonstrating text reading in a falsetto voice on stage; f_o range documented = c. 250–550 Hz.
[C-02-04-F03] ↗

Figure 1. Examples of intelligible speech produced by untrained speakers, illustrating extensive fo contours in contexts of everyday life. [C-02-04-F01]



1–1 [Speech] 2172-A-w
R204217

1–2 [Speech] 2172-A-w
R204263

1–3 [Speech] 2172-A-w
R204328

1–4 [Speech] 2220-A-w
R204779

1–5 [Speech] 2336-A-w
R215903

1–6 [Speech] 2379-A-w
R203947

1–7 [Speech] 2420-A-m
R203999

1–8 [Speech] 2420-A-m
R203992

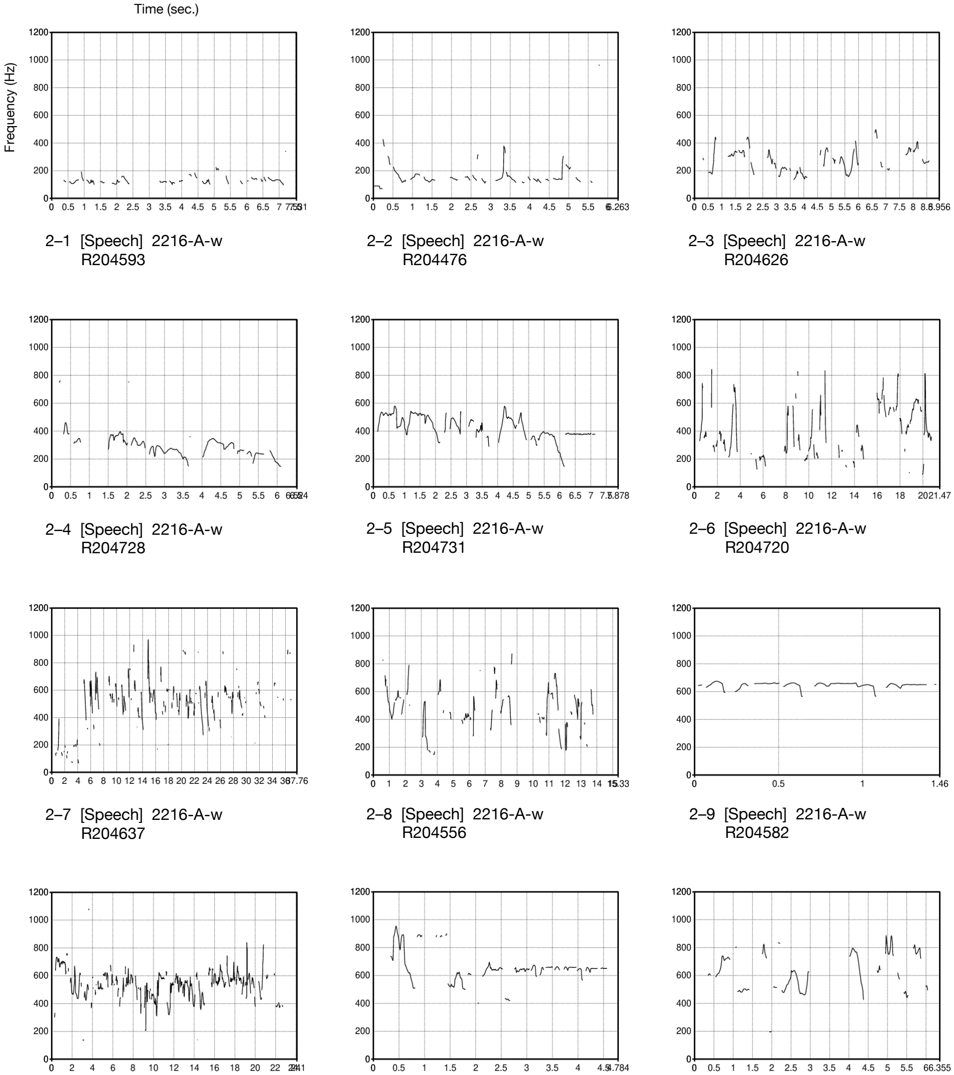
1–9 [Speech] 2420-A-m
R204028

1–10 [Speech] 2439-A-m
R204078

1–11 [Speech] 2498-A-m
R212239

1–12 [Speech] 2380-A-w
R204810

Figure 2. Examples of intelligible speech produced by actresses (ST style) and female singers (CS and EC styles), illustrating extensive fo contours in contexts of artistic performance. [C-02-04-F02]



2–1 [Speech] 2216-A-w
R204593

2–2 [Speech] 2216-A-w
R204476

2–3 [Speech] 2216-A-w
R204626

2–4 [Speech] 2216-A-w
R204728

2–5 [Speech] 2216-A-w
R204731

2–6 [Speech] 2216-A-w
R204720

2–7 [Speech] 2216-A-w
R204637

2–8 [Speech] 2216-A-w
R204556

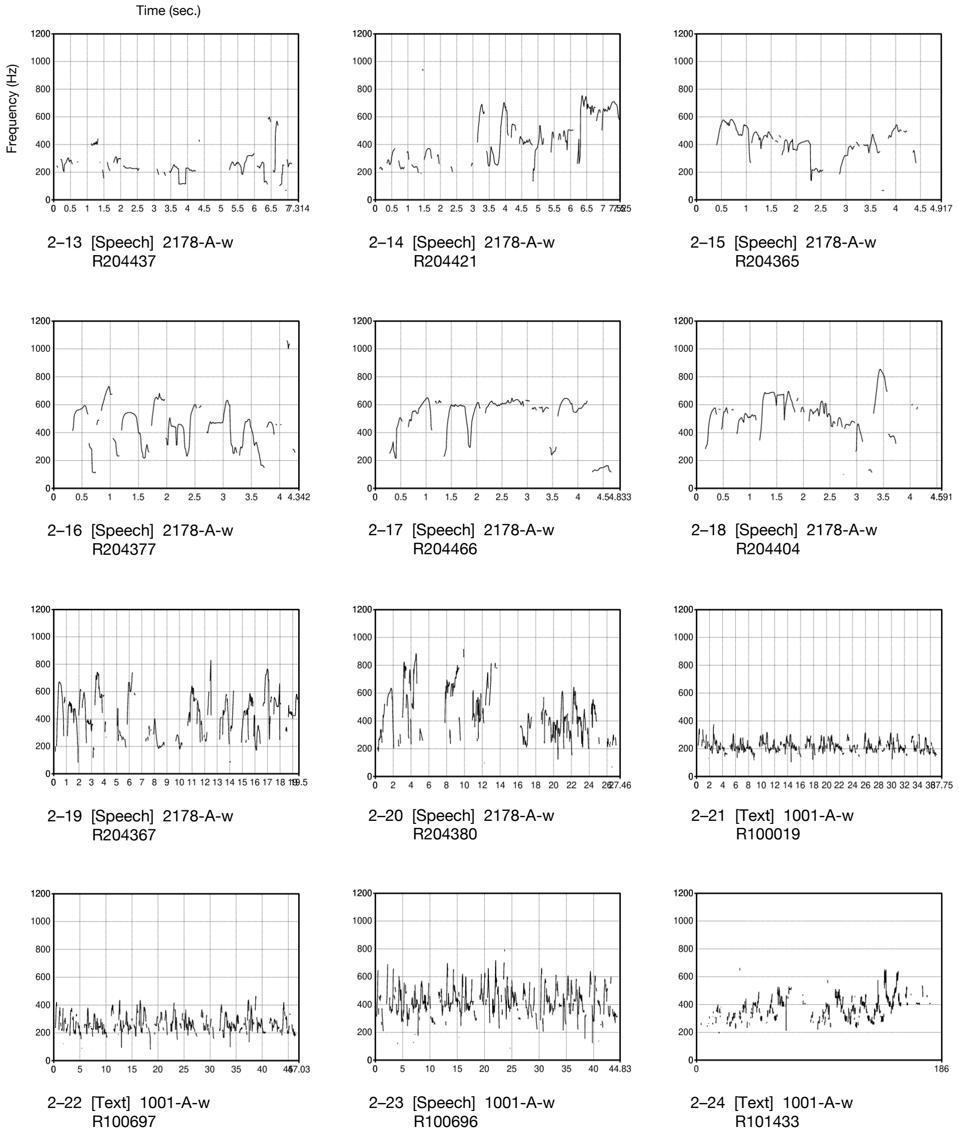
2–9 [Speech] 2216-A-w
R204582

2–10 [Speech] 2216-A-w
R204639

2–11 [Speech] 2216-A-w
R204560

2–12 [Speech] 2216-A-w
R204511

Figure 2 (continuation). [C-02-04-F02]



2–13 [Speech] 2178-A-w
R204437

2–14 [Speech] 2178-A-w
R204421

2–15 [Speech] 2178-A-w
R204365

2–16 [Speech] 2178-A-w
R204377

2–17 [Speech] 2178-A-w
R204466

2–18 [Speech] 2178-A-w
R204404

2–19 [Speech] 2178-A-w
R204367

2–20 [Speech] 2178-A-w
R204380

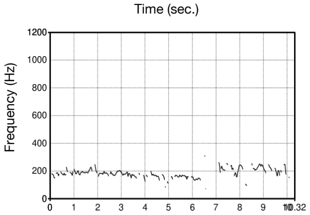
2–21 [Text] 1001-A-w
R100019

2–22 [Text] 1001-A-w
R100697

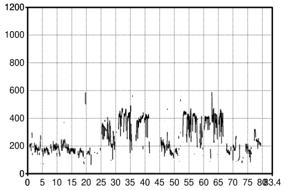
2–23 [Speech] 1001-A-w
R100696

2–24 [Text] 1001-A-w
R101433

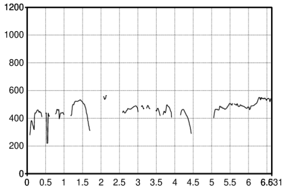
Figure 2 (continuation). [C-02-04-F02]



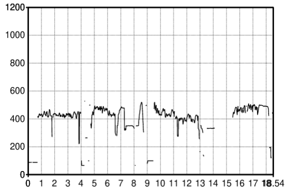
2–25 [Speech] 2177-A-w
R215422



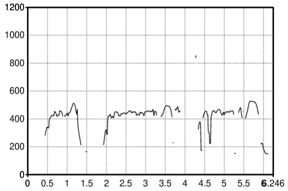
2–26 [Speech] 2177-A-w
R215482



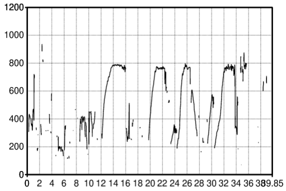
2–27 [Speech] 2177-A-w
R215695



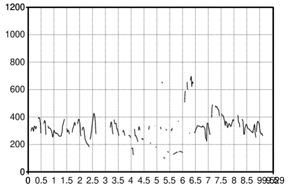
2–28 [Speech] 2177-A-w
R215494



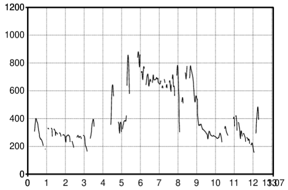
2–29 [Speech] 2177-A-w
R215469



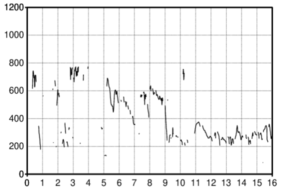
2–30 [Speech] 2177-A-w
R215646



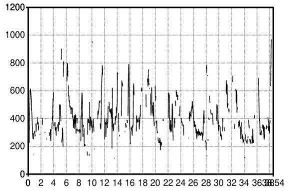
2–31 [Speech] 2234-A-w
R215570



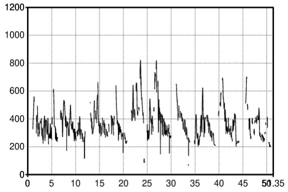
2–32 [Speech] 2234-A-w
R215568



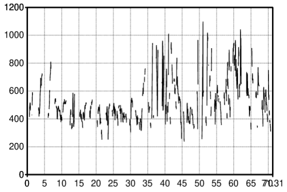
2–33 [Speech] 2234-A-w
R215569



2–34 [Speech] 2223-A-w
R215561

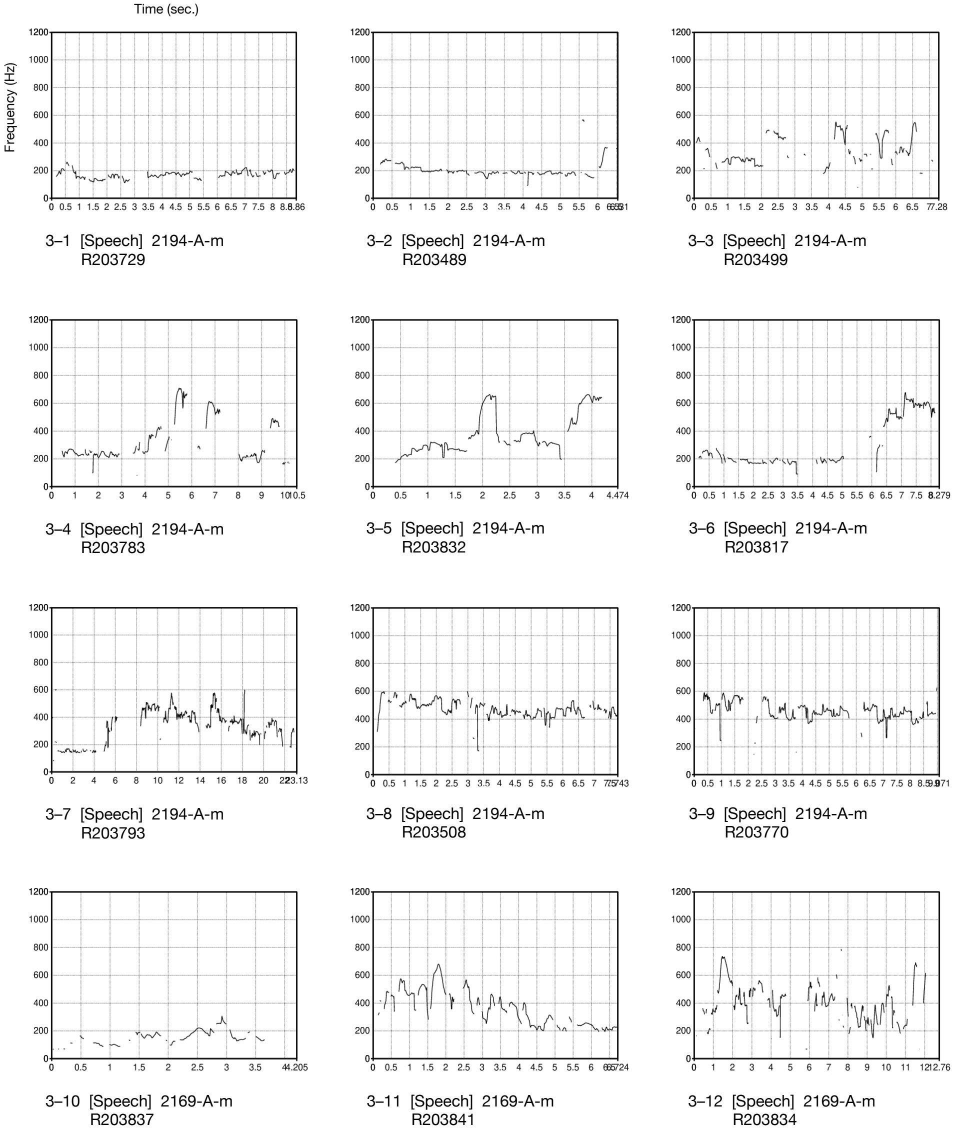


2–35 [Speech] 1005-A-w
R108403



2–36 [Speech] 1052-A-w
R142104

Figure 3. Examples of intelligible speech produced by actors (ST style) and male singers (CS style), illustrating extensive fo contours in contexts of artistic performance. [C-02-04-F03]



3–1 [Speech] 2194-A-m
R203729

3–2 [Speech] 2194-A-m
R203489

3–3 [Speech] 2194-A-m
R203499

3–4 [Speech] 2194-A-m
R203783

3–5 [Speech] 2194-A-m
R203832

3–6 [Speech] 2194-A-m
R203817

3–7 [Speech] 2194-A-m
R203793

3–8 [Speech] 2194-A-m
R203508

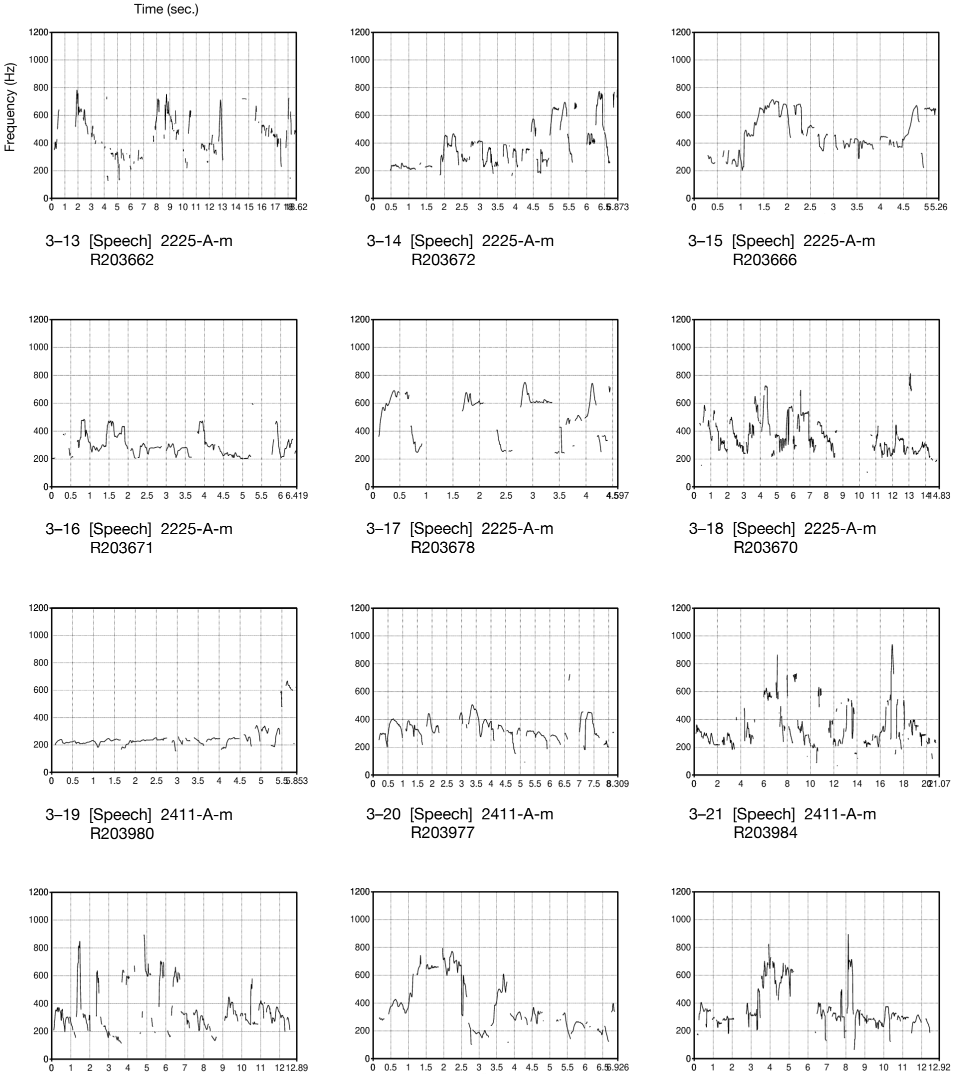
3–9 [Speech] 2194-A-m
R203770

3–10 [Speech] 2169-A-m
R203837

3–11 [Speech] 2169-A-m
R203841

3–12 [Speech] 2169-A-m
R203834

Figure 3 (continuation). [C-02-04-F03]



3–13 [Speech] 2225-A-m
R203662

3–14 [Speech] 2225-A-m
R203672

3–15 [Speech] 2225-A-m
R203666

3–16 [Speech] 2225-A-m
R203671

3–17 [Speech] 2225-A-m
R203678

3–18 [Speech] 2225-A-m
R203670

3–19 [Speech] 2411-A-m
R203980

3–20 [Speech] 2411-A-m
R203977

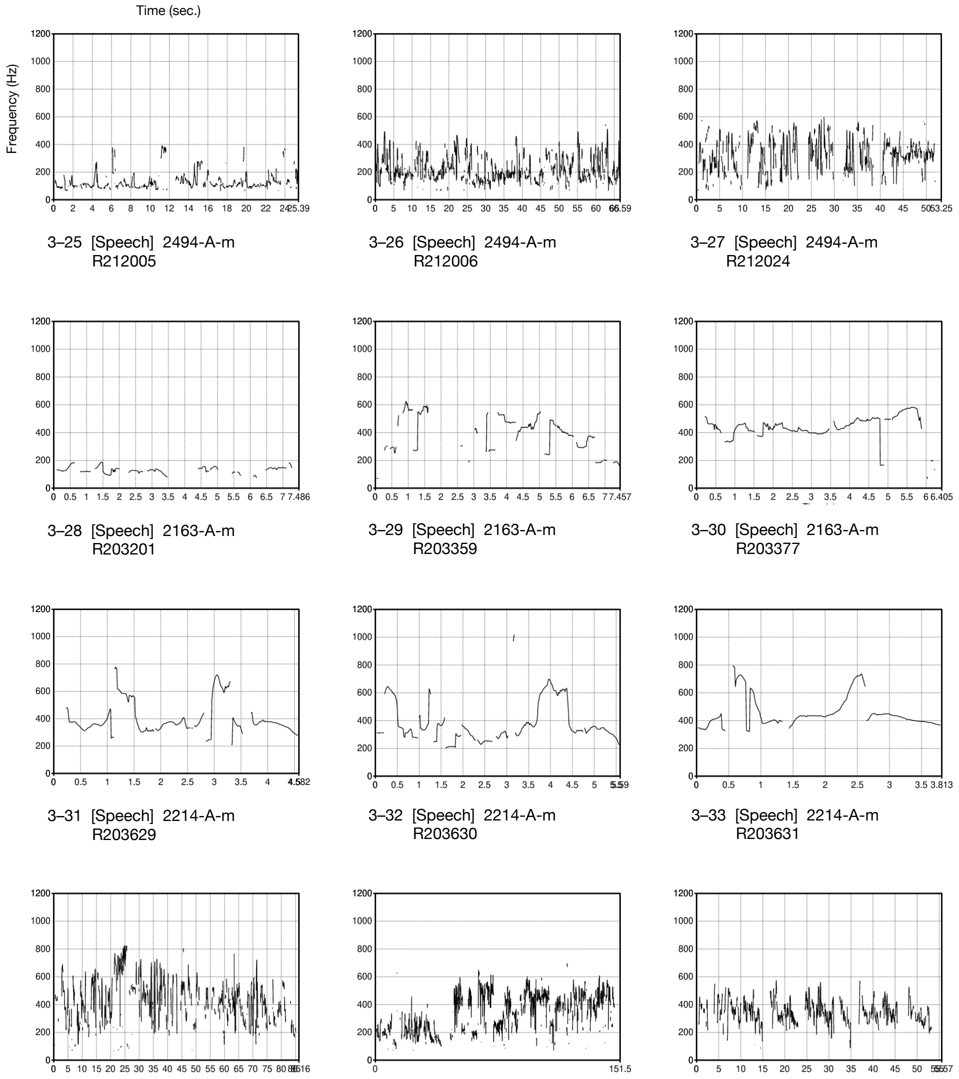
3–21 [Speech] 2411-A-m
R203984

3–22 [Speech] 2411-A-m
R203988

3–23 [Speech] 2411-A-m
R203986

3–24 [Speech] 2411-A-m
R203979

Figure 3 (continuation). [C-02-04-F03]



3–25 [Speech] 2494-A-m
R212005

3–26 [Speech] 2494-A-m
R212006

3–27 [Speech] 2494-A-m
R212024

3–28 [Speech] 2163-A-m
R203201

3–29 [Speech] 2163-A-m
R203359

3–30 [Speech] 2163-A-m
R203377

3–31 [Speech] 2214-A-m
R203629

3–32 [Speech] 2214-A-m
R203630

3–33 [Speech] 2214-A-m
R203631

3–34 [Speech] 2294-A-m
R203870

3–35 [Speech] 2297-A-m
R203897

3–36 [Speech] 1003-A-m
R104935

2.5 Conclusion

As demonstrated, extensive f_o variation is not a side but a core phenomenon of speech and, thus, of the vowel sound. That's why this treatise introduces its course of argument with the documentation and discussion of (i) vocalises, here in terms of recognisable vowel sounds produced by single speakers with extensive variation of f_o exceeding a range of two octaves for adults and one and a half octaves for children, (ii) recognisable vowel sounds produced at high f_o levels above statistically given levels of F_1 for almost all vowels but above all for close and close-mid vowels, (iii) examples of intelligible minimal pairs produced at these high f_o levels, and (iv) speech extracts with f_o contours also ranging up to these high f_o levels. Against such a background, we conclude that these four observations and experiences form an ensemble of phenomena to which an acoustic theory of the vowel has to relate as a general reference and premise.

Formulated in general terms, firstly, an increase of f_o from lower to higher levels did not principally impede the differentiation of long vowels even if on- and offsets of the vowel sounds were deleted, and up to an f_o level of c. 1 kHz, the vowel quality of the sounds of corner vowels (including /y/) could be differentiated perceptually by untrained listeners. Our studies further indicated that the vowel quality of unmanipulated sounds of all long Standard German vowels (isolated sounds including on- and offsets as well as minimal pairs) produced by speakers with exceptional vocal abilities at f_o levels of c. 1 kHz could also be recognised successfully by well-trained listeners. Secondly, as a tendency, the studies indicated that an increase of f_o from lower to higher levels had a more substantial effect on successful vowel recognition for the sounds of close-mid than for the sounds of the other vowel qualities. Thirdly, vowel sounds in the upper ranges of f_o, as discussed here, proved to be not only a phenomenon of singing but also of speech, and they can often be encountered not only in artistic performances but also in everyday life. Indeed, artistic performances very often mirror and represent vocal expressions from everyday life.

A future acoustic theory of the vowel has to refer to these reference statements conceptually, methodologically and experimentally: The theory should embrace the vowel phenomenon as such, independent of the f_o levels of recognisable vowel sounds.

The facts that (i) many speakers only speak within a somewhat limited vocal range and are probably not able to produce vowels at the high f_o levels mentioned above, (ii) many speakers are not capable of maintaining the pronunciation of vowel quality independent of f_o levels of production (an ability professional speakers and singers train for extensively), (iii) many listeners may not recognise and differentiate vowel sounds produced at high f_o levels, and (iv) recordings in an experimental setting with restricted conditions and at high levels of f_o often cause stress for phonation and articulation do not relativise the reference statements made. Many other speakers do possess excellent vocal abilities, including large vocal ranges and clear vowel pronunciation, experienced listeners can provide consistent vowel recognition results for sounds at high f_o levels, and future studies that are not constrained by an experimental setting – e.g. studies based on recordings made in everyday situations or during stage performances or recordings extracted from films and on extracted vowel sounds of the utterances – may provide more evidence for vowel recognition of sounds with less stress of phonation and articulation produced at the f_o levels discussed here.

The observations and experiences documented also clarify the debate on two hypotheses often supported in the literature which claim that spectral undersampling and/or oversinging F_1 imperatively impair vowel recognition: The simple fact that the sounds of the corner vowels (including /y/) – and with them the sounds of close vowels associated with the lowest statistical F_1 of all vowel qualities – can be produced and recognised up to an f_o level of c. 1 kHz contradicts both hypotheses. An increase of f_o levels per se neither results in a general impairment of recognised vowel quality nor in a general vowel quality shift towards /a/, even if f_o far surpasses statistical F_1.

Because the f_o range of recognisable natural vowel sounds covers the entire range of statistical F_1 for all vowel qualities generally given for adults, concerning the spectral energy distribution of the sounds in general and the estimated spectral peaks or F-patterns or entire spectral envelopes in particular (as far as their estimation is methodologically substantiated), these characteristics relate to f_o, above all concerning the frequency range below c. 1.5–2 kHz. Further, for the same reason, it is evident that no existing concept of general vowel-specific patterns of spectral peaks or formants or vowel-specific spectral envelopes can account for the finding of the actual f_o range of recognisable vowel sounds, not only because of the resulting variation of the vowel spectrum but also because of the lack of a methodological substantiation to estimate these spectral characteristics for all recognisable vowel sounds independent of f_o.

Note that, in this first state of exposition of indices, f_o and pitch are parallel aspects. Only further investigation will show that these two characteristics must be considered as distinct characteristics.

3 Ambiguity of Spectral Peaks, Estimated Formant Patterns and Spectral Shapes

3.1 Source–Filter Synthesis Based on Statistical Formant Frequencies, Including Variation of f_o

One way to interpret estimated statistical *F*-patterns of natural vowel sounds and evaluate their role in vowel quality recognition is to use a source–filter synthesiser to reproduce sounds based on these *F*-patterns and then test the recognition thereof.

According to the results of two published studies on two different reference statistics of *F*-patterns related to American English (for details, see Chapter M3.1), synthesised vowel sounds of this type remained recognisable even if the vowel recognition rate when compared to natural sounds dropped depending on whether or not they were based on dynamic sound properties used in synthesis (contours of *F*-patterns and f_o). (Notably, pronounced recognition differences for different vowel qualities also occurred in these studies.) Because the average recognition rate for all production conditions was found to be above 70%, it was concluded that static *F*-patterns represent the primary acoustic characteristics of vowel quality and that dynamic properties play a secondary – although quite important – role in vowel quality recognition.

However, the two studies did not include pronounced f_o variation, neither when investigating statistical *F*-patterns of natural sounds of a speaker group nor when investigating vowel synthesis based on these patterns. Therefore, it is not evident that the matching of intended vowel qualities of natural sounds and recognised vowel qualities of synthesised replicas – both types of sounds produced at equal f_o – leads to the conclusion that the *F*-pattern itself is vowel quality-related *in general,* independent of the f_o of the natural sounds that the measurement of the *F*-pattern is based on and independent of the f_o of synthesis.

In order to tackle this question, we examined the effect of f_o variation on vowel recognition for vowel synthesis that was based on the statistical average *F*-patterns of three different studies, Maurer et al. (1992, hereafter referred to as MA), Pätzold and Simpson (1997, hereafter referred to as PS) and Fant (1959, hereafter referred to as FA). The *F*-patterns given by MA were selected because they report values for men, women and children, including different and similar f_o levels for the sounds produced by the speakers of the three different speaker groups; however, values are given for the long Swiss German vowels /i, e, a, o, u/ only (region of the canton of Zurich). The *F*-patterns given by PS were selected because they report values for men and women of the eight vowels /i, y, e, ø, ɛ, a, o, u/ of Standard German. (Note that, in this study, formant measurement was conducted without editing calculated formant tracks, and the vowel /ɛ/ was investigated as a short vowel.) Finally, the *F*-patterns given by FA were selected because they represent a historical basis within the context of the formulation of the source–filter theory of speech production and, at the same time, the vowels /i, y, e, ø, ɛ, a, o, u/ of Swedish selected for the present study are comparable to variants of long Standard German vowels.

For each of the reported single *F*-patterns, seven steady-state sounds of 1 sec. were synthesised at seven different f_o levels of 65–131–220–262–330–440–523 Hz using a Klatt synthesiser (cascade mode; for details of formant frequencies and bandwidths applied, see Chapter M3.1). The f_o range of 131–523 Hz covers both the f_o levels of recognisable vowel sounds to observe in everyday speech and the statistical F_1 levels often given in the literature for sounds of close and close-mid vowels; the f_o level of 65 Hz was added to imitate creaky phonation. Vowel recognition was assessed in a listening test according to the standard procedure of the Zurich Corpus and involving the five standard listeners.

The main results regarding vowel recognition are shown in Table 1. Recognised vowel qualities with a labelling majority (recognition rate ≥ 60%) are given. In general terms and formulated as a tendency, the results for this kind of synthesis indicated two types of effects of f_o variation on vowel recognition: (1) If, for a given statistical *F*-pattern, f_o was increased incrementally from a lower level of 65 Hz to a higher level of 330 Hz in synthesis, this variation either had no effect on the recognised vowel quality or the quality shifted in an open–close direction (sometimes including an additional unrounded–rounded shift). Open–close shifts were pronounced above all for sounds of the close-mid vowels /e, o/. (2) If f_o was further increased and thereby substantially surpassed the levels of statistical F_1 of close vowels and, subsequently, also of some close-mid vowels, in some cases, the open–close shifts newly occurred or continued (pronounced for /e, ɛ, o/), while in other cases, they were inverted or reverted to close–open shifts (pronounced for /i, y/). For the rest of the cases, no shifts were found.

In these terms, the main indication provided by this experimentation is that vowel recognition does not relate to statistical *F*-patterns independent of the f_o levels of the synthesised sounds. Figure 1 illustrates the main finding of occurring vowel quality shifts in an open–close direction as a result of f_o variation in synthesis, with the related *F*-patterns kept unchanged.

However, within the limits of the general tendencies of vowel quality shift directions mentioned, the individual shifts observed depended on vowel qualities of the natural reference sounds, different studies and their statistical *F*-patterns, as well as on speaker groups and related statistical f_o. Thus, spectral representation of vowel quality was again indicated to be nonuniform.

Finally, inverted or reverted shifts in a close–open direction for sounds synthesised at f_o of 440 and 523 Hz may have to be considered with regard to two different aspects: On the one hand, these shifts may have been related to the specific fine structure of the harmonic spectra of the sounds in question resulting from the relation between the LPC filter curve and the f_o level applied in synthesis (consider, above all, the frequency distance of the harmonics and the resulting sampling of the filter curve including the match or mismatch of harmonics and filter frequencies). But on the other hand, above all for sounds of close front vowels for which these inverted shifts mainly occurred in the present investigation, the change in the relation between lower and higher spectral energy may have been the cause of the inverted shifts. Notably, if sounds of close (and sometimes also of close-mid) vowels were HP filtered applying CFs below c. 1 kHz, which in its turn resulted in a change in the relation between the lower and higher spectral energy, close–open shifts were observed (see Chapter 8.2). As we will argue below, this supports the thesis of the vowel being a kind of foreground–background phenomenon.

For references, extended background information, details of experimental design, method and results, an extended discussion (including some relativisations) and documentation of results (tables including sound links), see the Materials, Chapter M3.2.

Table 1: Source–filter synthesis based on statistical *F*-patterns, including variation of f_o: Summary of the vowel recognition results. x-axis = *F*-pattern-related studies of Maurer et al. (1992, MA), Pätzold and Simpson (1997, PS) and Fant (1959, FA) for the speaker groups of men (m), women (w) and children (c); y-axis = f_o of synthesis. Vowel qualities indicated = qualities labelled by a majority of the five listeners (empty positions in the graphs correspond to cases without a majority). Colour code: Green = f_o of synthesis corresponded to statistical levels as given for the investigated speaker groups and sounds in the MA and FA studies and correspondingly also related to *F*-patterns of the PS study; no colour = matching intended vowel quality of a natural sound and recognised quality after synthesis, or no labelling majority; red = recognition mismatch of a synthesised sound and resulting vowel quality shift in an open–close direction with increasing f_o from low to high; purple = recognition mismatch of a synthesised sound and resulting vowel quality shift in an inverted or reverted close–open direction with increasing f_o to 440 and/or 523 Hz.
[C-03-01-T01]

Figure 1. Source–filter synthesis based on statistical *F*-patterns, including variation of f_o: Illustration of occurring vowel quality shifts in an open–close direction due to an increase of f_o from lower to higher levels. Extract of Chapter M3.1, Table 3 (see Series 1–3 in this table). Sounds 1 to 2 = synthesised sounds related to a statistical *F*-pattern of /e/ of 335–2050–2633 Hz (MA study, men) associated with a statistical f_o of 131 Hz (according to the musical C-major scale); f_o of synthesis of 131–330 Hz resulting in a recognised vowel quality shift from /e/ to /y/. Sounds 3 to 5 = synthesised sounds related to a statistical *F*-pattern of /ø/ of 363–1690–2200 Hz (FA study, men) associated with a statistical f_o of 131 Hz; f_o of synthesis of 131–330–440 Hz resulting in a recognised vowel quality shift from /ø/ to /y/. Sounds 6 to 8 = synthesised sounds related to a statistical *F*-pattern of /o/ of 346–700–2600 Hz (MA study, men) associated with a statistical f_o of 131 Hz; f_o of synthesis of 131–330–440 Hz resulting in a recognised vowel quality shift from /o/ to /u/.
[C-03-01-F01]

Table 1. Source–filter synthesis based on statistical F-patterns, including variation of fo: Summary of the vowel recognition results. [C03-01-T01]

Vowel recognition

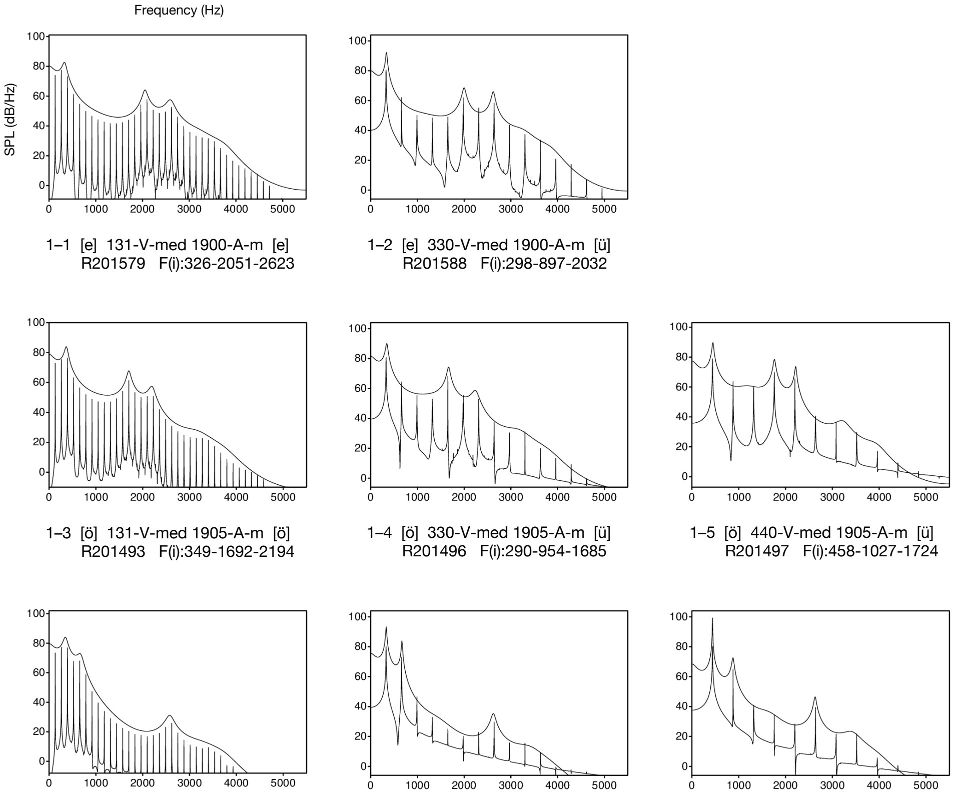
fo of vowel resynthesis (in Hz)	/i/										/u/										/y/			
523	–	–	–	e	e	e	e	e	e	e	u	u	u	u	u	o	u	u	u	u	ø	–	ø	e
440	e	–	e	e	e	e	e	e	e	i	u	–	u	u	u	u	u	o	u	u	ø	ø	ø	e
	F1 limit (close vowels)																							
330	i	y	i	i	i	i	i	i	i	i	u	u	u	u	u	u	u	u	u	u	y	y	y	i
262	i	y	i	i	i	–	i	i	i	i	u	u	u	u	u	u	u	u	u	u	y	y	ø	i
220	i	y	i	i	i	–	i	i	i	i	u	u	u	u	u	–	u	u	u	u	y	y	ø	i
131	i	y	i	i	i	e	i	i	i	e	u	u	u	u	u	u	o	u	u	u	ø	y	ø	i
65	i	y	i	i	i	e	i	e	i	e	u	u	u	u	u	–	–	u	u	u	ø	y	ø	i
	131			220				262			131			220				262			131		220	
	m	m	m	m	w	w	w	m	w	c	m	m	m	m	w	w	w	m	w	c	m	m	w	w
	MA	PS	FA	MA		PS	FA	MA			MA	PS	FA	MA		PS	FA	MA			PS	FA	PS	FA

fo of vowel resynthesis (in Hz)	/e/										/o/										/ø/			
523	–	–	–	–	e	–	e	y	–	i	u	u	u	u	u	u	u	u	u	u	ø	ø	ø	–
440	e	y	e	–	i	–	i	–	e	e	u	o	u	u	u	u	u	–	–	u	ø	y	y	e
	F1 limit (close vowels)																							
330	y	y	y	e	e	e	i	e	e	e	u	u	u	o	o	u	–	o	o	o	øy	y	ø	y
262	–	ø	y	e	e	e	e	e	e	e	u	u	o	o	o	o	o	o	o	o	ø	ø	ø	–
220	–	ø	y	e	e	e	e	e	e	–	o	o	o	o	o	o	o	o	o	ɔ	ø	ø	ø	e
131	e	ø	e	ɛ	ɛ	e	e	ɛ	ɛ	ɛ	o	o	o	ɔ	ɔ	o	o	ɔ	ɔ	ɔ	ø	ø	ø	ø
65	e	ø	e	ɛ	ɛ	ɛ	e	ɛ	ɛ	ɛ	–	o	o	ɔ	ɔ	o	o	ɔ	ɔ	ɔ	ø	ø	ø	e
	131			220				262			131			220				262			131		220	
	m	m	m	m	w	w	w	m	w	c	m	m	m	m	w	w	w	m	w	c	m	m	w	w
	MA	PS	FA	MA		PS	FA	MA			MA	PS	FA	MA		PS	FA	MA			PS	FA	PS	FA

fo of vowel resynthesis (in Hz)	/ɛ/										/a/									
523		ø	–			y	y				–	–	–	a	a	a	a	a	–	a
440		ø	y			–	e				a	a	ɔ	a	a	a	ɔ	a	a	a
	F1 limit (close vowels)																			
330		ø	ø			–	e				o	–	o	–	–	a	a	a	a	a
262		ø	ø			–	e				–	–	aɔ	a	a	a	a	a	a	a
220		ø	ø			ɛ	ɛe				a	a	–	a	a	a	a	a	a	a
131		–	ø			ɛ	ɛ				a	a	a	a	a	a	a	a	a	a
65		–	ø			ɛ	ɛ				a	–	a	a	a	a	a	a	a	a
	131			220				262			131			220				262		
	–	m	m	–	–	w	w	–	–	–	m	m	m	m	w	w	w	m	w	c
	MA	PS	FA	MA		PS	FA	MA			MA	PS	FA	MA		PS	FA	MA		

F-patterns and related speaker groups and fo (in Hz)

Figure 1. Source–filter synthesis based on statistical F-patterns, including variation of fo: Illustration of occurring vowel quality shifts in an open–close direction due to an increase of fo from lower to higher levels. [C-03-01-F01]



1–1 [e] 131-V-med 1900-A-m [e]
R201579 F(i):326-2051-2623

1–2 [e] 330-V-med 1900-A-m [ü]
R201588 F(i):298-897-2032

1–3 [ö] 131-V-med 1905-A-m [ö]
R201493 F(i):349-1692-2194

1–4 [ö] 330-V-med 1905-A-m [ü]
R201496 F(i):290-954-1685

1–5 [ö] 440-V-med 1905-A-m [ü]
R201497 F(i):458-1027-1724

1–6 [o] 131-V-med 1900-A-m [o]
R201663 F(i):344-701

1–7 [o] 330-V-med 1900-A-m [u]
R201672 F(i):307-666

1–8 [o] 440-V-med 1900-A-m [u]
R201675 F(i):438-851

3.2 Source–Filter Resynthesis Based on Estimated Spectral Envelopes of Single Natural Sounds, Including Variation of f_o

Pursuing the investigation of vowel recognition for synthesised and resynthesised sounds (see the general Introduction to this treatise regarding this terminological differentiation) based on supposed vowel quality-related spectral characteristics but varying f_o, in a further study, a resynthesis was conducted relating to spectral envelopes of single natural vowel sounds instead of using estimated statistical *F*-patterns and, referring to Hillenbrand et al. (2006), using spectral envelope synthesis.

Based on the Zurich Corpus, sounds of three untrained speakers (non-professionals), one man, one woman and one child, were selected. The speakers produced sounds of the eight long Standard German vowels in nonstyle mode with voiced phonation and medium vocal effort in V context at different intended f_o levels of 220–262–330–440–523–659 Hz (all speakers), 165 Hz (man and woman) and 131 Hz (man only). As a result, a sample of 168 natural reference vowel sounds was created. Except for a few sounds, the vowel recognition rate obtained in the listening test conducted when creating the corpus was 100% matching vowel intention.

For each single natural reference sound, the dynamic contours of the harmonic envelope and f_o were calculated (entire sounds, including on- and offsets). Subsequently, for each single spectral envelope contour, sounds as replicas were resynthesised using a spectral envelope synthesiser (for details, see Chapter M3.2) at eight f_o levels of 131–165–220–262–330–440–523–659 Hz (these values given as approximations to the musical C-major scale), that is, at the f_o level (and related contour) of the natural reference sound and then at f_o (and related contours) shifted up or down to the other seven remaining levels. Note that the f_o level of 65 Hz was not applied in the present experiment because, in the previous experiment, no marked vowel quality shifts were observed for the f_o variation of 65–131 Hz, except for one single sound. However, the high f_o level of 659 Hz was included to extend the frequency range of f_o variation.

Vowel recognition of natural and resynthesised sounds was assessed in a listening test according to the standard procedure of the Zurich Corpus and involving the five standard listeners, with the test divided into eight subtests, each subset presenting sounds at similar f_o. Natural and resynthesised sounds were not separated in the test. (Note that even though the vowel quality recognition of the natural vowel sounds used in this experiment had already been tested when creating the database of the Zurich Corpus, they were tested anew here in a different sound context. Hence, the vowel quality recognition results given below concern the results of this experiment-specific test.)

Notably, vowel recognition was examined within this experimental setting for f_o variation of both the natural reference sounds and their replicas, in contrast to the previous experiment relating to statistical *F*-patterns. As a consequence, f_o-related vowel quality shifts for resynthesised sounds were not examined mainly for increasing f_o from a lower to a higher frequency level when compared with the f_o level of the natural reference sound, but for both directions of f_o variation.

Table 1 shows the results of the listening test. The recognised vowel qualities with a labelling majority (recognition rate ≥ 60%) are given for the sounds investigated. In general terms and formulated as a tendency, with few exceptions, the resynthesis of the natural vowel sounds based on the related spectral envelopes and applying a step-by-step increase of f_o from 131 to 659 Hz resulted in a vowel quality shift in an open–close direction (and sometimes also in additional unrounded–rounded shifts), the shifts sometimes involving more than two adjacent vowel qualities. However, as was the case in the previous synthesis experiment relating to statistical *F*-patterns, the occurrence and the extent of the shifts proved to be nonuniform: They varied among the vowel qualities of the natural reference sounds, the f_o levels of these reference sounds, the range of f_o variation of the synthesised replicas as well as the individual course of the spectral envelope, possibly due to intra- or inter-speaker differences of sound production. Figures 1 to 4 illustrate these two main findings of a general open–close vowel quality shift direction due to increasing f_o levels in resynthesis and the nonuniform character of the shifts in relation to individual natural reference sounds.

Exceptions to these findings concerned rare cases of inverted or reverted close–open shifts and front–back or back–front confusions (for details, see Chapter M3.2).

Comparing the results of this resynthesis experiment relating to single natural sounds and their spectral envelopes with the results of the previous synthesis experiment relating to statistical *F*-patterns and LPC curves, two main differences were indicated. The first difference concerned resynthesis related to the spectral envelopes of the natural vowel sounds produced at f_o below 330 Hz: If f_o was raised in resynthesis from a lower to a higher level and thereby substantially surpassed the frequency level commonly assumed as the statistical F_1 of the vowel quality in question, inverted or reverted close–open shifts following previous open–close shifts occurred very rarely, in contrast to the synthesis related to statistical F-patterns. The second difference concerned the extent of the vowel quality shifts: These shifts were far more pronounced in the present than in the previous experimentation. Both kinds of recognition differences may be due either to the difference in the f_o ranges of the natural vowel sounds and their replicas investigated or to the difference between vowel synthesis based on averaged LPC curves and vowel resynthesis based on a spectral envelope of a single natural sound. Furthermore, the methodological differences in estimating the LPC curve or the spectral envelope and the resulting effects on experiments of this kind also have to be considered.

In conclusion, the results of this study extended the findings of the previous experiment and confirmed that vowel recognition did not relate to a measured spectral envelope of a natural sound independent of f_o levels. Thus, the spectral envelope of a natural sound, *per se,* indeed proves to be an ambiguous representation of vowel quality because of the role of its relation to f_o (and pitch) for vowel sounds.

For references, extended background information, details of experimental design, method and results, an extended discussion (including some relativisations) and documentation of results (tables including sound links), see the Materials, Chapter M3.2.

Table 1: Source–filter resynthesis based on spectral envelopes of single natural reference sounds: Summary of the vowel recognition results. x-axis = average f_o levels of the natural reference vowel sounds (in Hz, according to the musical C-major scale), given separately for the utterances of the three speakers (m = man, w = woman, c = child). y-axis = f_o levels of resynthesis (in Hz, according to the musical C-major scale). Vowel qualities indicated = qualities with a labelling majority (recognition rate ≥ 60%; empty positions in the table correspond to cases with no majority). Colour code and "*" marks: Green = resynthesis at f_o of the natural reference sound; no colour = resynthesis at f_o differing from f_o of the natural reference sound but with recognised vowel quality matching the quality of the natural reference sound, or no labelling majority; red = resynthesis at f_o differing from f_o of the natural reference sound, associated with a vowel quality shift in an open–close direction with increasing f_o from a lower to a higher level in resynthesis; purple = resynthesis at f_o differing from f_o of natural reference sound, associated with a vowel quality shift in a close–open direction with increasing f_o from a lower to a higher level in resynthesis; grey and/or "*" marks = front–back or back–front confusions.
[C-03-02-T01]

Figure 1: Source–filter resynthesis based on spectral envelopes of single natural reference sounds: First illustration of occurring vowel quality shifts in an open–close direction due to an increase in f_o from the level of the reference sound to higher levels. Extract of Chapter M3.2, Table 1. Sounds 1 to 3 = resynthesised sounds related to the spectral envelope of a natural reference sound of /e/ produced by a woman at an intended f_o of 220 Hz, with f_o of resynthesis of 220–330–440 Hz resulting in a recognised vowel quality shift from /e/ to /i/. Sounds 4 to 6 = resynthesised sounds related to the spectral envelope of a natural reference sound of /ø/ produced by a man at an intended f_o of 131 Hz, with f_o of resynthesis of 131–262–330 Hz resulting in a recognised vowel quality shift from /ø/ to /y/. Sounds 7 to 9 = resynthesised sounds related to the spectral envelope of a natural reference sound of /o/ produced by a woman at an intended f_o of 220 Hz, with f_o of resynthesis of 220–440–523 Hz resulting in a recognised vowel quality shift from /o/ to /u/.
C-03-02-F01]

Figure 2: Source–filter resynthesis based on spectral envelopes of single natural reference sounds: Second illustration of occurring vowel quality shifts in an open–close direction due to an increase in f_o from a low level up to the level of the reference sound. Extract of Chapter M3.2, Table 1. Sounds 1 to 3 = resynthesised sounds related to the spectral envelope of a natural reference sound of /i/ produced by a woman at an intended f_o of 440 Hz, with f_o of resynthesis of 220–262–440 Hz resulting in a recognised vowel quality shift from /e/ to /i/. Sounds 4 and 5 = resynthesised sounds related to the spectral envelope of a natural reference sound of /y/ produced by a woman at an intended f_o of 330 Hz, with f_o of resynthesis of 220–330 Hz resulting in a recognised vowel quality shift from /ø/ to /y/. Sounds 6 and 7 = resynthesised sounds related to the spectral envelope of a natural reference sound of /u/ produced by a child at an intended f_o of 440 Hz, with f_o of resynthesis of 262–440 Hz resulting in a recognised vowel quality shift from /o/ to /u/.
[C-03-02-F02]

Figure 3: Source–filter resynthesis based on spectral envelopes of single natural reference sounds: Third illustration of occurring vowel quality shifts in an open–close direction due to an increase in f_o from lower to higher levels, the shifts including adjacent and non-adjacent vowel qualities. Extract of Chapter M3.2, Table 1. Sounds 1 to 3 = resynthesised sounds related to the spectral envelope of a natural reference sound of /e/ produced by a man at an intended f_o of 330 Hz, with f_o of resynthesis of 165–330–659 Hz resulting in a recognised vowel quality shift from /ɛ/ to /e/ and /y/. Sounds 4 to 7 = resynthesised sounds related to the spectral envelope of a natural reference sound of /o/ produced by a man at an intended f_o of 330 Hz, with f_o of resynthesis of 165–220–330–659 Hz resulting in a recognised vowel quality shift from /ɔ/ to /o/ and /u/.
[C-03-02-F03] ↗

Figure 4: Source–filter resynthesis based on spectral envelopes of single natural reference sounds: Illustration of the nonuniform character of occurring vowel quality shifts related to f_o variation. Extract of Chapter M3.2, Table 2 (see Series 9–12 in this table). Two series of sounds resynthesised based on two natural reference sounds of /ɛ/ and two corresponding series of /a/ are shown, with pronounced vowel quality shifts occurring for only the first sound series. Sounds 1 to 4 = resynthesised sounds related to the spectral envelope of a natural reference sound of /ɛ/ produced by a woman at an intended f_o of 220 Hz, with f_o of resynthesis of 220–330–440–659 Hz resulting in a recognised vowel quality shift from /ɛ/ to /e/ and /i/. Sounds 5 to 9 = resynthesised sounds related to the spectral envelope of a natural reference sound of /ɛ/ produced by a child at an intended f_o of 330 Hz, with f_o of resynthesis of 220–330–440–523–659 Hz resulting in no pronounced vowel quality shifts. Sounds 10 to 12 = resynthesised sounds related to the spectral envelope of a natural reference sound of /a/ produced by a woman at an intended f_o of 220 Hz, with f_o of resynthesis of 165–220–523 Hz resulting in no pronounced vowel quality shifts. Sounds 13 to 15 = resynthesised sounds related to the spectral envelope of a natural reference sound of /a/ produced by a child at an intended f_o of 440 Hz, with f_o of resynthesis of 220–440–659 Hz resulting in no pronounced vowel quality shifts.
[C-03-02-F04] ↗

Table 1. Source–filter resynthesis based on spectral envelopes of single natural reference sounds: Summary of the vowel recognition results. [C03-02-T01]

fo of vowel resynthesis (in Hz)	Vowel recognition																				
	/i/																				
659	i	i	i	i	i	i	i	i	i	i	i	i	i	i	–	i	i	u*	–	i	–
523	i	y	i	i	i	i	i	i	i	i	i	i	i	i	i	i	i	u*	ε	–	e
440	i	i	i	i	i	i	i	i	i	i	i	i	i	i	i	e	e	o*	e	e	ε
330	i	i	i	i	i	i	i	i	i	i	i	i	e	e	e	e	e	o*	ε	ε	ε
262	i	i	i	i	i	i	i	i	i	e	e	i	e	e	e	e	e	o*	ε	–	ε
220	i	e	e	i	i	i	e	i	i	e	e	i	–	e	–	ε	ε	o*	ε	ε	ε
165	i	i	i	i	e	i	i	i	i	e	e	i	–	e	–	ε	ε	ɔ*	ε	–	ε
131	i	e	e	e	e	i	e	i	e	–	e	i	ε	e	–	–	e	–	ε	ε	ε
	131	165		220			262			330			440			523			659		
	m	m	w	m	w	c	m	w	c	m	w	c	m	w	c	m	w	c	m	w	c
	/y/																				
659	y	y	y	y	y	y	y	y	y	y	y	–	y	y	u*	y	–	y	–	y	y
523	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y	–	ε	–
440	y	y	y	y	y	y	y	y	y	i	y	y	y	y	y	ε	e	–	ε	ε	ε
330	y	i	y	–	y	y	y	y	–	y	y	y	e	ø	–	e	e	–	ε	ε	ε
262	y	y	y	e	y	y	y	y	y	e	ø	y	–	ø	–	ε	e	–	ε	–	ɔ*
220	y	e	y	y	y	y	y	ø	y	e	ø	ø	ø	ø	–	ε	–	–	ε	ε	–
165	y	y	y	y	y	y	y	y	–	e	ø	ø	–	–	–	ε	ε	–	ε	–	–
131	y	–	ø	e	–	ø	–	ø	ø	e	ø	–	ε	–	–	ε	ε	–	ε	a	a
	131	165		220			262			330			440			523			659		
	m	m	w	m	w	c	m	w	c	m	w	c	m	w	c	m	w	c	m	w	c
	/u/																				
659	u	u	i*	u	u	u	u	u	u	u	u	u	u	u	u	u	u	u	–	u	u
523	u	–	i*	u	u	u	u	u	u	u	u	u	u	u	u	u	u	u	a	o	o
440	u	u	i*	u	u	u	u	u	u	u	u	u	u	u	u	o	o	–	a	o	o
330	u	u	i*	u	u	u	u	u	u	u	u	u	o	o	o	o	o	o	a	o	o
262	u	u	u	u	u	u	u	u	u	u	u	u	o	u	o	–	ɔ	–	–	o	ɔ
220	u	u	u	u	u	u	u	–	u	u	–	u	ɔ	o	o	ɔ	–	o	a	ɔ	ɔ
165	u	u	u	u	u	u	u	u	u	u	–	–	–	o	ɔ	–	ɔ	–	a	ɔ	ɔ
131	u	u	u	u	o	u	u	u	u	u	–	o	–	o	o	–	–	–	a	ɔ	ɔ
	131	165		220			262			330			440			523			659		
	m	m	w	m	w	c	m	w	c	m	w	c	m	w	c	m	w	c	m	w	c
Speaker groups and fo of the natural sounds (m=man, w=woman, c=child)																					

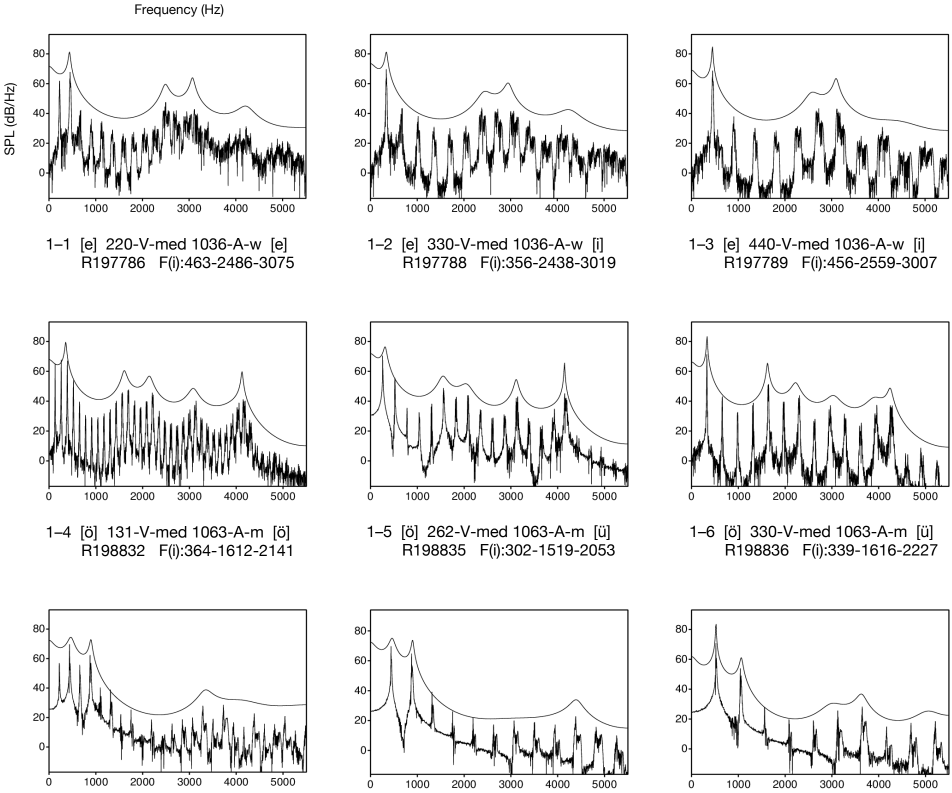
Table 1 (continuation). [C03-02-T01]

fo of vowel resynthesis (in Hz)	Vowel recognition																				
	/e/																				
659	y	y	i	y	i	i	y	i	i	y	i	i	y	–	–	e	e	e	–	e	e
523	i	y	i	y	i	i	y	i	–	e	e	e	ε	e	e	e	e	e	ε	ε	e
440	y	y	i	i	i	i	y	i	–	e	e	e	e	e	e	e	e	e	ε	ε	ε
330	i	e	i	e	i	e	e	e	e	e	e	e	ε	e	e	e	–	ε	ε	ε	ε
262	e	e	e	e	e	e	e	e	e	ε	e	ε	ε	e	–	ε	ε	–	ε	ε	ε
220	e	e	e	e	e	e	–	ε	e	ε	e	ε	ε	e	e	ε	ε	ε	–	ε	ε
165	e	e	e	e	e	e	–	e	e	ε	–	ε	ε	ε	–	ε	ε	–	–	ε	ε
131	e	e	e	e	e	e	–	ε	–	ε	–	ε	ε	e	e	ε	ε	ε	–	ε	ε
	131	165		220			262			330			440			523			659		
	m	m	w	m	w	c	m	w	c	m	w	c	m	w	c	m	w	c	m	w	c
	/ø/																				
659	y	–	y	y	y	y	–	y	y	–	–	y	–	y	y	ø	–	–	–	ø	–
523	y	y	y	y	y	y	y	y	y	y	y	y	ø	–	–	ø	ø	ø	–	ε	–
440	y	y	y	y	y	y	y	y	y	ø	ø	ø	–	ø	ø	ø	–	ø	–	ε	ø
330	y	–	ø	ø	–	y	ø	ø	ø	ø	ø	ø	–	ø	ø	ø	–	ε	ɔ*	ε	–
262	y	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	–	–	–	–	–	ε	ε	ɔ*	ε	–
220	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ø	ε	ø	ø	ε	–	ε	a	–	a
165	ø	ø	ø	ø	ø	ø	ø	ø	–	–	ø	–	–	ε	–	ε	–	ε	–	–	a
131	ø	ø	ø	ø	ø	ø	–	–	o*	–	ø	–	ε	ø	–	–	–	ε	a	a	a
	131	165		220			262			330			440			523			659		
	m	m	w	m	w	c	m	w	c	m	w	c	m	w	c	m	w	c	m	w	c
	/o/																				
659	–	u	u	u	u	u	u	u	u	u	u	u	u	u	u	–	ɔ	–	o	–	o
523	–	u	–	u	u	u	u	u	u	–	u	–	o	–	o	o	o	o	a	a	–
440	–	u	u	u	u	u	o	u	u	o	–	u	o	o	o	ɔ	o	o	ɔ	a	a
330	–	u	u	o	u	u	o	o	u	o	o	o	o	ɔ	o	–	–	ɔ	ɔ	a	a
262	o	u	u	o	o	u	o	o	o	ɔ	–	ɔ	ɔ	ɔ	ɔ	–	–	–	ɔ	a	a
220	o	–	u	o	o	u	o	o	o	ɔ	o	–	ɔ	ɔ	ɔ	a	ɔ	–	–	a	a
165	o	o	u	o	o	o	o	o	o	ɔ	o	–	ɔ	ɔ	ɔ	a	a	ɔ	a	a	a
131	o	o	u	–	o	–	–	o	o	ɔ	o	–	a	a	ɔ	a	a	–	ɔ	a	a
	131	165		220			262			330			440			523			659		
	m	m	w	m	w	c	m	w	c	m	w	c	m	w	c	m	w	c	m	w	c
Speaker groups and fo of the natural sounds (m=man, w=woman, c=child)																					

Table 1 (continuation). [C03-02-T01]

fo of vowel resynthesis (in Hz)	Vowel recognition																					
	/ɛ/																					
	659	–	y	y	e	i	ɛ	y	e	–	–	e	ɛ	–	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ
	523	e	–	–	ɛ	e	ɛ	–	e	ɛ	ɛ	e	ɛ	ɛ	ɛ	ɔ*	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ
	440	e	–	i	ɛ	e	ɛ	ɛ	e	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ
	330	e	–	–	ɛ	e	ɛ	ɛ	e	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ
	262	ɛ	–	e	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	–	ɛ	ɛ	–	ɛ	–	ɛ
	220	ɛ	ɛ	e	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	–	ɛ	ɛ	ɛa	ɛ	ɛ	–
	165	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	a	–	ɛ	–	a	–	–	–
	131	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	a	ɛ	a	ɛa	ɛ	ɛ	a
		131	165		220			262			330			440			523			659		
		m	m	w	m	w	c	m	w	c	m	w	c	m	w	c	m	w	c	m	w	c
	/a/																					
	659	a	a	o	–	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a
	523	a	a	a	a	a	a	a	–	a	a	a	a	a	a	a	a	a	a	ɛ	a	a
	440	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	–	a	a
	330	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a
	262	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a
	220	a	a	–	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a
	165	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a
	131	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	–	a	a	–	a	a
		131	165		220			262			330			440			523			659		
		m	m	w	m	w	c	m	w	c	m	w	c	m	w	c	m	w	c	m	w	c
Speaker groups and fo of the natural sounds (m=man, w=woman, c=child)																						

Figure 1. Source–filter resynthesis based on spectral envelopes of single natural reference sounds: First illustration of occurring vowel quality shifts in an open–close direction due to an increase in fo from the level of the reference sound to higher levels. [C-03-02-F01]



1–1 [e] 220-V-med 1036-A-w [e]
R197786 F(i):463-2486-3075

1–2 [e] 330-V-med 1036-A-w [i]
R197788 F(i):356-2438-3019

1–3 [e] 440-V-med 1036-A-w [i]
R197789 F(i):456-2559-3007

1–4 [ö] 131-V-med 1063-A-m [ö]
R198832 F(i):364-1612-2141

1–5 [ö] 262-V-med 1063-A-m [ü]
R198835 F(i):302-1519-2053

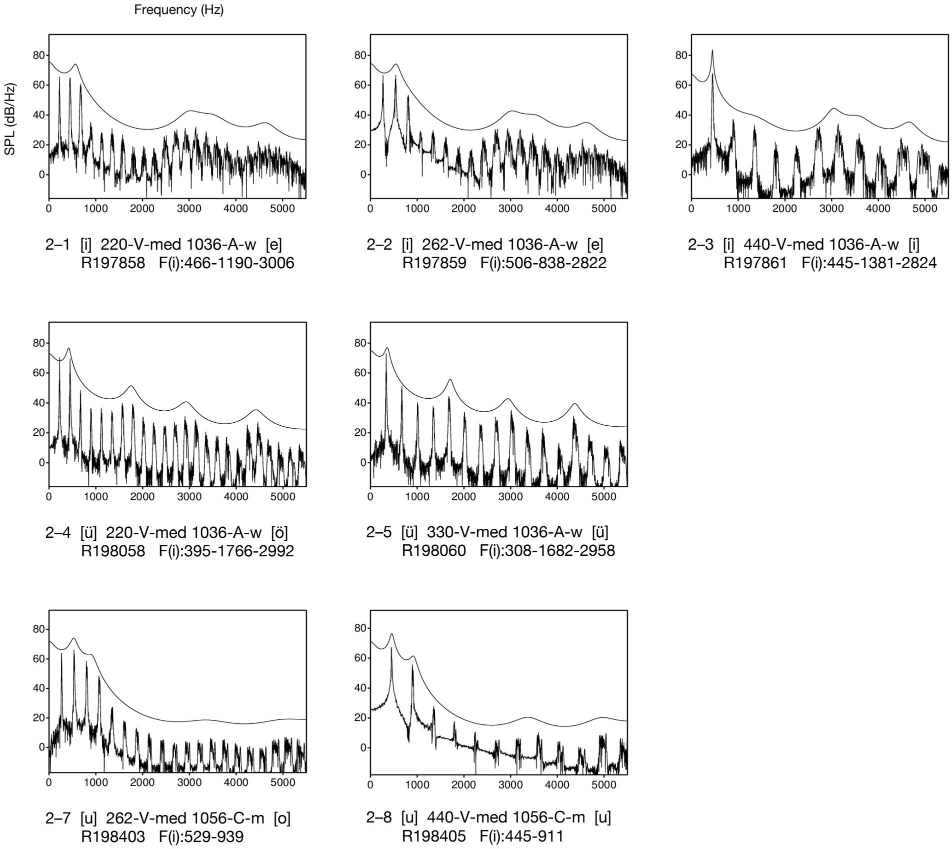
1–6 [ö] 330-V-med 1063-A-m [ü]
R198836 F(i):339-1616-2227

1–7 [o] 220-V-med 1036-A-w [o]
R197914 F(i):431-898

1–8 [o] 440-V-med 1036-A-w [u]
R197917 F(i):432-885

1–9 [o] 523-V-med 1036-A-w [u]
R197918 F(i):508-1062

Figure 2. Source–filter resynthesis based on spectral envelopes of single natural reference sounds: Second illustration of occurring vowel quality shifts in an open–close direction due to an increase in fo from a low level up to the level of the reference sound. [C-03-02-F02]



2–1 [i] 220-V-med 1036-A-w [e]
R197858 F(i):466-1190-3006

2–2 [i] 262-V-med 1036-A-w [e]
R197859 F(i):506-838-2822

2–3 [i] 440-V-med 1036-A-w [i]
R197861 F(i):445-1381-2824

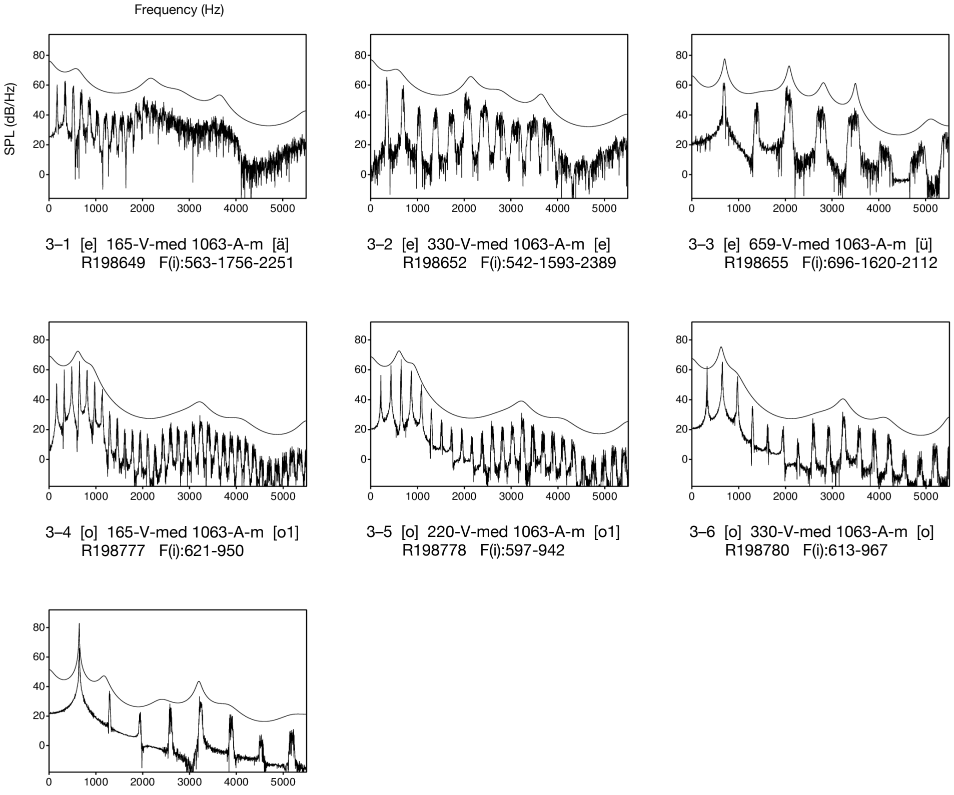
2–4 [ü] 220-V-med 1036-A-w [ö]
R198058 F(i):395-1766-2992

2–5 [ü] 330-V-med 1036-A-w [ü]
R198060 F(i):308-1682-2958

2–7 [u] 262-V-med 1056-C-m [o]
R198403 F(i):529-939

2–8 [u] 440-V-med 1056-C-m [u]
R198405 F(i):445-911

Figure 3. Source–filter resynthesis based on spectral envelopes of single natural reference sounds: Third illustration of occurring vowel quality shifts in an open–close direction due to an increase in fo from lower to higher levels, the shifts including adjacent and non-adjacent vowel qualities. [C-03-02-F03]



3–1 [e] 165-V-med 1063-A-m [ä]
R198649 F(i):563-1756-2251

3–2 [e] 330-V-med 1063-A-m [e]
R198652 F(i):542-1593-2389

3–3 [e] 659-V-med 1063-A-m [ü]
R198655 F(i):696-1620-2112

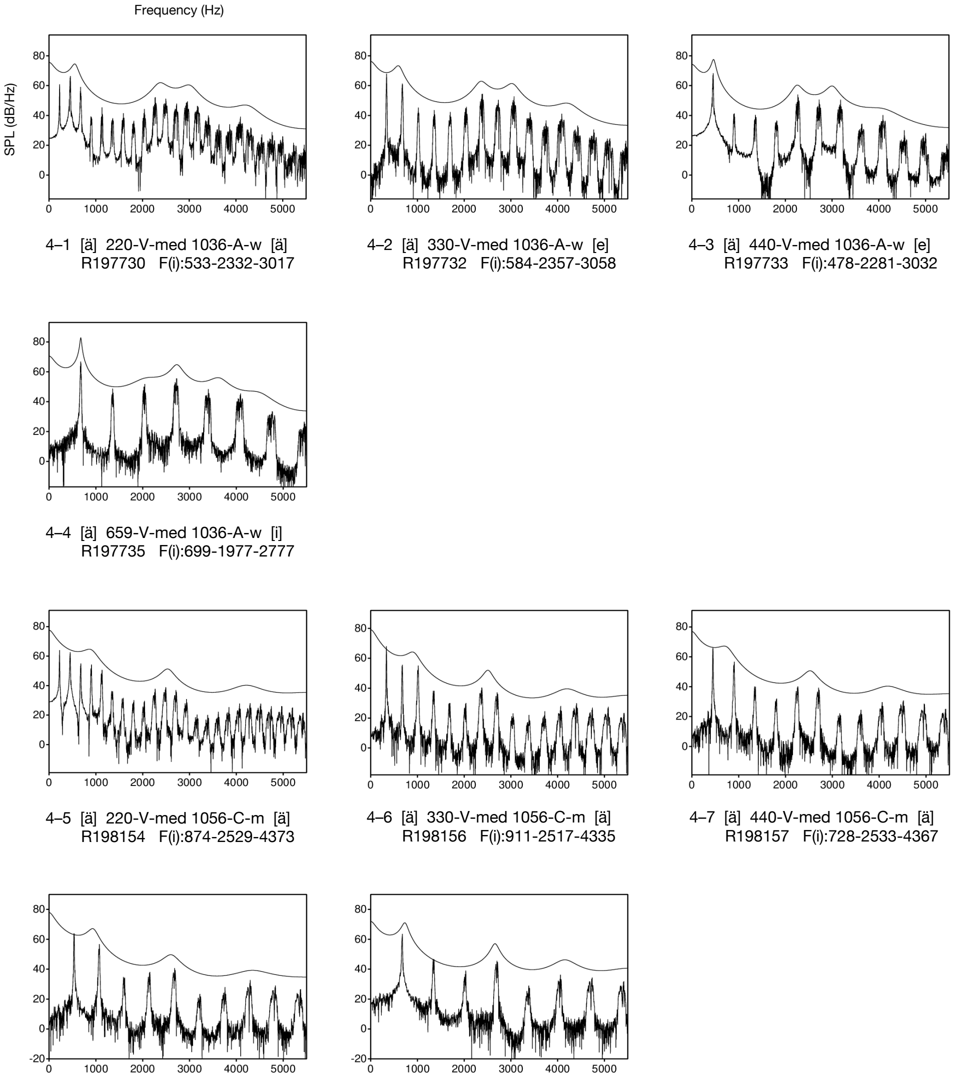
3–4 [o] 165-V-med 1063-A-m [o1]
R198777 F(i):621-950

3–5 [o] 220-V-med 1063-A-m [o1]
R198778 F(i):597-942

3–6 [o] 330-V-med 1063-A-m [o]
R198780 F(i):613-967

3–7 [o] 659-V-med 1063-A-m [u]
R198783 F(i):638-1049

Figure 4. Source–filter resynthesis based on spectral envelopes of single natural reference sounds: Illustration of the nonuniform character of occurring vowel quality shifts related to fo variation. [C-03-02-F04]



4–1 [ä] 220-V-med 1036-A-w [ä]
R197730 F(i):533-2332-3017

4–2 [ä] 330-V-med 1036-A-w [e]
R197732 F(i):584-2357-3058

4–3 [ä] 440-V-med 1036-A-w [e]
R197733 F(i):478-2281-3032

4–4 [ä] 659-V-med 1036-A-w [i]
R197735 F(i):699-1977-2777

4–5 [ä] 220-V-med 1056-C-m [ä]
R198154 F(i):874-2529-4373

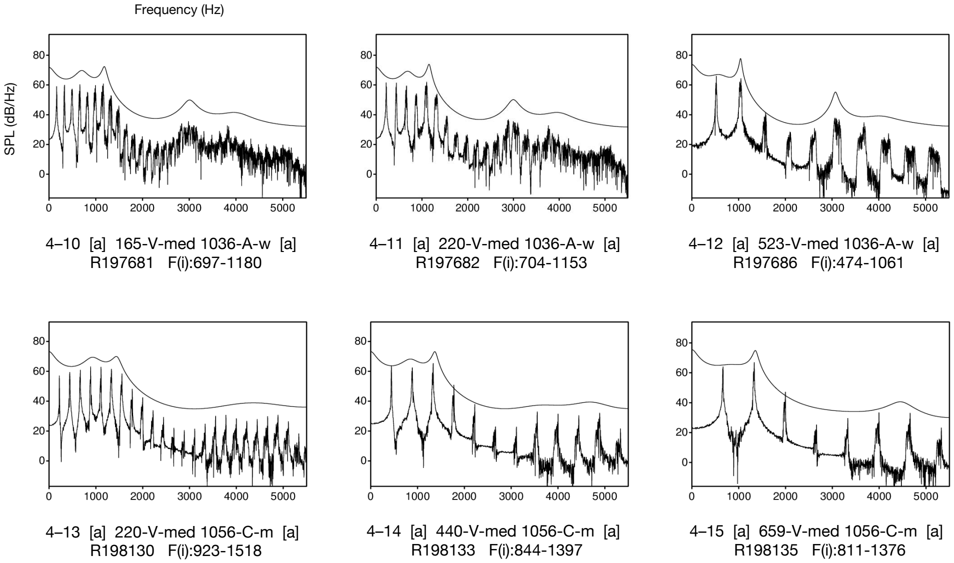
4–6 [ä] 330-V-med 1056-C-m [ä]
R198156 F(i):911-2517-4335

4–7 [ä] 440-V-med 1056-C-m [ä]
R198157 F(i):728-2533-4367

4–8 [ä] 523-V-med 1056-C-m [ä]
R198158 F(i):909-2590-4348

4–9 [ä] 659-V-med 1056-C-m [ä]
R198159 F(i):718-2652-4347

Figure 4 (continuation). [C-03-02-F04]



4–10 [a] 165-V-med 1036-A-w [a]
R197681 F(i):697-1180

4–11 [a] 220-V-med 1036-A-w [a]
R197682 F(i):704-1153

4–12 [a] 523-V-med 1036-A-w [a]
R197686 F(i):474-1061

4–13 [a] 220-V-med 1056-C-m [a]
R198130 F(i):923-1518

4–14 [a] 440-V-med 1056-C-m [a]
R198133 F(i):844-1397

4–15 [a] 659-V-med 1056-C-m [a]
R198135 F(i):811-1376

3.3 Source–Filter Synthesis Based on Model Filter Patterns of Long Standard German Vowels, Including Variation of f_o

In a third type of experimentation, a model synthesis was conducted that allows for a straightforward replication and testing of vowel quality shifts that occur with rising f_o. The basic idea of the model synthesis was to create interrelated filter patterns and f_o levels in terms of all filter frequencies of a pattern being multiples (in whole numbers) of f_o for two or three f_o levels. These models represent "ideal" cases of filter curves and harmonic spectra, the dominant harmonics always coinciding with the filters, and the only acoustic differences between the sounds being f_o (and pitch) and frequency distances of harmonics.

Three vowel synthesis experiments were conducted using a Klatt synthesiser (parallel mode, steady-state sounds of 1 sec.). In the first experiment, model $F1$–$F2$–$F3$ patterns for sounds of the back vowels /o, ɔ/ and the front vowels /e, ø, ɛ/ were created, in approximate relation to observed F-patterns of natural sounds produced at an f_o level of c. 200 Hz. Approximations were based on an extensive analysis of sounds of these vowels in the Zurich Corpus. For the synthesis of close-mid vowel sounds produced at 200 Hz, the filter frequencies were set as multiples of 400 Hz, and for open-mid vowel sounds, the filter frequencies were set as multiples of 600 Hz. The formant bandwidths and formant levels were set to bring the resulting sound spectra into line with the observed spectra of natural sounds of the vowels in question, as documented in the Zurich Corpus. (Two higher formants with low levels were also added to smoothen the higher frequencies > 3.5 kHz; for details, see Chapter M3.3.) For F-patterns with filter frequencies as multiples of 400 Hz, the two f_o levels 200 and 400 Hz were investigated in synthesis. For F-patterns with filter frequencies as multiples of 600 Hz, the three f_o levels of 200, 300 and 600 Hz were investigated in synthesis.

In the second experiment, this synthesis was repeated with f_o levels halved, that is, applying f_o levels of 100 and 200 Hz and 100, 150 and 300 Hz, respectively.

In the third experiment, the synthesis was again repeated with the original f_o levels of experiment 1 but with the levels of the first two filters altered to the following settings: L_{F1} = -10 dB and L_{F2} = +10 dB, and L_{F1} = -20 dB and L_{F2} = +10 dB.

Vowel recognition of the synthesised sounds was examined for each of the three experiments separately according to the standard procedure of the Zurich Corpus and including the five standard listeners.

The model *F*1–*F*2–*F*3 patterns investigated, the f_o levels of experiment 1 and the main vowel recognition results for the synthesised sounds are shown in Table 1. Recognised vowel qualities with a labelling majority (recognition rate ≥ 60% for vowel openness) are given for the sounds.

The results of the first synthesis experiment showed consistent vowel quality shifts in an open–close direction with increasing f_o for all sound pairs and sound triples tested. Shifts to adjacent vowel qualities /o–u/, /e–y-i/ and /ø–y/ were found for *F*-patterns as multiples of 400 Hz and f_o variation of 200–400 Hz, and shifts to adjacent and non-adjacent vowel qualities /ɔ–o–u/, /ɛ–e–y-i/ and /ɛ-ə–ø–y/ were found for *F*-patterns as multiples of 600 Hz and f_o variation of 200–300–600 Hz. Thus, based on these model *F*-patterns and f_o level variations, the relation of spectral envelope representation of vowel quality to f_o was again, and in a paradigmatic way, made evident by the vowel recognition results. For a corresponding illustration, see Figures 1 and 2.

However, as highlighted in the various previous experiments, the vowel quality shifts resulting from f_o variation were again found to be non-uniform in the present experiment. Above all, contrary to the seven-semitone or one-octave f_o variation of 200–300 Hz or 200–400 Hz in experiment 1, no vowel quality shifts were found in experiment 2 for the seven-semitone or one-octave f_o variation of 100–150 Hz or 100–200 Hz. This result indicated that not only the extent but also the frequency range of f_o variation had an impact on vowel quality shifts. In addition, vowel recognition was also influenced by the level ratio of the investigated filters (although to a limited degree), above all for synthesised sounds related to *F*-patterns of natural sounds of back vowels (compare the vowel recognition results of experiments 1 and 3 in Table 1).

As said, the general aim of this type of experimentation was to contribute to the creation of model experiments that allow for a simple verification of a perceptual change in recognised vowel quality caused by f_o changes only, keeping the resonance curve of vowel production unchanged and applying different f_o levels in a way that allows for an exact spectral sampling of resonance frequencies for all sounds. At the same time, this type of experimentation also highlights the nonuniform character of the effect of f_o variation.

For details of experimental design, method and results, an extended discussion and documentation of results (tables including sound links), see the Materials, Chapter M3.3.

Table 1: Source–filter synthesis based on model filter patterns of Standard German back and front vowels, including f_o variation in synthesis: Model *F*-patterns and f_o variation investigated and vowel recognition results (summary). Columns 1 to 13 = vowel synthesis (VO = vowel openness; S = sound series; fo = f_o levels applied in synthesis in experiment 1, in Hz; Δfo = f_o level differences in reference to the first sound of a series, in semitones, ST; F(i), L(i) and B(i) = frequencies, levels and bandwidths of the formant patterns used in synthesis); Column 14 = vowel recognition result for experiment 1. Column 15 = vowel recognition result for experiment 2. Columns 16 and 17 = vowel recognition result for experiments 3a and 3b (two level variations, see text). Vowel recognition without a labelling majority is given as "–". Colour code: Red = vowel quality shifts related to f_o variation. Note that the recognition results are given in relation to the openness of the vowels (see Series 3, 4 and 6).
[C-03-03-T01]

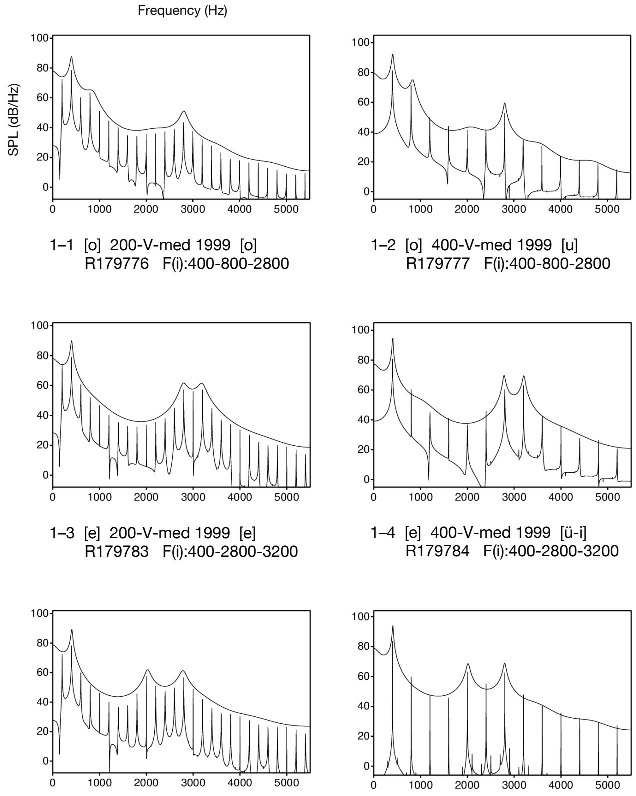
Figure 1: Sound pairs related to the model *F*-patterns 1, 4 and 5 presented in Table 1, with f_o variation of 200–400 Hz applied in synthesis: Illustration of open–close vowel quality shifts due to an increase in f_o, involving adjacent qualities. For the sounds of the model *F*-patterns shown, increasing f_o by one octave within this frequency range resulted in vowel quality shifts /o–u/, /e–y-i/, /ø–y/.
[C-03-03-F01] ↗

Figure 2: Sound triplets related to the model *F*-patterns 2 and 6 presented in Table 1, with f_o variation of 200–300–600 Hz applied in synthesis: Illustration of open–close vowel quality shifts due to an increase in f_o, involving adjacent and non-adjacent qualities. For the sounds of the model *F*-patterns shown, increasing f_o by 7 and 19 semitones within this frequency range resulted in vowel quality shifts of /ɔ–o–u/ and /ɛ–e–y-i/.
[C-03-03-F02] ↗

Table 1. Source–filter synthesis based on model filter patterns of Standard German back and front vowels, including fo variation in synthesis: Model F-patterns and fo variation investigated and vowel recognition results (summary). [C-03-03-T01]

Vowel synthesis (fo and F-patterns)													Vowel recognition			
VO	**S**	**fo**	**Δfo**	**F1**	**L1**	**B1**	**F2**	**L2**	**B2**	**F3**	**L3**	**B3**	**Experiments 1 to 3**			
		Hz	**ST**	**Hz**	**dB**	**Hz**	**Hz**	**dB**	**Hz**	**Hz**	**dB**	**Hz**	**1**	**2**	**3a**	**3b**
back	1	200	ref	400	100	100	800	105	100	2800	90	200	o	o	o	o
		400	12										u	o	o	o
	2	200	ref	600	100	100	1200	95	100	3000	85	200	ɔ	a	a	a
		300	7										o	a	o	o
		600	19										u	o	–	o
front	3	200	ref	400	100	100	2400	100	200	2800	100	200	e	e	e	e
		400	12										y-i	e	i	i
	4	200	ref	400	100	100	2800	100	200	3200	100	200	e	e	e	e
		400	12										y-i	e	y-i	y-i
	5	200	ref	400	100	100	2000	100	150	2800	100	200	ø	ø-e	ø	ø
		400	12										y	ø	y	y
	6	200	ref	600	100	100	2400	100	200	3000	100	200	ɛ	ɛ	ɛ	ɛ
		300	7										e	ɛ	e	e
		600	19										y-i	e	y-i	y-i
	7	200	ref	600	100	100	1800	100	150	3000	100	200	ə-ɛ	ɛ	ɛ	ɛ
		300	7										ø	ɛ	ø	ø
		600	19										y	ø	y	–

Figure 1. Sound pairs related to the model F-patterns 1, 4 and 5 presented in Table 1, with fo variation of 200–400 Hz applied in synthesis: Illustration of open–close vowel quality shifts due to an increase in fo, involving adjacent qualities. [C-03-03-F01]



1–1 [o] 200-V-med 1999 [o]
R179776 F(i):400-800-2800

1–2 [o] 400-V-med 1999 [u]
R179777 F(i):400-800-2800

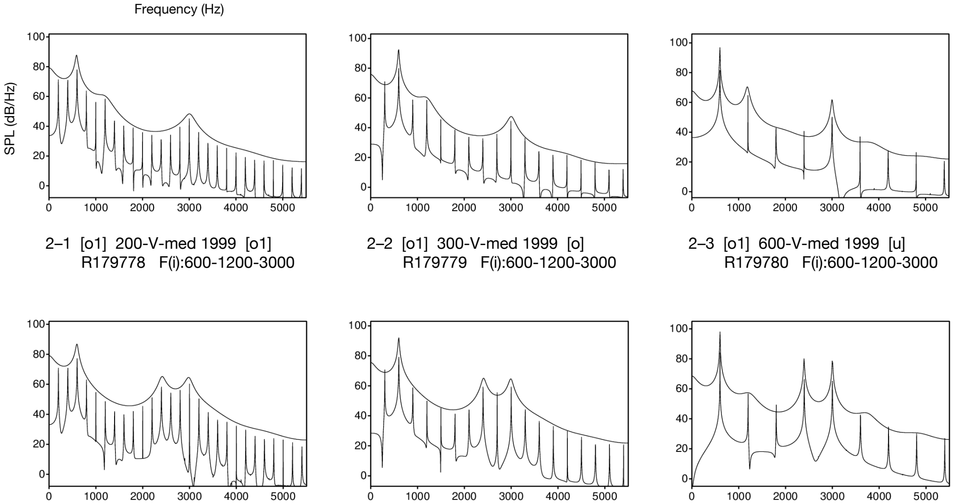
1–3 [e] 200-V-med 1999 [e]
R179783 F(i):400-2800-3200

1–4 [e] 400-V-med 1999 [ü-i]
R179784 F(i):400-2800-3200

1–5 [ö] 200-V-med 1999 [ö]
R179785 F(i):400-2000-2800

1–6 [ö] 400-V-med 1999 [ü]
R179786 F(i):400-2000-2800

Figure 2. Sound triplets related to the model F-patterns 2 and 6 presented in Table 1, with fo variation of 200–300–600 Hz applied in synthesis: Illustration of open–close vowel quality shifts due to an increase in fo, involving adjacent and nonadjacent qualities. [C-03-03-F02]



2–1 [o1] 200-V-med 1999 [o1]
R179778 F(i):600-1200-3000

2–2 [o1] 300-V-med 1999 [o]
R179779 F(i):600-1200-3000

2–3 [o1] 600-V-med 1999 [u]
R179780 F(i):600-1200-3000

2–4 [ä] 200-V-med 1999 [ä]
R179787 F(i):600-2400-3000

2–5 [ä] 300-V-med 1999 [e]
R179788 F(i):600-2400-3000

2–6 [ä] 600-V-med 1999 [ü-i]
R179789 F(i):600-2400-3000

3.4 Source–Filter Synthesis Based on Model Filter Patterns of Half-Open Tubes, Including Variation of f_o

Extending the previous experimentation, vowel synthesis based on model *F*-patterns was also related to three half-open tube resonance patterns commonly attributed to the three average vocal tract lengths of men, women and children, respectively: Steady-state sounds of 1 sec. were synthesised with F_1 set to 500, 600 or 700 Hz and with higher frequencies set to odd multiples of F_1 (Klatt synthesis, cascade mode, all bandwidths set to 100 Hz). For each of the three *F*-patterns, three f_o levels of 1/3, 1/2 and 1/1 of the first filter frequency were investigated. For the resulting synthesis, again, the frequencies of the dominant harmonics always coincided with the filter frequencies, and the sounds only differed in f_o (and pitch) and frequency distances of the harmonics. The vowel recognition of the synthesised sounds was examined in an experiment-specific listening test (for the test conditions, see Chapter M3.4).

The investigated *F*1–*F*2–*F*3–*F*4–*F*5 patterns and the main results of vowel recognition of the synthesised sounds are shown in Table 1. Recognised vowel qualities with a labelling majority (recognition rate ≥ 60%) are given for the sounds. According to these results, a vowel synthesis based on half-open tube resonance patterns approximating the resonances of the average vocal tract lengths of men, women and children combined with speaker group-specific f_o levels as given in formant statistics, that is, 1/3 of the first filter frequency, produced either a schwa sound (adults) or an /ɛ/–like sound (children). Yet, comparable to the results of the first experiment discussed in the previous chapter, the recognised vowel quality shifted in an open–close direction with increasing f_o, with shifts to an adjacent vowel quality associated with a one-octave or seven-semitone increase in f_o for sounds related to the *F*-patterns of adults, and shifts to a non-adjacent vowel quality associated with an f_o increase exceeding one octave for sounds related to all three *F*-patterns. Thus, the ambiguity of *F*-patterns and spectral envelopes also markedly affected half-open tube resonance patterns: The findings indicated that these patterns, commonly assumed to relate to neutral or centralised articulatory configurations, are not recognised consistently as neutral schwa vowels.

Some differences were found for synthesised sounds related to the *F*-patterns of adults and children: Above all, the sound synthesis based on the *F*-pattern of children and the lowest f_o level was not recognised as a clear schwa but rather as an /ɛ/–like sound, and only an increase of f_o by 19 semitones caused a pronounced vowel quality shift in an open–close direction. This finding again points to the nonuniform relation between vowel quality and spectral characteristics since the results depended on the *F*-patterns and the frequency ranges of f_o variation.

For references, extended background information, details of experimental design, method and results, an extended discussion and documentation of results (tables including sound links), see the Materials, Chapter M3.4.

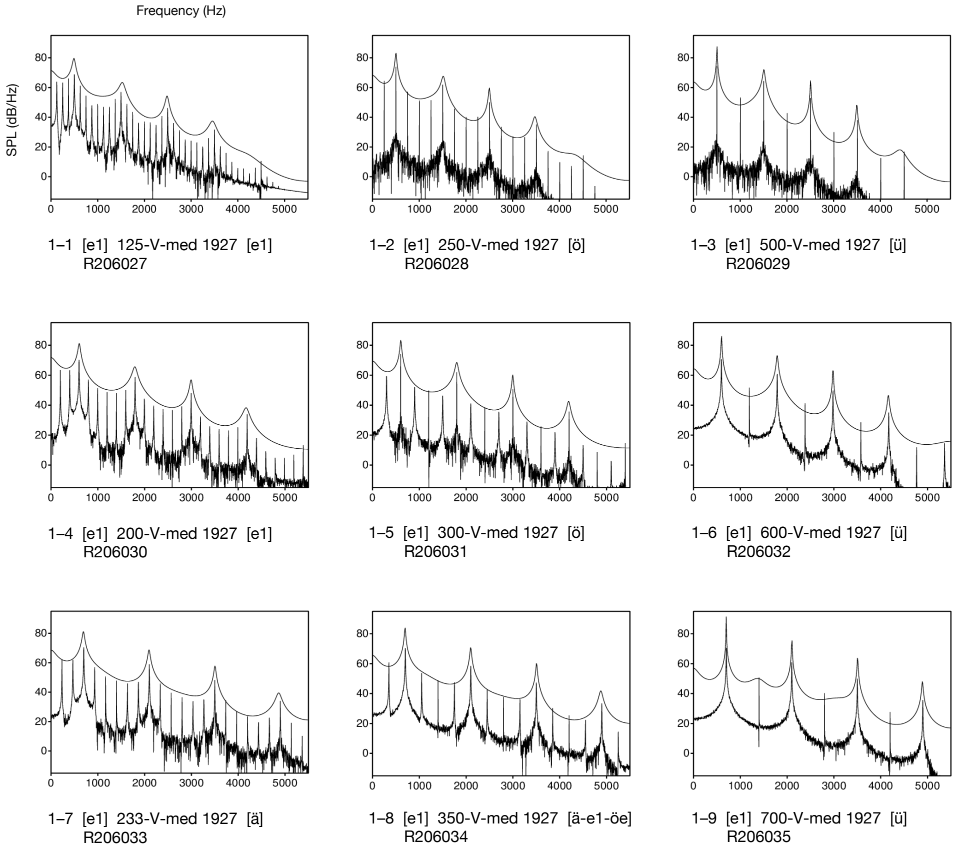
Table 1. Source–filter synthesis based on model filter patterns of half-open tubes, including f_o variation in synthesis: *F*-patterns and f_o variation investigated and vowel recognition results (summary). Columns 1 to 8 = vowel synthesis (SG = speaker group commonly related to the resonance pattern investigated; fo = level of f_o applied in synthesis; Δfo = f_o level differences in reference to the first sound of a series, in semitones, ST, with approximations given in parenthesis; F(i) = *F*-patterns investigated). Column 9 = vowel recognition results for recognition rates ≥ 60%. Note that the sound related to the *F*-pattern of children with f_o of synthesis of 350 Hz was somewhat confused, and the three labelled vowel qualities are shown.
[C-03-04-T01]

Figure 1: Source–filter synthesis based on model filter patterns of half-open tubes, including f_o variation in synthesis: Illustration of the main finding. Spectra (0–5.5 kHz) and filter curves of the three sound triplets, as listed in Table 1, are shown.
[C-03-04-F01]

Table 1. Source–filter synthesis based on model filter patterns of half-open tubes, including fo variation in synthesis: F-patterns and fo variation investigated and vowel recognition results (summary). [03-04-T01]

Vowel synthesis								**Vowel recognition ≥ 60%**
SG	**f_o**	**Δf_o**	**F_1**	**F_2**	**F_3**	**F_4**	**F_5**	
	Hz	**ST**	**Hz**	**Hz**	**Hz**	**Hz**	**Hz**	
men	125	ref	500	1500	2500	3500	4500	ə
	250	12						ø
	500	24						y
women	200	ref	600	1800	3000	4200	5400	ə
	300	(7)						ø
	600	(19)						y
children	233	ref	700	2100	3500	4900	6300	ε
	350	(7)						ε–ə–øe boundary
	700	(19)						y

Figure 1. Source–filter synthesis based on model filter patterns of half-open tubes, including fo variation in synthesis: Illustration of the main finding. [C-03-04-F01]



1–1 [e1] 125-V-med 1927 [e1]
R206027

1–2 [e1] 250-V-med 1927 [ö]
R206028

1–3 [e1] 500-V-med 1927 [ü]
R206029

1–4 [e1] 200-V-med 1927 [e1]
R206030

1–5 [e1] 300-V-med 1927 [ö]
R206031

1–6 [e1] 600-V-med 1927 [ü]
R206032

1–7 [e1] 233-V-med 1927 [ä]
R206033

1–8 [e1] 350-V-med 1927 [ä-e1-öe]
R206034

1–9 [e1] 700-V-med 1927 [ü]
R206035

3.5 Paradigmatic Examples of Formant Pattern and Spectral Shape Ambiguity in Natural Vocalisations and Their Resynthesised Replicas

Neither *F*-patterns nor spectral envelopes were found to be stable within the vowel-related frequency ranges when looking at natural sounds of single vowels produced at different f_o levels, even for sounds produced by a single speaker (see Chapter 2). In a vowel synthesis or resynthesis that was based on one single unchanged *F*-pattern and related LPC filter curve or one single unchanged spectral envelope, depending on the experimental setting, the recognised vowel quality for a substantial portion or even the majority of sounds changed if f_o was substantially altered (see Chapters 3.1 to 3.4). These findings led to the conclusion that *F*-patterns and spectral envelopes *per se* are ambiguous acoustic representations of vowel quality.

In the preceding Chapters 3.1 to 3.4, the ambiguity became evident in synthesis and resynthesis experiments. In the present chapter, the ambiguity is addressed as a core phenomenon of natural vowel sounds: Against the background of earlier studies on natural, resynthesised and synthesised vowel sounds and of the preceding experiments and results, and on the basis of the new Zurich Corpus, paradigmatic series of voiced vowel sounds produced by children, women and men in V context were compiled and are documented below, providing evidence for formant pattern and spectral shape ambiguity in natural vocalisations, according to the following design of selection and documentation.

Sound pairs of two adjacent vowels /e–i/, /ø–y/ and /o–u/ were compiled according to the following criteria:

- Both sounds of a pair were produced by the same speaker.
- For both sounds of a pair, the upper limit of calculated f_o was 400 Hz for men, 450 Hz for women and 500 Hz for children, reflecting the average everyday vocal range of women and children and the chest and “mixed” voice for men.
- A 100% vowel recognition rate matching vowel intention was obtained in the standard listening test conducted when creating the Zurich Corpus.
- Vowel-related peaks of the harmonic spectra, spectral envelopes and estimated *F*-patterns of both sounds were appraised as being comparable; above all, differences of estimated vowel quality-related spectral peaks and estimated *F*-patterns were considered to remain within the commonly assumed range of variation for sounds of a single vowel quality.

For the entire f_o range of recognisable vowel sounds, sound triplets of adjacent and non-adjacent vowels /ɛ–e–i/ or /ɛ–e–y/, /a–o–u/ and /ɔ–o–u/ were compiled according to the following criteria:

- All three sounds were produced by speakers of a single speaker group (children, women or men).
- An 80–100% vowel recognition rate matching vowel intention was obtained in the standard listening test conducted when creating the Zurich Corpus. (Note that for these adjacent and non-adjacent vowels produced by speakers of a single age- and gender-related speaker group, the compilation of sound triplets with similar vowel-related spectral peaks, *F*-patterns and spectral envelopes proved to be much more difficult than merely compiling sound pairs of adjacent vowels; therefore, f_o was not limited and vowel sounds with recognition rates of 80–100% were investigated.)
- Vowel-related peaks of the harmonic spectra, spectral envelopes and estimated *F*-patterns of all three sounds were appraised as comparable.

Note that the vowel /ɔ/ was also included in the investigation in order to examine possible *F*-pattern and spectral shape ambiguity for the non-adjacent back vowels /ɔ/ and /u/. Note also that production style and vocal effort of the sounds were disregarded.

Acoustic analysis of the sounds accorded to the standard procedure of the Zurich Corpus. For sounds of front vowels, the spectral comparison was related to a frequency range of up to 3 kHz for adults and up to 3.5 kHz for children. For sounds of back vowels and /a/, the spectral comparison was related to a frequency range of up to 2 kHz for all speakers.

In the first step, an extensive sound sample was compiled, consisting of several sound comparisons related to the vowel pairs and triplets investigated. In the second step, this sample was reduced for exemplary documentation of the ambiguity phenomenon in natural vocalisations in this treatise: For each of the pairs of the adjacent vowels /e–i/, /e–y/ and /o–u/ and each of the triplets of the adjacent and non-adjacent vowels /ɛ–e–i/ or /ɛ–e–y/, /a–o–u/ and /ɔ–o–u/, three sound comparisons were selected, resulting in a total of nine sound pairs and nine sound triplets of natural sounds or 45 natural and 117 resynthesised sounds in total. (For the numerical indications of the *F*-patterns, see Chapter M3.5.)

For these selected sounds compared within a pair or triplet, Klatt synthesis (cascade mode, steady-state sounds of 1 sec.) was applied to further examine the ambiguity of the estimated *F*-patterns. For every single *F*-pattern of a natural sound, two or three f_o levels (taken from the opposing sounds of the pair or triplet in question) were applied in resynthesis according to three reflections: Firstly, suppose a vowel quality of a natural reference sound is maintained in a resynthesis that is based on both its estimated LPC curve and its average f_o level. In this case, the maintained vowel quality can be considered a possible validation criterion for the LPC curve and, thus, for the related estimated *F*-pattern and spectral envelope. Secondly, suppose an increase or decrease of the f_o level applied in resynthesis causes a vowel quality shift while keeping the LPC curve unchanged. In that case, the LPC curve is indicated to be an ambiguous representation of vowel quality. Thirdly, suppose the same vowel quality is recognised for the two or three sounds of a sound pair or triplet, the sounds resynthesised based on the two or three LPC curves but applying equal f_o levels. In that case, vowel recognition validates an assessment of spectral peak pattern, *F*-pattern and spectral envelope similarity to the natural reference sounds of different vowels that stand in comparison.

Therefore, in a separate listening test involving the five standard listeners of the Zurich Corpus, vowel recognition of the resynthesised sounds was investigated according to the standard procedure of the corpus, except for the condition of sound presentation: Instead of featuring single sounds, each test item consisted of two sounds, the natural reference sound (first sound) and one of the resynthesised replicas of the sound pair or triplet (second sound), resulting in two test items per pair and three items per triplet. The two sounds of a test item were presented (separated by a 1 sec. pause), and the listeners were asked to assign the recognised vowel quality of the second sound only.

According to the results of acoustic analysis (estimation of *F*-patterns; for details, see Chapter M3.5, Table 1), for all natural sounds of the sound pairs, the F_1 difference between the two *F*-patterns was < 35 Hz and the F_2 difference was < 100 Hz, except for one pair. The F_3 difference for the pairs of front vowel sounds was < 140 Hz. However, for all pairs of front vowel sounds, the difference in either F_2 or F_3 was < 60 Hz. Likewise, for all sounds of the sound triplets (for details, see Chapter M3.5, see Table 2), the F_1 (or lowest single peak frequency) difference was < 65 Hz, the F_2 difference was in the range of 28–146 Hz for the triplets of sounds of /a/ or /ɔ/ compared with /o/ and /u/, and the differences in F_2–F_3 for the triplets of front vowel sounds were 22–16, 214–145 and 170–144 Hz, respectively. On this basis, the F_1 differences for sounds of the two or three different vowels compared were markedly smaller than the differences of statistical F_1 of these two or three vowels generally given in the literature. Further, if the range of statistical F_2 and F_3 variation reported in the literature for sounds of a single vowel quality produced at a given single f_o level is taken into account (for reference, see Chapter M3.5), the differences of the entire F_1–F_2–F_3 patterns of the documented sounds were found within this statistical variation range for allophones. In these terms, the *F*-patterns of the natural sounds of different vowels compared were appraised as similar.

Resynthesis also confirmed the estimated similarity, above all for the sound pairs of close-mid and close vowels and the sound triplets of open-mid, close-mid and close vowels: For these sounds and vowels, in a resynthesis based on the calculated *F*-patterns and the calculated f_o of the natural reference sounds, the recognition rate for vowel quality or vowel openness matching vowel intention was ≥ 80%, with only two exceptions for which the rate dropped to 60%. For the sounds of /a–o–u/, the recognition rate was ≥ 80% for four and 60% for five sounds. In contrast, in a resynthesis based on the calculated *F*-patterns of the natural reference sounds but varying f_o in terms of applying the opposing f_o levels of a sound pair or triplet and then testing the vowel recognition of the resynthesised replicas, the recognised vowel qualities shifted for almost all replicas of the natural reference sounds investigated, with shifts being very pronounced for the sound pairs and triplets of close-mid and close and of open-mid, close-mid and close vowels. Again, a general open–close shift direction resulting from an increase of f_o from low to high levels was found, the shifts involving adjacent and non-adjacent vowel qualities for the sound triplets.

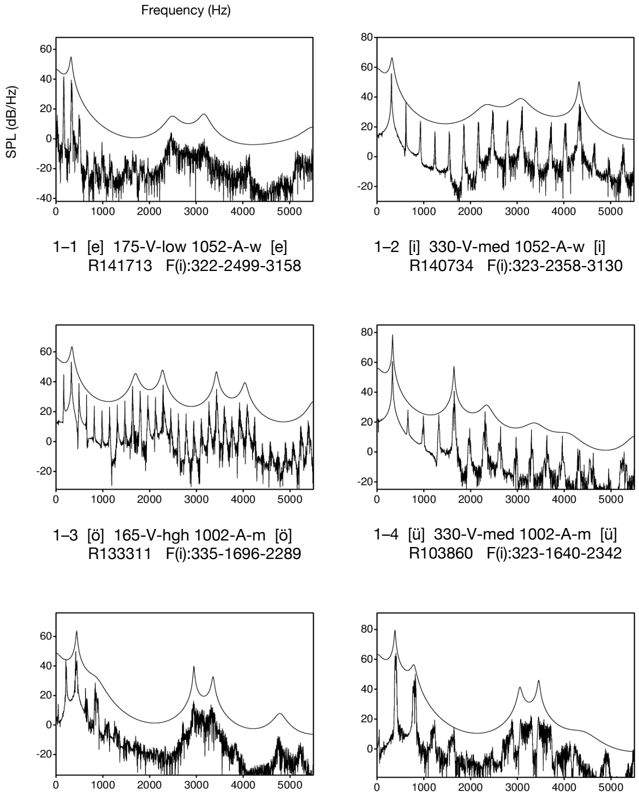
Based on such a form of comparison of natural vowel sounds and related highly restricted conditions for speakers and resynthesised replicas, the nine sound pairs and nine sound triplets documented provide paradigmatic examples of formant pattern and spectral shape ambiguity in natural vocalisations. As an example of the full documentation created, Figure 1 shows similar spectral peaks and/or estimated *F*-patterns and spectral envelopes for natural sound pairs of the adjacent vowels /e–i/, /ø–y/ and /o–u/, the sounds of a pair produced by a single speaker. Figure 2 shows similar spectral peaks and/or estimated *F*-patterns and spectral envelopes for natural sound triplets of the adjacent and non-adjacent vowels /ɛ–e–i/, /a–o–u/ and /ɔ–o–u/, all three sounds of a triplet produced by speakers of the same age- and gender-related speaker group.

For references, extended background information, details of experimental design, method and results, an extended discussion and documentation of results (tables including sound links), see the Materials, Chapter M3.5.

Figure 1: Formant pattern and spectral shape ambiguity in natural vocalisations: Similar estimated *F*-patterns and spectral envelopes for sound pairs of close-mid and close vowels. Extract of Chapter M3.5, Table 1 (see Series 1, 4 and 7 in this table, including the respective details of vowel recognition). First pair = sounds of /e/ and /i/ produced by a woman. Second pair = sounds of /ø/ and /y/ produced by a man. Third pair = sounds of /o/ and /u/ produced by a woman. Note that, in the figures, intended f_o levels according to the musical C-major scale are given.
[C-03-05-F01] ↗

Figure 2: Formant pattern and spectral shape ambiguity in natural vocalisations: Similar estimated *F*-patterns and spectral envelopes for sound triplets of open or open-mid, close-mid and close vowels. Extract of Chapter M3.5, Table 2 (see Series 2, 4 and 9 in this table, including the respective details of vowel recognition). First triplet = sounds of /ɛ/, /e/ and /y/ produced by women. Second triplet = sounds of /a/, /o/ and /u/ produced by men (note that the sounds of /a/ and /o/ were produced by the same speaker). Third triplet = sounds of /ɔ/, /o/ and /u/ produced by women.
[C-03-05-F02] ↗

Figure 1. Formant pattern and spectral shape ambiguity in natural vocalisations: Similar estimated F-patterns and spectral envelopes for sound pairs of close-mid and close vowels. [C-03-05-F01]



1–1 [e] 175-V-low 1052-A-w [e]
R141713 F(i):322-2499-3158

1–2 [i] 330-V-med 1052-A-w [i]
R140734 F(i):323-2358-3130

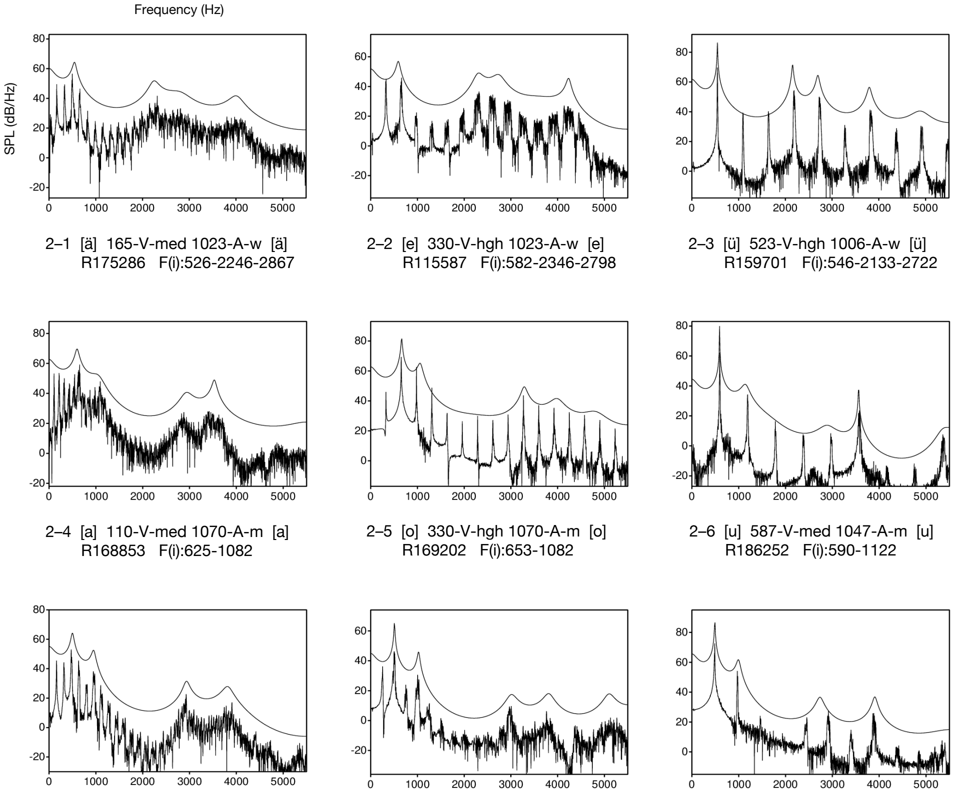
1–3 [ö] 165-V-hgh 1002-A-m [ö]
R133311 F(i):335-1696-2289

1–4 [ü] 330-V-med 1002-A-m [ü]
R103860 F(i):323-1640-2342

1–5 [o] 220-V-med 1005-A-w [o]
R181101 F(i):410-780

1–6 [u] 392-V-med 1005-A-w [u]
R181070 F(i):393-809

Figure 2. Formant pattern and spectral shape ambiguity in natural vocalisations: Similar estimated F-patterns and spectral envelopes for sound triplets of open or open-mid, close-mid and close vowels. [C-03-05-F02]



2–1 [ä] 165-V-med 1023-A-w [ä]
R175286 F(i):526-2246-2867

2–2 [e] 330-V-hgh 1023-A-w [e]
R115587 F(i):582-2346-2798

2–3 [ü] 523-V-hgh 1006-A-w [ü]
R159701 F(i):546-2133-2722

2–4 [a] 110-V-med 1070-A-m [a]
R168853 F(i):625-1082

2–5 [o] 330-V-hgh 1070-A-m [o]
R169202 F(i):653-1082

2–6 [u] 587-V-med 1047-A-m [u]
R186252 F(i):590-1122

2–7 [o1] 165-V-med 1048-A-w [o1]
R129590 F(i):497-981

2–8 [o] 247-V-hgh 1053-A-w [o]
R148746 F(i):496-1009

2–9 [u] 494-V-hgh 1004-A-w [u]
R157784 F(i):490-988

3.6 Conclusion

In the Preliminaries (pp. 187–216), based on previous studies and an older sample of vowel sounds, we have already documented numerous comparisons of natural sounds of two or three different vowels produced at different f_o levels with similar patterns of relative spectral energy maxima and/or similar estimated *F*-patterns within their supposed vowel-specific frequency range. However, in this third main chapter, we have addressed the formant pattern and spectral shape ambiguity phenomenon in a broader perspective: Firstly, by means of an investigation into vowel synthesis and resynthesis experiments related to statistical *F*-patterns, spectral envelopes of single natural vowel sounds, and model *F*-patterns including resonance patterns of half-open tubes commonly attributed to the average vocal tract lengths of men, women and children; secondly, against the background of the (re-) synthesis experiments and on the basis of the newly compiled Zurich Corpus, by means of a renewed documentation of natural sounds of different vowels produced at different f_o levels with similar patterns of relative spectral energy maxima, similar estimated *F*-patterns and similar estimated spectral envelopes within their supposed vowel-specific frequency range, these similarities being crosschecked in vowel resynthesis.

The results of all previous and present experiments lead to the conclusion that no matter how single spectral peak patterns or single *F*-patterns or single spectral envelopes are estimated, they are, as such, not related to single vowel qualities *in general.* In most cases, they represent sounds of different vowels. At the same time, the experiments and their results also lead to the conclusion that the ambiguity depends, above all, on the levels and ranges of f_o variation as well as on the vowel qualities in question: While ambiguous *F*-patterns and spectral shapes were hardly found for f_o variation below c. 200 Hz, they were widespread for higher frequency ranges, and while these spectral characteristics were rarely found as ambiguous for sounds of /a/, they were ambiguous in a quasi-systematic way for sounds of close-mid and close vowels.

The experiments investigating *F*-patterns and spectral envelopes in this third chapter did not address and discuss spectral changes due to variations in phonation, vocal effort, and additional production modes (above all, variation in speaking and singing styles, including register changes). This limitation of the investigation does not relativise the general conclusion that the formant pattern and the spectral shape are ambiguous acoustic representations of vowel quality, as the results of model vowel synthesis and of resynthesis have demonstrated. Nevertheless, attention should be given to the fact that the evidence was provided for specific sets of sounds and related production parameters only and that these settings have to be accounted for when further investigating the ambiguity phenomenon. Also, concerning the significance of the results, the limitations of the method of investigation have to be considered. Above all, for the sounds produced with the Klatt synthesiser and, although to a lesser degree, with the spectral envelope synthesiser, the sound quality was limited in general and some artefacts occurred, sometimes making listening to the sounds unpleasant to the ear. These artefacts, although not under further investigation here, also have to be taken into account with regard to an interpretation of results, and their occurrence points again to the need for improved resynthesis and synthesis tools.

However, to repeat, evidence is given anew for the main conclusion that formant patterns and spectral shapes *per se* are ambiguous spectral representations of recognised vowel quality. In our understanding, this ambiguity phenomenon is at the core of any reflection on the acoustics of the vowel and of a future acoustic theory.

But how is it that spectral peaks, estimated *F*-patterns and estimated spectral envelopes prove to be ambiguous acoustic representations of vowel quality? And how is it that this ambiguity proves to be non-uniform among levels and ranges of f_o variation and vowel qualities? From a purely observational perspective, the ambiguity had to be expected because recognisable vowel sounds can be produced on f_o levels in a frequency range that encompasses the range of statistical F_1 of sounds of almost all vowels. In contrast, when looked at from a purely productional perspective based on a source–filter model in which only the filter configuration is understood as vowel-related, the ambiguity is difficult to understand: Why should f_o variation as a characteristic of the source affect the recognised vowel quality if the resonances of the vocal tract are kept unchanged and if the resonances are mirrored in the resulting spectral peak patterns? To the best of our knowledge, there is no indication given in the literature that could serve as an explanation of the ambiguity phenomenon within the prevailing source–filter model of speech production. Note in this context that although source–filter interactions are discussed in the literature, these interactions do not concern the ambiguity shown here. Further, the supposed effect of vocal tract size in terms of different *F*-patterns for sounds of a given vowel produced by speakers different in age and gender also does not concern the phenomenon in question since the ambiguity of *F*-patterns and spectral shapes was demonstrated for synthesised and resynthesised vowel sounds as well as for natural vowel sounds produced by single speakers or speakers of a given age- and gender-related speaker group. Finally, the supposed effect of “spectral undersampling” of resonance or filter curves for sounds produced at middle or higher f_o levels does not serve as an explanation for the ambiguity phenomenon either since the ambiguity was confirmed in model synthesis experiments in which the harmonic frequencies of the sounds matched the filter frequencies of sound production.

Below, the question of why spectral peaks, estimated *F*-patterns and estimated spectral envelopes prove to be ambiguous acoustic representations of vowel quality is addressed in two steps: Firstly, an excursus presents a reflection on the difference between f_o and pitch – f_o as an aspect of sound production or of an acoustic measure of the radiated sound and its periodicity, and pitch as an aspect of sound perception and sound quality recognition. By doing so, the role of perception will come into focus. Secondly, in Chapter 6 and based on the reflections of the excursus, experiments and their results are presented that lead to the thesis of a vowel–pitch relation: We will argue that it is not f_o but pitch (or a comparable perceptual referencing to a sound pattern repetition over time) that the vowel spectrum relates to and that the observed relation of spectral peaks, estimated *F*-patterns and spectral shapes to f_o is but an indication of this relation. This argument will offer an answer to the above question. However, preceding the excursus and the sixth chapter, the matter of *F*-pattern and spectral shape differences for sounds of a vowel commonly attributed to differences in either the age and gender of the speakers, or phonation type or vocal effort, is discussed in the fourth and fifth chapters to provide an extended phenomenological basis for the direct exposition of the vowel–pitch relation thesis.

4 Vowel Spectrum and Age and Gender of the Speakers

4.1 Similar Lower Spectral Peak Frequencies and Estimated Formant Frequencies for Vowel Sounds Produced by Children, Women and Men at a Similar f_o

Vowel quality-related *F*-patterns, as given in formant statistics for voiced vowel sounds, differ according to the age and gender of the speakers: Generally, statistical average *F*-patterns were reported as highest for children, intermediate for women and lowest for men. These differences are commonly understood to be primarily due to the different average vocal tract sizes of these age- and gender-related groups.

However, to the best of our knowledge, in almost all statistical investigations of average *F*-patterns, the sounds were produced at f_o levels comparable to relaxed speech, these levels representing only the lower part of the vocal range of the speakers in general and of recognisable vowel sounds in particular (see Chapter 2). Therefore, existing formant statistics do not provide empirical evidence of whether or not supposed age- and gender-related spectral differences remain if f_o is substantially varied. In this context, two types of comparisons are of specific interest: Firstly, vowel sounds produced by speakers different in age and gender at similar f_o levels; secondly, vowel sounds produced by adults at f_o levels that are higher than the levels of those produced by children, and vowel sounds produced by men at f_o levels that are higher than the levels of those produced by women. The first aspect is addressed in this chapter; the second will be addressed in the following chapter in terms of adult–children sound comparisons.

In earlier studies, on the bases of either investigating statistical average *F*-patterns of speakers different in age and gender or directly comparing single sounds and their spectra of single speakers different in age and gender, with sounds produced by the speakers at different and similar f_o levels, we have in many cases observed a decrease or even a disappearance of expected speaker-group differences in spectral peaks, estimated formant frequencies and the entire spectral envelope < 1.5–2 kHz, that is, a decrease or even a disappearance of the spectral differences commonly related to the F_1–F_2 of sounds of back vowels and /a/ and to the F_1 of sounds of front vowels. Therefore, we have concluded that the variation of f_o has a much stronger effect on the lower vowel-related formants < 1.5–2 kHz than the assumed average vocal tract size of the speakers does.

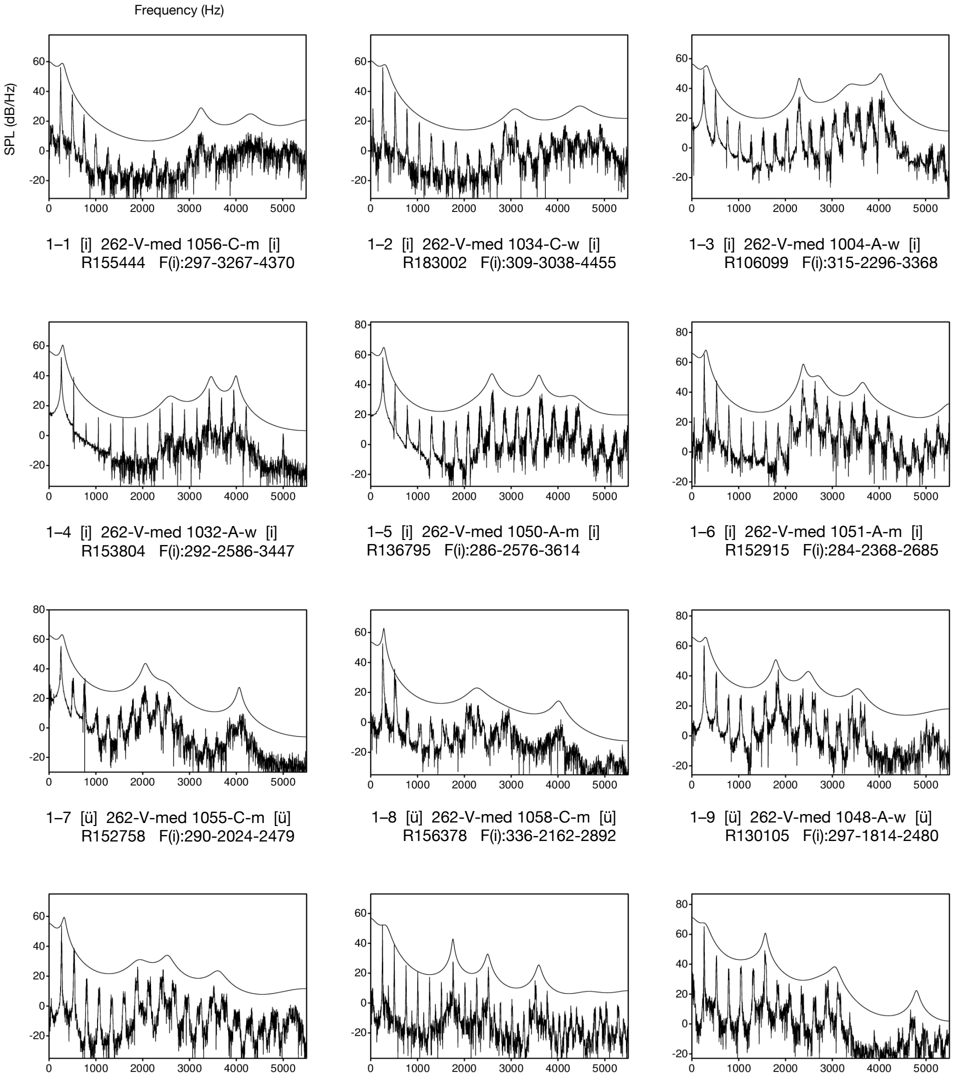
In the Preliminaries, we have already given exemplary documentation of corresponding sound comparisons based on recordings of earlier studies. However, as explained in the Introduction and Chapter 1.1, these sounds were recorded under varying conditions and with varying sound qualities, and the rights for online playback could not be obtained retrospectively from all speakers. Therefore, the documentation was renewed and extended based on the Zurich Corpus and integrated here in this treatise: Firstly, for each of the eight long Standard German vowels relating to the vocalises presented and discussed in Chapter 2.1, one sound of each of the three speakers produced at f_o within a range of 220–330 Hz was selected, the three selected sounds manifesting similar spectral peak frequencies (indicated by similar prominent harmonics) and similar estimated formant frequencies < 1.5 kHz, that is, similar F_1 for sounds of all vowels, and also similar F_2 for sounds of back vowels and /a/. Secondly, on the basis of other sounds of the Zurich Corpus, for each of the eight long Standard German vowels and the corresponding sounds produced with a medium vocal effort at an intended f_o of 262 Hz, two sounds produced by children, two by women and two by men were selected that manifested similar spectral peak frequencies and similar estimated formant frequencies < 1.5 kHz and that were unambiguously recognised in the standard listening test conducted when creating the corpus, matching vowel intention.

As a result, two samples of 24 and 48 sounds were created and are documented in this treatise. Both compilations demonstrate comparisons of vowel sounds produced by speakers different in age or gender at similar f_o levels, not manifesting general age- or gender-related differences in estimated spectral peaks and formant frequencies < 1.5 kHz: For almost all sounds of all speakers, estimated F_1 differences were either < 70 Hz or related to higher values for adults than children, and estimated F_2 were comparable for children, women and men. Indeed, for this lower frequency range < 1.5 kHz, the entire harmonic configuration was comparable for the sounds produced by all speakers. Figure 1 illustrates this finding for the sounds of the second compilation: For each of the eight vowels investigated, the spectra of six sounds produced by two children, two women and two men are compared with each other directly. No vowel-related spectral differences < 1.5 kHz are indicated in these comparisons. (Concerning the first compilation, please refer to Chapter M4.1.)

For references, extended background information, details of experimental design, method and results, an extended discussion and documentation of results (tables including sound links), see the Materials, Chapter M4.1.

Figure 1. Comparison of sounds and sound spectra of the long Standard German vowels produced by children, women and men at a similar f_o level: Occurrence of similar lower spectral peak frequencies and estimated formant frequencies in contrast to commonly expected age- and gender-related *F*-pattern differences. Extract of Chapter M4.1, Table 2 (see this table for the estimate *F*-patterns). For each vowel, six sounds produced by two children, two women and two men with a medium vocal effort at an intended f_o of 262 Hz (indicated in the figure) are shown, not manifesting supposed general age- and gender-related spectral differences < 1.5 kHz.
[C-04-01-F01] ↗

Figure 1. Comparison of sounds and sound spectra of the long Standard German vowels produced by children, women and men at a similar fo level: Occurrence of similar lower spectral peak frequencies and estimated formant frequencies in contrast to commonly expected age- and gender-related F-pattern differences. [C-04-01-F01]



1–1 [i] 262-V-med 1056-C-m [i]
R155444 F(i):297-3267-4370

1–2 [i] 262-V-med 1034-C-w [i]
R183002 F(i):309-3038-4455

1–3 [i] 262-V-med 1004-A-w [i]
R106099 F(i):315-2296-3368

1–4 [i] 262-V-med 1032-A-w [i]
R153804 F(i):292-2586-3447

1–5 [i] 262-V-med 1050-A-m [i]
R136795 F(i):286-2576-3614

1–6 [i] 262-V-med 1051-A-m [i]
R152915 F(i):284-2368-2685

1–7 [ü] 262-V-med 1055-C-m [ü]
R152758 F(i):290-2024-2479

1–8 [ü] 262-V-med 1058-C-m [ü]
R156378 F(i):336-2162-2892

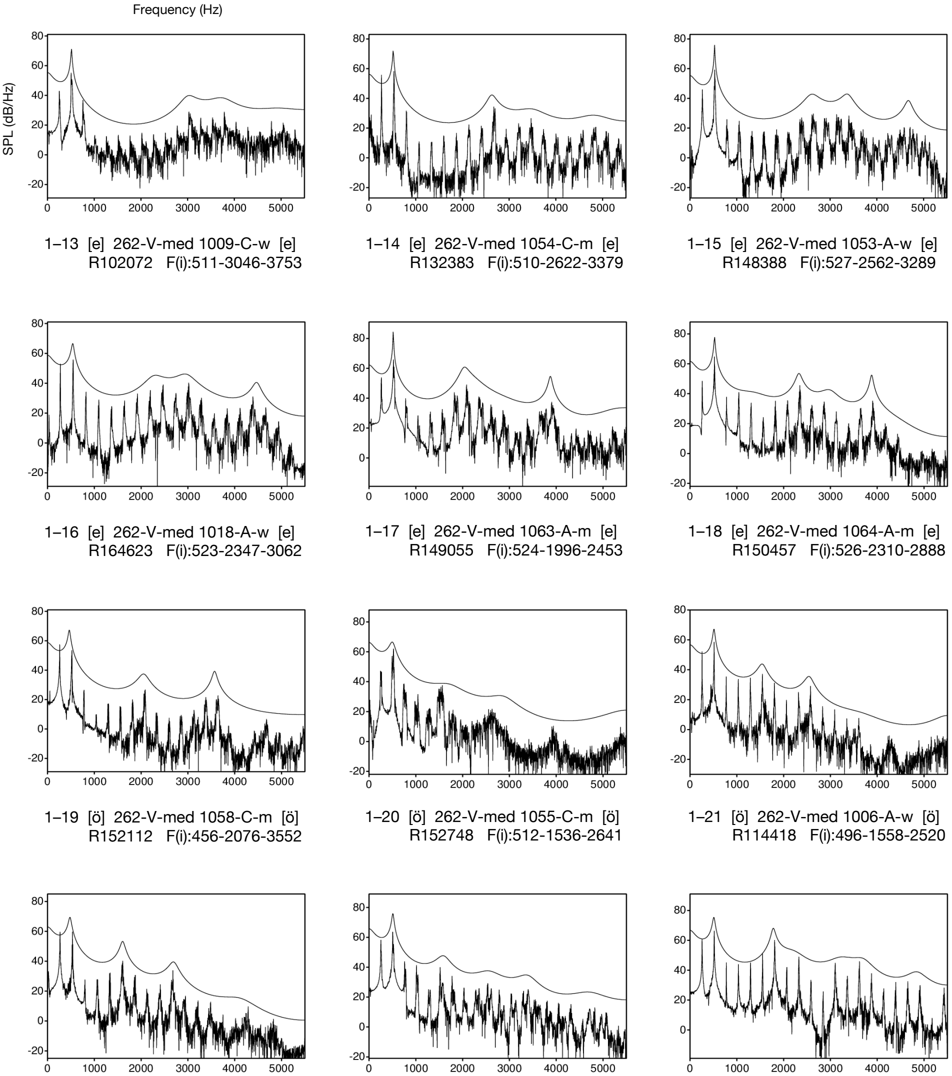
1–9 [ü] 262-V-med 1048-A-w [ü]
R130105 F(i):297-1814-2480

1–10 [ü] 262-V-med 1053-A-w [ü]
R148444 F(i):340-1819-2480

1–11 [ü] 262-V-med 1030-A-m [ü]
R119150 F(i):351-1738-2491

1–12 [ü] 262-V-med 1049-A-m [ü]
R135236 F(i):316-1536-1802

Figure 1 (continuation). [C-04-01-F01]



1–13 [e] 262-V-med 1009-C-w [e]
R102072 F(i):511-3046-3753

1–14 [e] 262-V-med 1054-C-m [e]
R132383 F(i):510-2622-3379

1–15 [e] 262-V-med 1053-A-w [e]
R148388 F(i):527-2562-3289

1–16 [e] 262-V-med 1018-A-w [e]
R164623 F(i):523-2347-3062

1–17 [e] 262-V-med 1063-A-m [e]
R149055 F(i):524-1996-2453

1–18 [e] 262-V-med 1064-A-m [e]
R150457 F(i):526-2310-2888

1–19 [ö] 262-V-med 1058-C-m [ö]
R152112 F(i):456-2076-3552

1–20 [ö] 262-V-med 1055-C-m [ö]
R152748 F(i):512-1536-2641

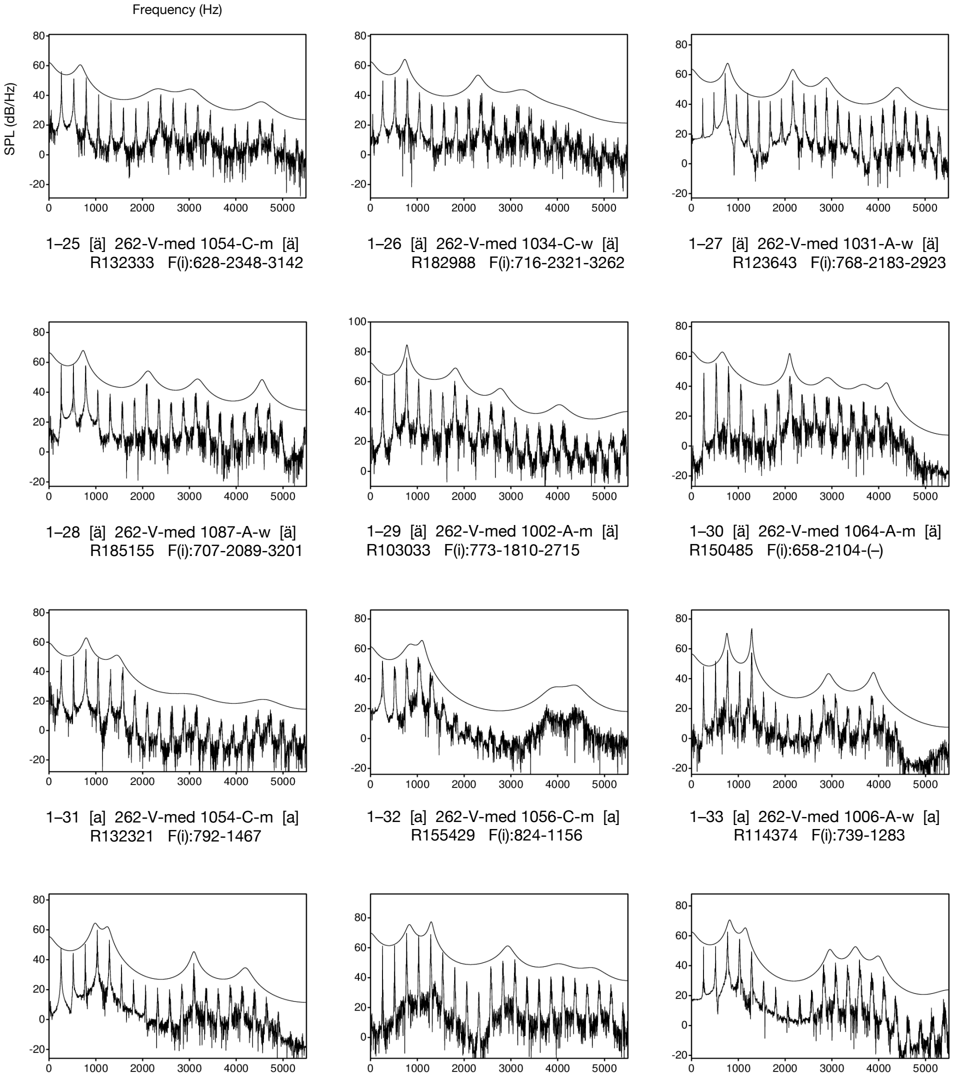
1–21 [ö] 262-V-med 1006-A-w [ö]
R114418 F(i):496-1558-2520

1–22 [ö] 262-V-med 1036-A-w [ö]
R138935 F(i):452-1638-2702

1–23 [ö] 262-V-med 1045-A-m [ö]
R124574 F(i):508-1552-2536

1–24 [ö] 262-V-med 1069-A-m [ö]
R165532 F(i):502-1775-2176

Figure 1 (continuation). [C-04-01-F01]



1–25 [ä] 262-V-med 1054-C-m [ä]
R132333 F(i):628-2348-3142

1–26 [ä] 262-V-med 1034-C-w [ä]
R182988 F(i):716-2321-3262

1–27 [ä] 262-V-med 1031-A-w [ä]
R123643 F(i):768-2183-2923

1–28 [ä] 262-V-med 1087-A-w [ä]
R185155 F(i):707-2089-3201

1–29 [ä] 262-V-med 1002-A-m [ä]
R103033 F(i):773-1810-2715

1–30 [ä] 262-V-med 1064-A-m [ä]
R150485 F(i):658-2104-(–)

1–31 [a] 262-V-med 1054-C-m [a]
R132321 F(i):792-1467

1–32 [a] 262-V-med 1056-C-m [a]
R155429 F(i):824-1156

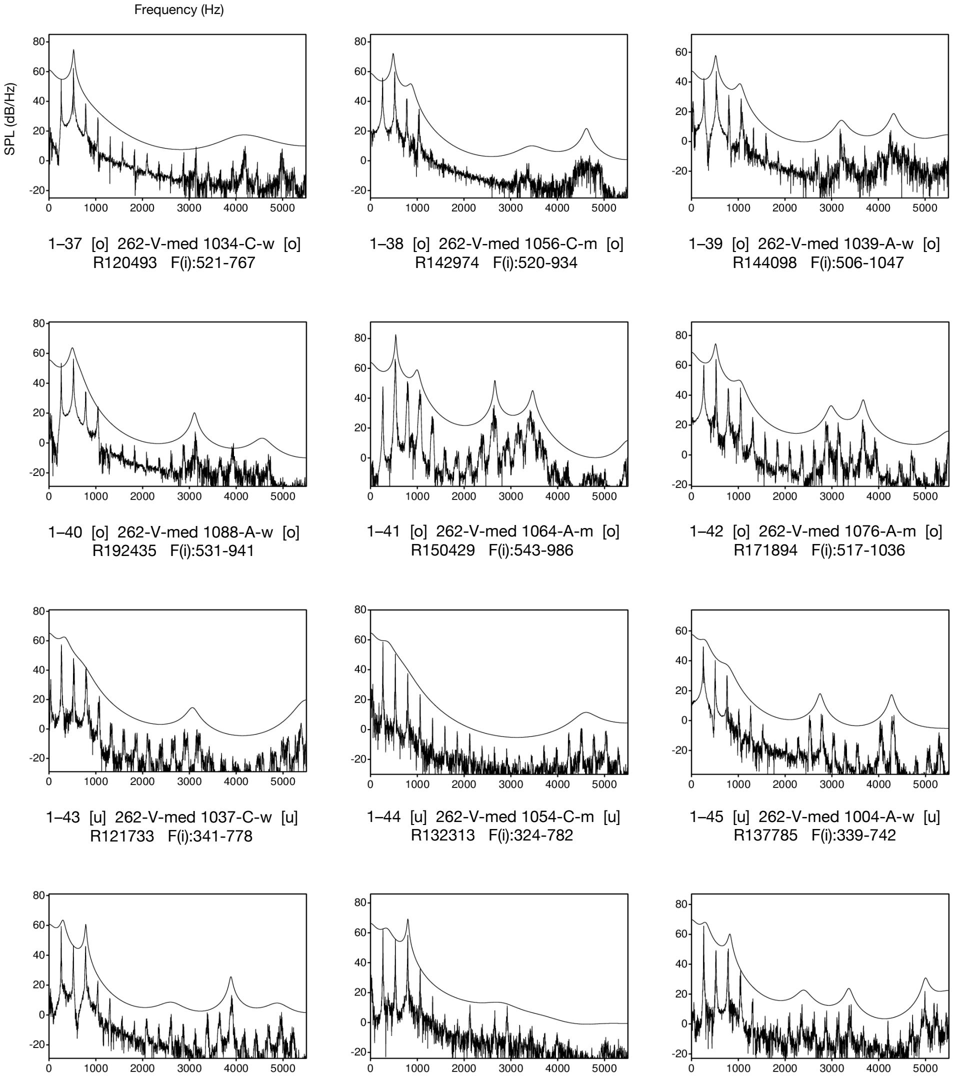
1–33 [a] 262-V-med 1006-A-w [a]
R114374 F(i):739-1283

1–34 [a] 262-V-med 1046-A-w [a]
R159927 F(i):974-1209

1–35 [a] 262-V-med 1069-A-m [a]
R165480 F(i):831-1340

1–36 [a] 262-V-med 1047-A-m [a]
R183724 F(i):815-1143

Figure 1 (continuation). [C-04-01-F01]



1–37 [o] 262-V-med 1034-C-w [o]
R120493 F(i):521-767

1–38 [o] 262-V-med 1056-C-m [o]
R142974 F(i):520-934

1–39 [o] 262-V-med 1039-A-w [o]
R144098 F(i):506-1047

1–40 [o] 262-V-med 1088-A-w [o]
R192435 F(i):531-941

1–41 [o] 262-V-med 1064-A-m [o]
R150429 F(i):543-986

1–42 [o] 262-V-med 1076-A-m [o]
R171894 F(i):517-1036

1–43 [u] 262-V-med 1037-C-w [u]
R121733 F(i):341-778

1–44 [u] 262-V-med 1054-C-m [u]
R132313 F(i):324-782

1–45 [u] 262-V-med 1004-A-w [u]
R137785 F(i):339-742

1–46 [u] 262-V-med 1086-A-w [u]
R185614 F(i):303-785

1–47 [u] 262-V-med 1049-A-m [u]
R135152 F(i):380-800

1–48 [u] 262-V-med 1045-A-m [u]
R169617 F(i):302-836

4.2 Cases of Inverted Vowel-Related Spectral Differences for Sounds Produced by Children and Adults

In the previous chapter, cases of vowel sounds are shown that do not manifest age- and gender-related differences in estimated spectral peaks and formant frequencies < 1.5 kHz, the compared sounds produced by all speakers at similar f_o levels. In pursuit of the question of whether vocal tract size stands in an imperative, direct relation to the entire vowel spectrum, a further experiment was devised to investigate the occurrence of inversions of commonly expected age-related F-pattern differences, that is higher vowel-related formant frequencies for sounds of adults than of children.

In an earlier study, in the documentation of the Preliminaries and the vocalises shown in Chapter 2.1, such inversions within a spectral range < 1.5 kHz were indeed indicated. Above all, they concerned sounds of close and close-mid vowels produced by children at lower f_o levels of their vocal range and by adults at middle or higher f_o levels. Further, considering the variation extent of the higher formants reported for the sounds of front vowels (for reference, see Chapter M4.2), the question of the occurrence of inverted spectral age and gender differences can be extended to the comparison of F_1–F_2 patterns for sounds of all vowels, independent of the frequency range of F_2, and it can even be extended to the comparison of F_1–F_2–F_3 for sounds of front vowels. In this context, sound comparisons between children and adults are of particular interest because of the very pronounced differences in their vocal tract sizes.

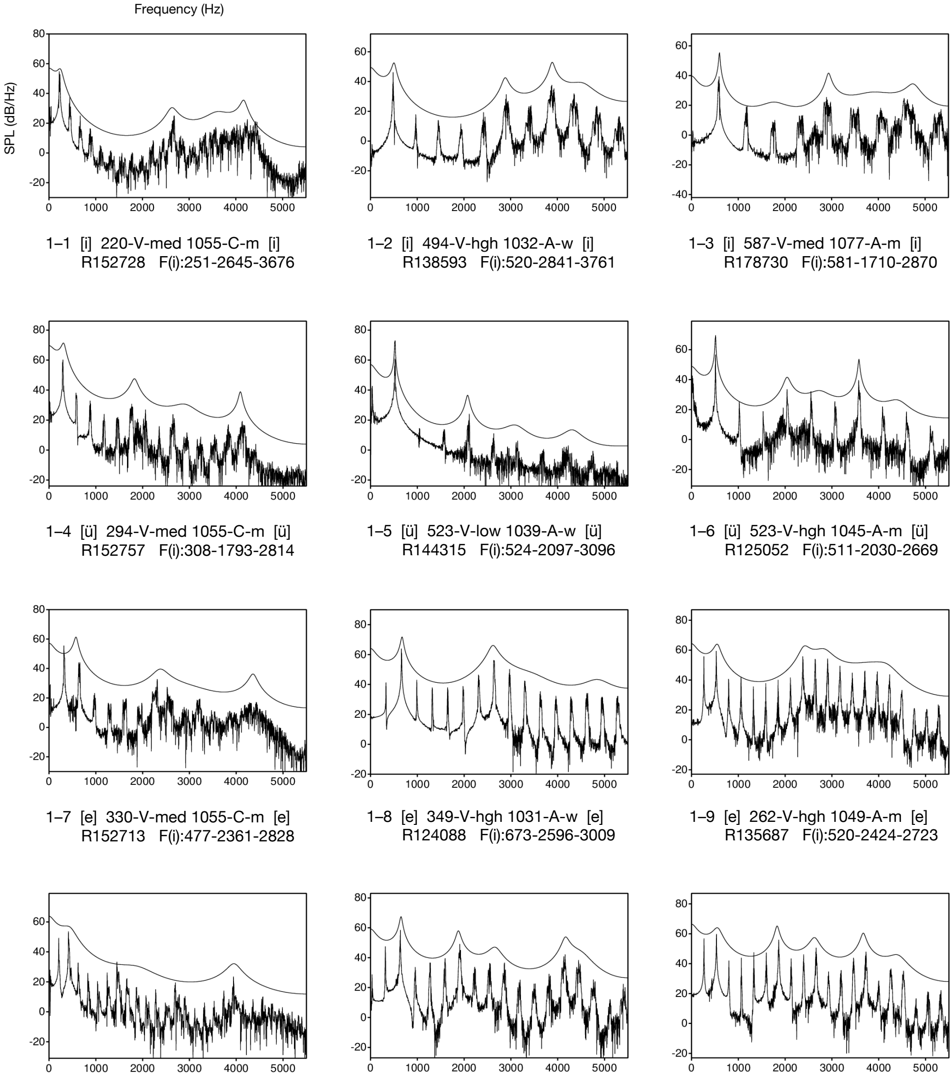
This being the case, to highlight the inversion phenomenon and to document and embed it in the context of our line of argument, an attempt was made to compile sound series manifesting inverted F_1 or F_1–F_2 or even F_1–F_2–F_3 standing in contrast to vocal tract size-related assumptions. The attempt was based on the above indications and on the assumption that the effect of f_o variation on the lower part of the vowel spectrum is more important than the vocal tract size of the speaker in question. Vocal effort and production style were disregarded for the sound selection. According to this experimental approach, for each of the eight long Standard German vowels and each of the three speaker groups of children, women and men, two sounds produced in V context were selected from the Zurich Corpus, the sounds compared manifesting children–adult inversions with regard to vocal tract size-related expectations for one or more estimated spectral peaks and estimated formant frequencies. All selected sounds were fully recognised in the standard listening test conducted when creating the corpus, matching vowel intention. As a result, a compilation of a total of 48 sounds for the eight vowels investigated was created.

According to the results of the acoustic analysis, for all sounds of all vowels, pronounced children–adult inversions of F_1 were found, that is, lower F_1 for the sounds of children than of the adults. Except for two sounds produced by men, the same held true for pronounced children–adult inversions of F_1–F_2. Besides, cases of children–women inversions of F_1–F_2–F_3 were also found for sounds of /i/, /ø/ and /ɛ/. Thus, concerning commonly assumed age- and gender-related spectral differences, the compilation documents anew that these differences may not only decrease or disappear if f_o variation of the sounds is included in an investigation, but that the differences may also be inverted. Figure 1 illustrates this finding: For each of the eight vowels investigated, as a sample of the entire compilation made, triplets of sounds produced by a child at lower f_o and by a woman and a man at middle or higher f_o are shown. For sound comparisons, F_1–F_2 for the child occurred on lower frequency levels than F_1–F_2 for the adults.

For references, extended background information, details of experimental design, method and results, an extended discussion and documentation of results (tables including sound links), see the Materials, Chapter M4.2.

Figure 1: Comparison of sounds and sound spectra of the long Standard German vowels produced by children and adults at different f_o levels and/or with different vocal efforts: Occurrence of inversions of commonly expected age-related *F*-pattern differences. Extract of Chapter M4.2, Table 1 (in this table, see Series 1, sounds 1, 3 and 5; Series 2, sounds 1, 4 and 6; Series 3, sounds 2, 4 and 5; Series 4, sounds 2, 4 and 6; Series 5, sounds 2, 3 and 6; Series 6, sounds 1, 3 and 5; Series 7, sounds 1, 4 and 5; Series 8, sounds 2, 3 and 5). Triplets of sounds of the vowels /i, y, e, ø, ɛ, a, o, u/ and related spectra, intended f_o levels and estimated *F*-patterns are shown, the sounds of a triplet produced by a child, a woman and a man (in this order). The documented sounds of a vowel manifest lower F_1–F_2 for the child compared to the adults. (For the estimated F_1–F_2 values referred to here, see Chapter M4.2; note also that, in the figure, intended levels of f_o are indicated, in contrast to calculated f_o levels for the sounds presented online.)
[C-04-02-F01] ↗

Figure 1. Comparison of sounds and sound spectra of the long Standard German vowels produced by children and adults at different fo levels and/or with different vocal efforts: Occurrence of inversions of commonly expected age-related F-pattern differences. [C-04-02-F01]



1–1 [i] 220-V-med 1055-C-m [i]
R152728 F(i):251-2645-3676

1–2 [i] 494-V-hgh 1032-A-w [i]
R138593 F(i):520-2841-3761

1–3 [i] 587-V-med 1077-A-m [i]
R178730 F(i):581-1710-2870

1–4 [ü] 294-V-med 1055-C-m [ü]
R152757 F(i):308-1793-2814

1–5 [ü] 523-V-low 1039-A-w [ü]
R144315 F(i):524-2097-3096

1–6 [ü] 523-V-hgh 1045-A-m [ü]
R125052 F(i):511-2030-2669

1–7 [e] 330-V-med 1055-C-m [e]
R152713 F(i):477-2361-2828

1–8 [e] 349-V-hgh 1031-A-w [e]
R124088 F(i):673-2596-3009

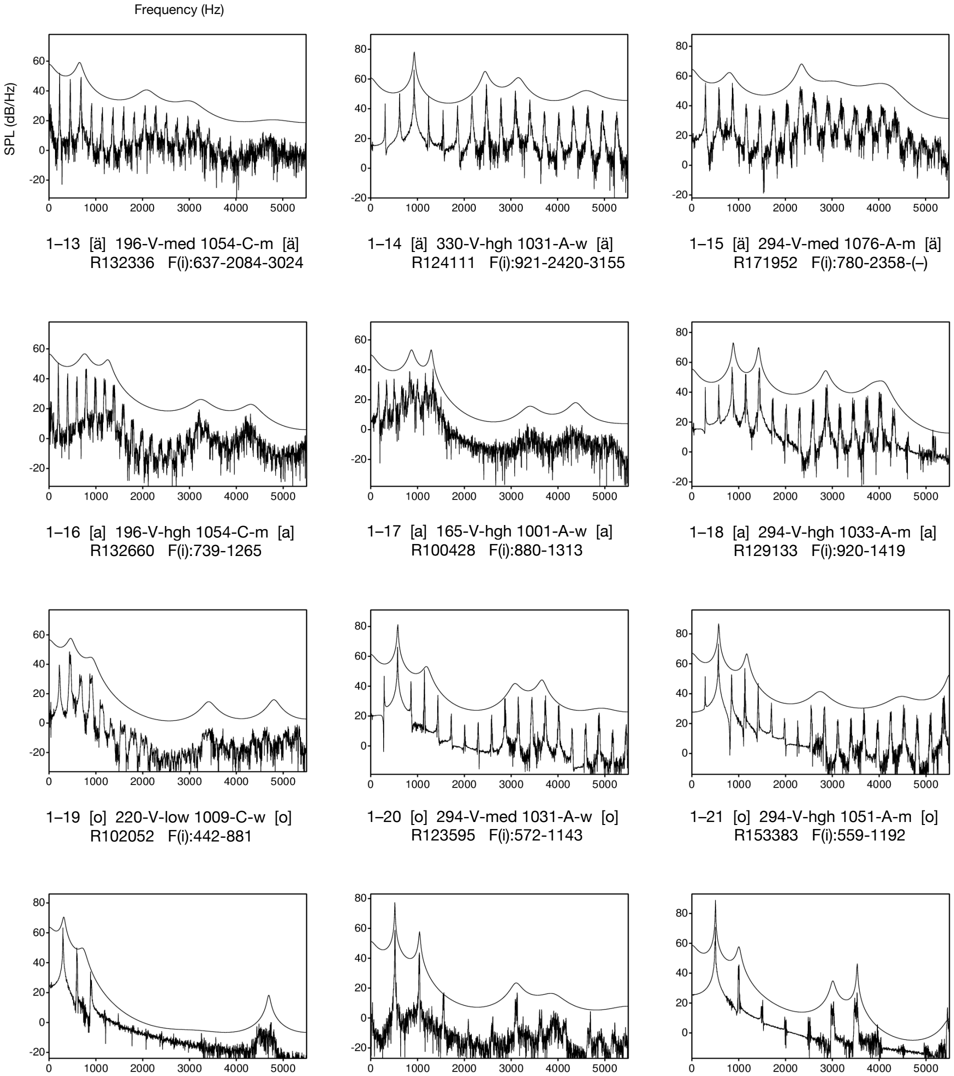
1–9 [e] 262-V-hgh 1049-A-m [e]
R135687 F(i):520-2424-2723

1–10 [ö] 220-V-med 1055-C-m [ö]
R152750 F(i):417-1702-3687

1–11 [ö] 330-V-med 1031-A-w [ö]
R123651 F(i):612-1893-2701

1–12 [ö] 262-V-hgh 1049-A-m [ö]
R135721 F(i):516-1815-2631

Figure 1 (continuation). [C-04-02-F01]



1–13 [ä] 196-V-med 1054-C-m [ä]
R132336 F(i):637-2084-3024

1–14 [ä] 330-V-hgh 1031-A-w [ä]
R124111 F(i):921-2420-3155

1–15 [ä] 294-V-med 1076-A-m [ä]
R171952 F(i):780-2358-(–)

1–16 [a] 196-V-hgh 1054-C-m [a]
R132660 F(i):739-1265

1–17 [a] 165-V-hgh 1001-A-w [a]
R100428 F(i):880-1313

1–18 [a] 294-V-hgh 1033-A-m [a]
R129133 F(i):920-1419

1–19 [o] 220-V-low 1009-C-w [o]
R102052 F(i):442-881

1–20 [o] 294-V-med 1031-A-w [o]
R123595 F(i):572-1143

1–21 [o] 294-V-hgh 1051-A-m [o]
R153383 F(i):559-1192

1–22 [u] 294-V-med 1056-C-m [u]
R142961 F(i):312-711

1–23 [u] 523-V-med 1006-A-w [u]
R159433 F(i):519-1052

1–24 [u] 494-V-med 1047-A-m [u]
R186206 F(i):503-1002

4.3 Conclusion

The argument and documentation given in this main chapter on the vowel spectrum and its supposed relation to the age and gender of the speakers – and, with this, the supposed relation between the vowel spectrum and age- and gender-related average vocal tract sizes – make evident that, in fact, there is no such *general* relation for spectral characteristics < c. 1.5–2 kHz for sounds of back vowels and /a/ (frequency range of statistical F_1–F_2 of these vowels), as holds true for spectral characteristics < c. 1 kHz for sounds of front vowels (frequency range of statistical F_1 of all vowels): According to the results of the experiments conducted, for sounds of back vowels and /a/, age- and gender-related spectral peak and formant frequency differences ≤ c. 1.5–2 kHz, as generally given in formant statistics, were indicated to decrease or disappear if the speakers produced vowel sounds at a similar level of f_o. (Here, with regard to a general appraisal, a frequency range of < 1.5–2 kHz is discussed to include sounds of /a, o, u/ produced at high f_o levels above 500 Hz.) The same held true for sounds of front vowels and a frequency range of 0–1 kHz. Furthermore, and most importantly, commonly assumed age- and gender-related differences in F_1–F_2 for sounds of all vowels and for a frequency range exceeding 2 kHz were sometimes also inverted, with the sounds of the adults produced at an f_o level and/or a level of vocal effort that was markedly above the level(s) of the sounds produced by the children. (The same has to be expected for sounds of men produced at an f_o level markedly above the level of the sounds produced by women, an aspect that was not investigated here.) This finding was predictable because of the large range of f_o for recognisable vowel sounds, the relation of the lower spectrum to the f_o level of a sound, the variability of higher formants for sounds of front vowels reported in the literature, and the resulting formant pattern and spectral shape ambiguity, as these general aspects of the vowel sound and its spectrum were discussed in the previous main chapters.

Vocal effort, production style and register changes were disregarded in the experiment. However, no systematic speaker group-related vocal effort was manifest in the sound compilation investigated, and only 7 of the 32 selected sounds of the adults were produced in a particular style.

Concerning register changes and related changes in articulation, it may be tempting to argue that the sounds produced by men were subject to a change from the modal to the mixed or falsetto voice and that, therefore, the vocal tract was shortened and the articulation adjusted. It is then tempting to generalise that sounds produced by women may also be affected by register changes in a comparable way. However, such an interpretation would not account for the relation of the lower spectrum to the f_o of a sound and the resulting formant pattern and spectral shape ambiguity. An alternative explanation for the fact that commonly assumed age- and gender-related spectral differences can decrease, disappear or even be inverted (depending on f_o and additional characteristics of sound production) is to pose the question of a possible non-direct relation of the vocal tract and the acoustic characteristics of the produced vowel sound. We will return to this question in more detail below. In conclusion, aspects of vocal effort variation, production styles and register changes and related changes in articulation have to be considered when interpreting the above inversion phenomenon, but they can barely explain the phenomenon as such.

5 Vowel Spectrum, Phonation Type and Vocal Effort

5.1 Vowel Spectrum and Phonation Type I – Comparing the Spectra of Natural Vowel Sounds Produced With Voiced, Whispered, Creaky and Breathy Phonation, Excluding f_0 Variation of the Voiced Sounds

In the literature, the spectrum of whispered vowel sounds is reported as manifesting higher estimated *F*-patterns (with most pronounced differences for F_1 and with widened formant bandwidths) and a different spectral energy distribution (lower acoustic power, relatively flat noise-like spectrum) when compared with the spectrum of voiced sounds. These phonation-related spectral differences are understood as being mainly a consequence of changes in the geometry of the vocal tract around the glottis and the difference in the source characteristics.

The periodicity of vowel sounds produced with creaky phonation is reported as irregular, but the sound spectrum is reported to show no or only marginal spectral peak differences when compared with voiced sounds.

The spectral characteristics of vowel sounds produced with breathy phonation are reported to manifest an increased amplitude of the first harmonic, a lessened sound periodicity and weak levels of the harmonics in the upper frequency part of the spectrum combined with increased aspiration noise when compared with voiced vowel sounds. These phonation-related acoustic differences are understood as a consequence of the tendency of the glottal source function to be sinusoidal-like combined with a high level of aspiration noise due to nonsimultaneous closure along the length of the vocal folds.

To re-examine the reported spectral variation of sounds of a vowel due to different phonation types, document a set of vowel sounds and their spectra comparable to sound samples investigated in the literature and provide direct and accessible sound comparisons and a starting point and literature-related reference for subsequent experimentation, a corresponding study was conducted. Sound samples of a man, a woman and a child were selected from the Zurich Corpus that included recognised sounds of all long Standard German vowels produced with voiced, whispered, creaky and breathy phonation (one sound per speaker, vowel and phonation type, all sounds produced in V context, voiced sounds produced with medium vocal effort in nonstyle mode, 100% recognition rate for all sounds in the standard listening test conducted when creating the corpus, matching vowel intention). To facilitate a comparison of the sounds of the present study with the sounds in the studies discussed in the literature, voiced sounds produced at intended f_o of 131 Hz (man), 220 Hz (woman) and 262 Hz (child) were selected. According to the standard procedure of the Zurich Corpus, the speakers produced breathy sounds at an f_o level of their choice, reflecting their respective vocal comfort zones and vocal ranges. As a result, eight vowel-related comparisons of four sounds were compiled for each speaker, and a total sample of 96 sounds was investigated.

The acoustic analysis was conducted according to the standard procedure of the Zurich Corpus. For each comparison of sounds of a single vowel produced by a single speaker (four sounds produced with voiced, whispered, creaky and breathy phonation), the main spectral differences or similarities reported in the literature were re-examined based on spectra, spectrograms and manifest spectral peak patterns P_1–P_2 (for details of the method and the reasons for estimating spectral peaks based on a visual inspection of the sound spectra, see Chapter M5.1).

In this re-examination, the general assumptions made in the literature that the different modes of voiced, whispered and creaky phonation result in the above spectral differences or similarities were confirmed for the majority of the examined sounds, above all concerning:

- Higher P_1 and often also higher P_2 for whispered than for voiced sounds
- Comparable P_1 and often also comparable P_2 for creaky and voiced sounds
- Increased amplitude of the first harmonic, steeper spectral slope < 1 kHz, weak levels of the harmonics in the upper frequency range of the spectrum and increased aspiration noise for breathy sounds when compared with voiced sounds

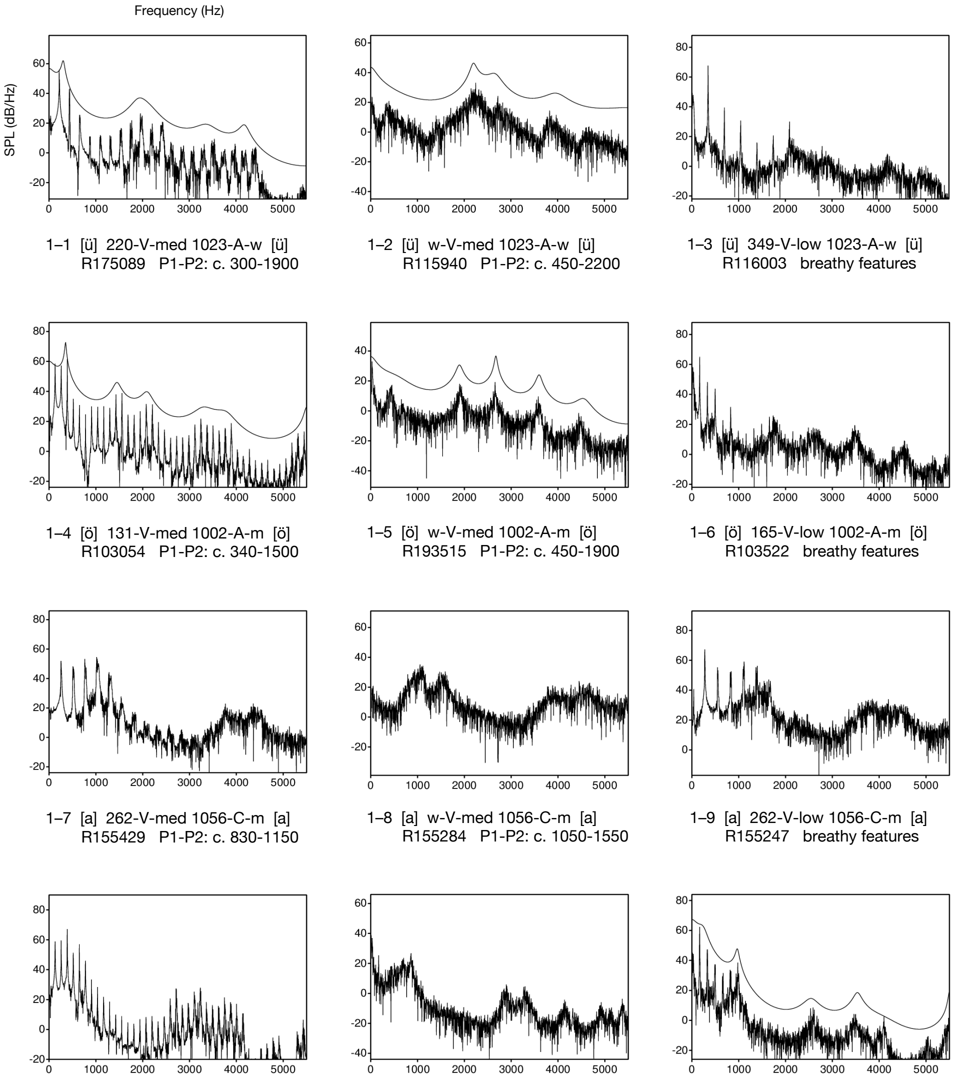
Figure 1 illustrates these spectral differences related to voiced, whispered and breathy phonation. However, exceptions to these tendencies of spectral differences also occurred.

As said, the aim of this experiment was limited to the documentation of a set of vowel sounds comparable to sets investigated in the literature, to provide direct and accessible sound comparisons and to create a starting point and literature-related reference for subsequent experiments. Therefore, no advanced discussion and relativisation of the results – relativisations that concern, above all, the small number of sounds and speakers, the lack of an inclusion of phonation subtypes in general and vocal effort variation for voiced and whispered sounds in particular, the lack of f_o variation for voiced and breathy sounds as well as the general methodological problem of spectral comparison – is made here. In these terms, the present study only provided illustrating examples based on an investigation comparable to most of the research published in the literature.

For references, extended background information, details of experimental design, method and results, an extended discussion and documentation of results (tables including sound links), see the Materials, Chapter M5.1.

Figure 1. Comparison of natural vowel sounds produced with voiced, whispered and breathy phonation, excluding f_o variation of the voiced sounds: Occurrence of phonation-related spectral differences. Extract of Chapter M5.1, Table 1 (see Series 10, 4, 22 and 7 in this table). For each sound triplet of a speaker presented, spectra of sounds produced with voiced, whispered and breathy phonation are shown. Sounds 1–3 = sounds of /y/ produced by the woman. Sounds 4–6 = sounds of /ø/ produced by the man. Sounds 7–9 = sounds of /a/ produced by the child. Sounds 10–12 = sounds of /o/ produced by the man. For details of acoustic analysis, see the sounds in the online corpus. [C-05-01-F01] ↗

Figure 1. Comparison of natural vowel sounds produced with voiced, whispered and breathy phonation, excluding fo variation of the voiced sounds: Occurrence of phonation-related spectral differences. [C-05-01-F01]



1–1 [ü] 220-V-med 1023-A-w [ü]
R175089 P1-P2: c. 300-1900

1–2 [ü] w-V-med 1023-A-w [ü]
R115940 P1-P2: c. 450-2200

1–3 [ü] 349-V-low 1023-A-w [ü]
R116003 breathy features

1–4 [ö] 131-V-med 1002-A-m [ö]
R103054 P1-P2: c. 340-1500

1–5 [ö] w-V-med 1002-A-m [ö]
R193515 P1-P2: c. 450-1900

1–6 [ö] 165-V-low 1002-A-m [ö]
R103522 breathy features

1–7 [a] 262-V-med 1056-C-m [a]
R155429 P1-P2: c. 830-1150

1–8 [a] w-V-med 1056-C-m [a]
R155284 P1-P2: c. 1050-1550

1–9 [a] 262-V-low 1056-C-m [a]
R155247 breathy features

1–10 [o] 131-V-med 1002-A-m [o]
R102984 P1-P2: c. 365-675

1–11 [o] w-V-med 1002-A-m [o]
R103490 P1-P2: >500 >700

1–12 [o] 165-V-low 1002-A-m [o]
R103526 breathy features

5.2 Vowel Spectrum and Phonation Type II – Comparing the First Lower Spectral Peak Frequency of Natural Vowel Sounds Produced With Voiced and Whispered Phonation, Including f_o Variation of the Voiced Sounds

As discussed in the previous chapter, according to the literature, the spectra of whispered vowel sounds are assumed to exhibit somewhat higher *F*-patterns in general and a pronounced increase for F_1 in particular when compared with voiced vowel sounds. However, this comparison is commonly discussed concerning voiced sounds produced at lower f_o levels of the vocal range of the speakers, corresponding to levels of citation-form words or relaxed speech. Thus, intra-speaker f_o variation for voiced sounds is not taken into account.

Hence, the following questions arise with regard to the relation between f_o, vowel-related spectral sound characteristics and phonation types: If natural voiced vowel sounds produced at different levels of f_o are compared with natural whispered vowel sounds (all produced by the same speaker), do assumed differences in vowel quality-related spectral peaks decrease or disappear? Or can they even be inverted, i.e., are there cases for which the vowel-related spectral peaks are higher for voiced sounds (produced at middle or high f_o levels) than for whispered sounds of comparison? (Note again that, when addressing such questions, the unsystematic character of spectral representation of vowel quality also has to be accounted for.) These questions were investigated in an extension of the previous experiment by including f_o variation for the voiced sounds. The investigation was limited to the examination of P_1 since, according to the literature, phonation-related differences between voiced and whispered vowel sounds were reported to be most pronounced for P_1/F_1. In addition, in this approach, the methodological problem of spectral peak and formant frequency estimation was limited to evaluating the lower frequency range < 1 kHz.

For each of the three speakers examined in the previous chapter and each of the eight long Standard German vowels, the voiced and whispered sounds were selected. Subsequently, for each sound comparison of a vowel and on the basis of other voiced sounds produced by the same speakers in V context with medium vocal effort (and in nonstyle mode for the two adult speakers) documented in the Zurich Corpus, two additional voiced sounds at medium and higher levels of f_o were selected for which their spectra showed either comparable or higher estimated P_1 than for the whispered sound. Finally, applying the same procedure, one sound per vowel and speaker produced in V context with a high vocal effort at a middle or higher f_o level was also added to each of the sound comparisons. All sounds selected and compared were fully recognised in the standard listening test conducted when creating the corpus, matching vowel intention. Sounds produced with high vocal effort were included to obtain a first indication of a possible additional effect of vocal effort variation for this type of sound comparison. As a result, a total sound sample of 120 sounds was created.

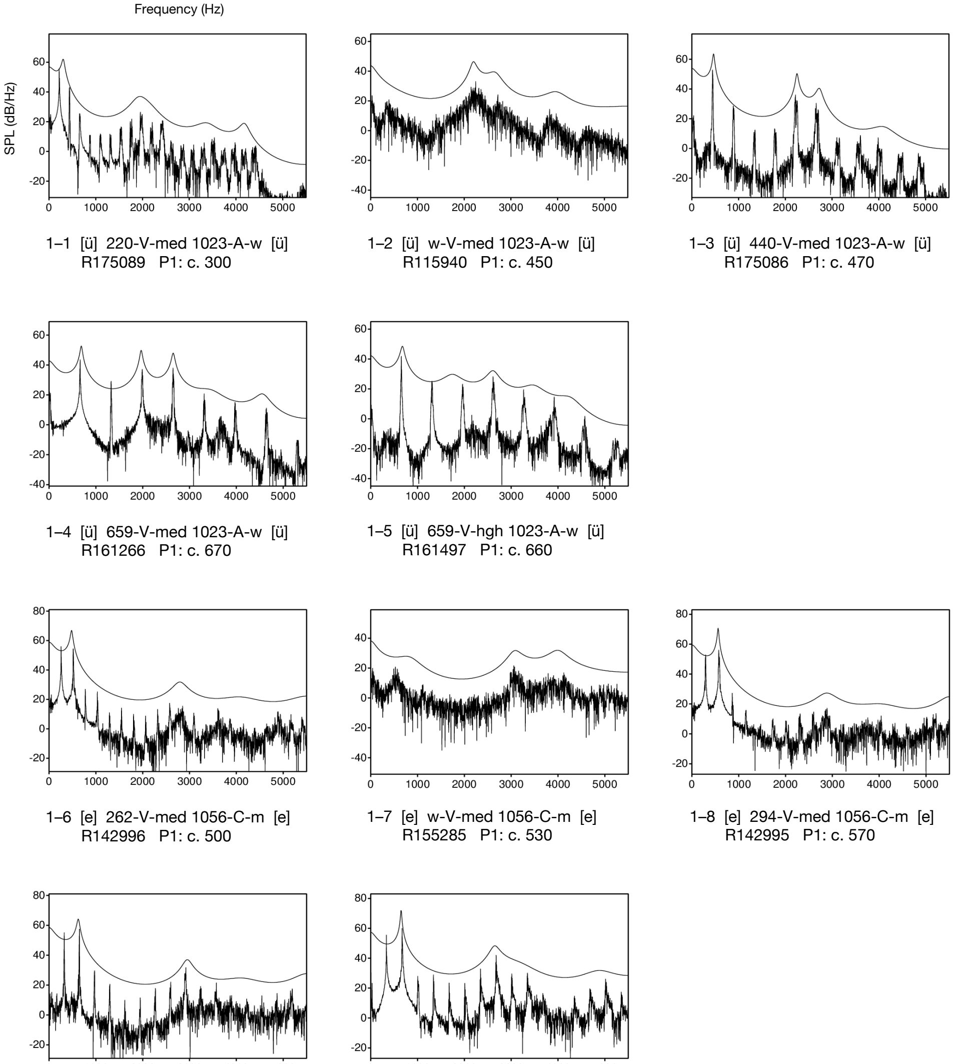
As shown in the previous experiment in Chapter 5.1, if the spectra of whispered and voiced vowel sounds produced by the three speakers were compared, and if the voiced sounds were produced at the lower f_o levels comparable to age- and gender-specific average levels as given in formant statistics, the P_1 of the whispered sounds were in most cases higher than the P_1 of the voiced sounds. This finding was in line with the general prediction given in the literature. However, if the voiced sounds were produced by speakers of all three speaker groups with a medium vocal effort at increased levels of f_o, for almost all comparisons investigated in this experiment, the voiced sounds manifested P_1 comparable to those of the whispered sounds. Moreover, if the f_o level of the voiced sounds was further increased and/or vocal effort was varied, the P_1 of these sounds surpassed the P_1 of the whispered sounds for the majority of comparisons. Thus, to say that whispered vowel sounds *in general* manifest somewhat higher *F*-patterns and a pronounced increase for F_1 in particular when compared with voiced sounds is empirically contradicted: The results of the present study indicated that the spectra of many voiced vowel sounds can be expected to manifest P_1/F_1 comparable to or even higher than P_1/F_1 of whispered sounds if f_o (and also vocal effort) is varied. Figure 1 illustrates this finding. (Note in this context that, for most sound comparisons investigated here, vocal effort variation had a somewhat limited impact on acoustic differences related to voiced and whispered phonation when compared to extensive f_o variation.)

In sum, and also taking into account the nonuniform relation between vowel quality and spectral characteristics, we conjecture that spectral peak frequency differences between natural voiced and whispered vowel sounds depend on the level of f_o of the voiced sounds, on vocal effort variation, on additional aspects of the course of the spectral envelope of a sound and possibly also on the vowel quality investigated. Special attention should also be given to the observation that the entire spectral envelope of whispered vowel sounds was often indicated to correspond to the envelope of voiced sounds produced at intermediate f_o levels of a speaker's vocal range. This matter will be addressed and discussed in more detail in the following chapter.

For references, extended background information, details of experimental design, method and results, an extended discussion and documentation of results (tables including sound links), see the Materials, Chapter M5.2.

Figure 1. Comparison of P_1 for natural voiced and whispered vowel sounds, including f_o variation: Occurrence of similar or inverted lower spectral characteristics with respect to expected phonation-related spectral differences. Extract of Chapter M5.2, Table 1 (see Series 10, 19, 4 and 23 in this table). Each of the four comparisons shown consists of one voiced sound produced at a lower f_o level of the speaker's vocal range, one whispered sound, two voiced sounds produced with medium vocal effort at middle or higher f_o levels and one voiced sound produced with a high vocal effort at middle or higher f_o levels. Illustration of estimated P_1 (in Hz) for voiced vowel sounds produced at middle or higher f_o levels comparable to or higher than the P_1 of whispered vowel sounds (for P_1 estimation, refer to the spectrograms of the sounds in the Zurich Corpus). Sounds 1–5 = sounds of /y/ produced by the woman. Sounds 6–10 = sounds of /e/ produced by the child. Sounds 11–15 = sounds of /ø/ produced by the man. Sounds 16–20 = sounds of /o/ produced by the child. The *P*-patterns given in parentheses were estimated based on the peaks of the sound spectrum.
[C-05-02-F01] ↗

Figure 1. Comparison of P1 for natural voiced and whispered vowel sounds, including fo variation: Occurrence of similar or inverted lower spectral characteristics with respect to expected phonation-related spectral differences. [C-05-02-F01]



1–1 [ü] 220-V-med 1023-A-w [ü]
R175089 P1: c. 300

1–2 [ü] w-V-med 1023-A-w [ü]
R115940 P1: c. 450

1–3 [ü] 440-V-med 1023-A-w [ü]
R175086 P1: c. 470

1–4 [ü] 659-V-med 1023-A-w [ü]
R161266 P1: c. 670

1–5 [ü] 659-V-hgh 1023-A-w [ü]
R161497 P1: c. 660

1–6 [e] 262-V-med 1056-C-m [e]
R142996 P1: c. 500

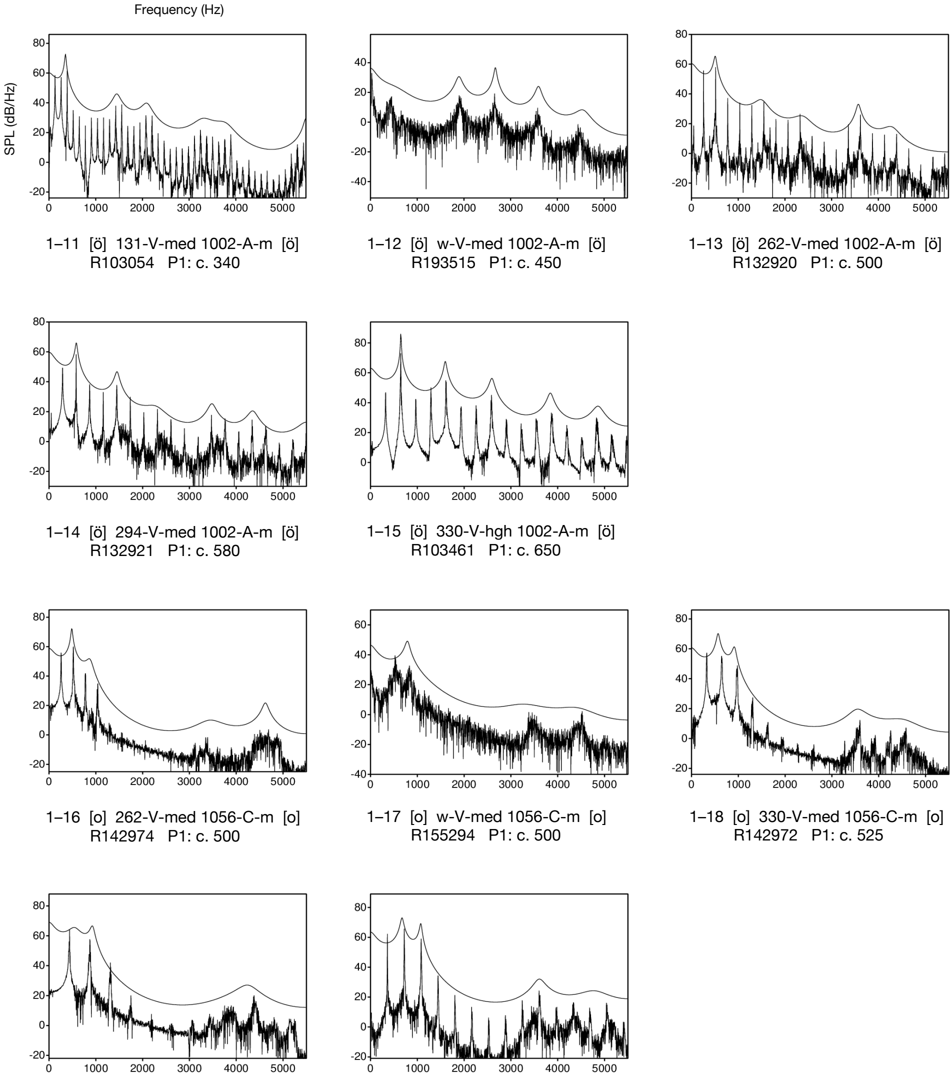
1–7 [e] w-V-med 1056-C-m [e]
R155285 P1: c. 530

1–8 [e] 294-V-med 1056-C-m [e]
R142995 P1: c. 570

1–9 [e] 330-V-med 1056-C-m [e]
R142994 P1: c. 620

1–10 [e] 330-V-hgh 1056-C-m [e]
R143143 P1: c. 650

Figure 1 (continuation). [C-05-02-F01]



1–11 [ö] 131-V-med 1002-A-m [ö]
R103054 P1: c. 340

1–12 [ö] w-V-med 1002-A-m [ö]
R193515 P1: c. 450

1–13 [ö] 262-V-med 1002-A-m [ö]
R132920 P1: c. 500

1–14 [ö] 294-V-med 1002-A-m [ö]
R132921 P1: c. 580

1–15 [ö] 330-V-hgh 1002-A-m [ö]
R103461 P1: c. 650

1–16 [o] 262-V-med 1056-C-m [o]
R142974 P1: c. 500

1–17 [o] w-V-med 1056-C-m [o]
R155294 P1: c. 500

1–18 [o] 330-V-med 1056-C-m [o]
R142972 P1: c. 525

1–19 [o] 440-V-med 1056-C-m [o]
R142969 P1: c. 550

1–20 [o] 349-V-hgh 1056-C-m [o]
R143119 P1: c. 640

5.3 Vowel Spectrum and Phonation Type III – Vowel Synthesis Based on Estimated Formant Patterns of Single Natural Vowel Sounds With Variation of Source Characteristics in Synthesis, and Related Vowel Recognition

Assuming that estimated *P*-patterns and *F*-patterns of natural vowel sounds produced with voiced and creaky phonation are comparable but that the patterns of whispered sounds increase in frequency levels, then both a synthesis of the creaky- and voiced-related patterns with a noise source and a synthesis of the whispered-related patterns with a voiced source should affect vowel quality recognition. (For the use of the term synthesis here, see below.) However, as explained, drawing such a straightforward conclusion does not take into account the relation of the spectrum of natural voiced vowel sounds to f_o: As indicated in the results of the experiments discussed in the previous chapter, assumed differences in the vowel quality-related P_1/F_1 for whispered and voiced sounds can disappear or even be inverted if the f_o level of the voiced sounds of comparison is varied. Moreover, the above direct and general conclusion also does not consider the unsystematic character of the spectral representation of vowel quality.

In light of this, an experiment was conducted investigating combinations of phonation-related *F*-patterns and source variation in vowel synthesis (Klatt synthesis) and their impact on vowel recognition. Based on the Zurich Corpus, for each of the eight long Standard German vowels, sound triplets of a man and a woman were compiled that included one whispered, one creaky and one voiced sound. All sounds were produced in V context with medium vocal effort and in nonstyle mode, and they were fully recognised according to the results of the standard listening test conducted when creating the corpus (100% vowel recognition rate matching vowel intention). Intended f_o for the voiced sounds related to 131 Hz for the man and 220 Hz for the woman, comparable to gender-specific average f_o levels as given in formant statistics for adult speakers. As a result, for each speaker and each of the eight vowels, three natural reference sounds were compiled, and a sample of 48 reference sounds in total was investigated.

In a visual examination of the spectrum, the spectrogram, the formant tracks and the LPC filter curve of these sounds, *F*-patterns for subsequent Klatt synthesis were estimated (for details, see Chapter M5.3). As a result, for each speaker and each of the eight vowels, three *F*-patterns related to the three natural reference sounds were assessed, and a total of 48 patterns (24 patterns per speaker) were investigated. Note that the estimation of *F*-patterns often proved to be difficult, above all for whispered vowel sounds, and manual corrections were often needed to approximate the estimated filter pattern for synthesis to the spectral envelope of the natural reference sound. Because of this, and because the source characteristics including phonation types were altered in the experiment, the sound production was termed synthesis and not resynthesis.

For vowel synthesis, every single *F*-pattern was combined with six different source characteristics, that is, noise and five levels of f_o: 65–131–220–262–393 Hz (*F*-patterns related to the sounds of the man) and 65–165–220–262–440 Hz (*F*-patterns related to the sounds of the woman), respectively. An f_o of 65 Hz was investigated in terms of imitating creaky sounds and f_o of 131 Hz (man) and 220 Hz (woman) were investigated referring to f_o of the natural voiced reference sounds and to gender-specific average f_o levels as given in formant statistics for adult speakers. As a result, for each speaker, each of the eight vowels and each of the three phonation-related *F*-patterns, six combinations of *F*-pattern and source characteristics were created, and a total of 288 sounds were produced using the KlattSyn synthesiser (cascade mode).

Finally, vowel recognition of the synthesised sounds was tested in two speaker-blocked subtests involving the five standard listeners, the subtests according to the standard procedure of the Zurich Corpus, with the exception of an experiment-specific sound presentation: One test item contained two sounds (separated by a 1 sec. pause), the first sound being the natural reference sound and the second sound being one of the six synthesised versions (*F*-pattern related to the natural reference sound, source characteristics set as one of the six options). The listeners were then asked to label the second sound only.

The main and most important indication resulting from the vowel recognition test concerned synthesised sounds related to *F*-patterns of natural whispered vowel sounds: If, in synthesis, these *F*-patterns were combined with either a creaky-like source (here voiced-like source with low f_o = 65 Hz) or a voiced-like source with gender-specific f_o levels as given in formant statistics for relaxed speech in adults (here f_o = 131 Hz for the man and 220 Hz for the woman), vowel recognition proved to be impaired for the sounds of some vowels, that is, vowel confusions in terms of recognised vowel qualities deviating from vowel intention occurred. But if these *F*-patterns were combined with either a noise source or a voiced-like source at an intermediate f_o level, higher than the one given in formant statistics for relaxed speech (here f_o = 262 Hz for the patterns of both speakers, except for one sound with f_o = 393 Hz), vowel quality was maintained for all sounds investigated independent of the source characteristic being noise or voiced. Notably, this intermediate level of f_o with the best match for vowel recognition was found for the sounds of both the man and the woman. For the highest f_o levels investigated in synthesis for a speaker, some vowel confusions occurred anew. Table 1 illustrates this main finding (compare the recognition results in Columns 9 and 10 with those in the other columns). In the table, sound links to the test items are included. This indication of vowel synthesis and vowel recognition paralleled the indication of the spectral comparison of natural whispered and voiced sounds reported in the previous chapter, that the lower part of the spectral envelope of whispered vowel sounds tended to correspond to the envelope of voiced sounds produced at intermediate levels of f_o of a speaker's vocal range.

But how should this finding be understood?

To introduce our conjecture, at this point of experimental findings and their interpretation, two reflections shall be anticipated in short that will be discussed in detail in the excursus on fundamental frequency and pitch and the following sixth main chapter: Firstly, although whispered vowel sounds have no periodicity (at least not in the sense in which a sound periodicity is generally defined), they are often perceived as having a pitch (for references, see Chapter M6.1). Secondly, measured f_o and pitch are two sound characteristics different in dimension: Fundamental frequency is a term used to refer either to a source characteristic of sound production or to an acoustic measure of radiated sound. These denotations pertain to physiology and to acoustics (physics). Pitch is a term used to refer to sound quality recognition. This denotation pertains to sound perception and recognition. Taking this into consideration, one possible explanation of the above indications is that pitch recognition of whispered sounds is comparable to pitch recognition of voiced sounds and that (as a rough approximation) the equivalent pitch level was indicated for the sounds investigated as lying near 262 Hz. (Note in this context that, according to the literature, recognised pitch levels of whispered sounds can vary to some extent. Here, the frequency estimate of the pitch levels of the whispered sounds is given only for the sounds investigated, with no further generalisation.)

If this is the case, for perception and acoustics, the link between whispered and voiced vowel sounds is not f_o but pitch (or if not pitch, then a comparable perceived sound characteristic; for this differentiation, see the following sixth main chapter).

This line of argument and conjecture leads to the need to separate f_o and pitch when discussing and analysing the acoustic characteristics of vowel sounds, a topic directly addressed in Chapter 6.

Besides whispering, if f_o levels of the synthesised voiced sounds corresponded to the levels as given in formant statistics, exchanging source characteristics for creaky- and voiced-related *F*-patterns had no substantial effect on vowel recognition for the sounds of the man, but it had an impact on the sounds of one close-mid and two close vowels of the woman in terms of an open–close vowel quality shift with increasing f_o (and pitch). However, as was to be expected based on previous experiments and the results thereof, more pronounced open–close shifts occurred for the voiced source condition with f_o levels of 262–393 Hz (sounds produced by the man) and 440 Hz (sounds produced by the woman; for details, see Chapter M5.3.)

In sum, if the sounds investigated here were set in a pitch order of 65–131–220–noise/262–393–440 Hz – the pitch at 65 Hz considered creaky-like and the pitch of the sounds with noise as a source presumed to be near 262 Hz – then the recognition results for the synthesised sounds of all *F*-patterns obtained accorded to the rule for vowel recognition of either being maintained or shifting (in a nonuniform manner) in an open–close direction with increasing pitch from low to high.

Why, then, not presume that all vowel sounds relate to pitch and that this relation is mirrored in the vowel spectrum, whether or not it manifests a harmonic structure and, with it, f_o? Why not speculate that pitch is not only the link between whispered and voiced vowel sounds but between all types of vowel sounds for both vowel recognition and acoustic vowel quality representation? Or, to be more cautious in speculation, if the actual perceptual reference of vowel quality does not prove to be pitch in general, then why not speculate that the reference is a perceived sound characteristic comparable to pitch?

For references, extended background information, details of experimental design, method and results, additional aspects of discussion and documentation of results (tables including sound links), see the Materials, Chapter M5.3.

Table 1. Synthesised vowel sounds related to *F*-patterns of natural whispered reference sounds, including source variation in synthesis: Vowel recognition results. Extract of Chapter M5.3, Table 3. Columns 1–5 = natural reference sounds (SP = gender of the speaker; S = sound series; V = intended and recognised vowel quality of the natural reference sound; Ref = number of the reference sound in the Zurich Corpus; P = phonation type of the reference sound). Columns 6–12 = synthesised sounds per source characteristics applied (see text; f_o levels for voiced-like source characteristic given in Hz) and links to the corresponding sound series in terms of the tested items (two sounds separated by a 1 sec. pause, see text). Colour code: Green = vowel recognition of the synthesised sound matched to the intended vowel quality of the reference sound (labelling majority, recognition rate of the synthesised sound ≥ 60%); dark red = vowel quality of the synthesised sound mismatched to the intended vowel quality of the reference sound (labelling majority, recognition rate of the synthesised sound ≥ 60%); light red = vowel quality of the synthesised sound mismatched to the intended vowel quality of the reference sound but only in terms of a limited shift to a vowel boundary.
[C-05-03-T01]

Table 1. Synthesised vowel sounds related to F-patterns of natural whispered reference sounds, including source variation in synthesis: Vowel recognition results. [C05-03-T01]

Natural reference sounds					Synthesised sounds vowel recognition ≥ 60%						
SP	S	V	Ref	P	▼	▼		rep	▼		L
					65	131	220	noise	262	393	
man	1	i	173175	whispered	i	i	i	i	i	i	↗
	2	y	173183		y	y	y	y	y	y	↗
	3	u	135837		o	o	o	u	o	u	↗
	4	e	135840		(ɛ–e)	(ɛ–e)	e	e	e	(e–i)	↗
	5	ø	135843		ɛ	(ɛ–ø)	ø	ø	ø	ø	↗
	6	o	173181		ɔ	o	o	o	o	u	↗
	7	ɛ	135860		ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	↗
	8	a	135839		a	a	a	a	a	a	↗
SP	S	V	Ref	P	▼		▼	rep	▼		L
					65	165	220	noise	262	440	
woman	9	i	141324	whispered	i	i	i	i	i	i	↗
	10	y	160347		e	(ø–y)	–	y	y	y	↗
	11	u	141320		o	o	o	u	u	u	↗
	12	e	141323		(ɛ–e)	e	e	e	e	i	↗
	13	ø	141326		(ɛ–ø)	ø	ø	ø	ø	y	↗
	14	o	141321		ɔ	ɔ	o	o	o	(o–u)	↗
	15	ɛ	160341		ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	↗
	16	a	160340		a	a	a	a	a	a	↗

5.4 Vowel Spectrum and Vocal Effort

The question of the acoustic differences of vowel sounds produced with different vocal efforts, including shouting, is a matter of debate. In the literature, the main reported acoustic features related to increased vocal effort are increased sound pressure level (SPL), increased f_o and F_1 (in some studies also F_2) and decreased spectral slope (emphasis of higher frequencies) combined with increased levels of the higher formants. These acoustic effects of vocal effort variation are generally understood as a consequence of a simultaneous change in respiratory, laryngeal and supralaryngeal behaviour (variation of subglottal pressure and vocal fold tension, adaptation of articulation).

In this treatise, the question regarding acoustic differences of vowel sounds produced with different vocal efforts is addressed in an observational manner only to extend the documentation and discussion of phonation- and articulation-related spectral variation of vowel sounds. In a corresponding experiment, vocal effort-related spectral differences of vowel sounds produced by men and women were investigated for two sound samples, the first sample involving sounds produced with low and high vocal efforts at f_o levels in the lower vocal range of the speakers, and the second sample involving sounds produced with low vocal effort and as shouted sounds at f_o levels in the middle vocal range of the speakers.

On the basis of the Zurich Corpus, for each of the eight long Standard German vowels, each of the two speaker groups of men and women and each of the two production parameters, low vocal effort and high vocal effort (systematically investigated in the corpus), two sounds of different speakers produced in nonstyle mode and V context at intended f_o levels in the ranges of 131–165 Hz (men) and 220–262 Hz (women) were selected. As a result, for each speaker group, for each of the eight vowels and f_o levels in the lower vocal range of the speakers, two sounds produced with low vocal effort were compared with two sounds produced with high vocal effort, resulting in two subsamples of 32 sounds per speaker group and in a sample of 64 sounds in total.

Likewise, for the same vowels, the same two speaker groups and each of the production parameters, low vocal effort and shouted, two sounds of different speakers produced in nonstyle mode and V context at intended f_o levels of 330 Hz (men) and 440 Hz (women), respectively, were selected. As a result, a second sample of 64 sounds produced at f_o levels in the middle vocal range of the speakers was created.

The selection of all sounds was based on two additional criteria: Full vowel recognition in the standard listening test conducted when creating the corpus (100% recognition rate matching vowel intention) and, if observable, marked differences in the lower part of the spectrum < c. 1 kHz related to vocal effort variation. The acoustic analysis of the sounds was conducted according to the standard procedure of the Zurich Corpus. On this basis, comparing the sounds of a vowel produced by speakers of a given speaker group, vocal effort-related differences of the acoustic characteristics were investigated concerning SPL, P_1/F_1 and P_2/F_2, L_{P2}/L_{F2} and alpha ratio (level difference between the average SPL of the 1–5.5 kHz frequency region and the average SPL of the 0.05–1 kHz frequency region).

As a main result and mostly in line with the indications given in the literature, an increase in vocal effort from low to high or shouted resulted in an increase of SPL in general and an increase of spectral energy in the higher frequency range > 1 kHz in particular (alpha ratio), the higher frequency range commonly assumed to be related to $F2$ of sounds of front vowels and /a/ and to the higher formants of sounds of all vowels. As a consequence, for almost all sound comparisons for which L_{P2}/L_{F2} could be estimated, L_{P2}/L_{F2} also increased with increasing vocal effort. Further, with only two exceptions, P_1/F_1 (or the spectral centre of gravity of the frequency range generally assumed to be related to P_1/F_1) in its turn increased markedly with vocal effort for the sounds of all eight vowels produced at f_o levels in the lower vocal range of the speakers. However, this only held true for the sounds of close-mid, open-mid and open vowels produced at f_o in the middle vocal range of the speakers (except for one comparison for which no estimate could be generated). For the sounds of close vowels, no such increase was observed. Thus, the increase of P_1/F_1 proved to be dependent on f_o levels and vowel qualities (vowel openness).

In the context of the present treatise, the sometimes striking spectral differences in the frequency region commonly assumed to be related to P_1/F_1 deserve special attention. For the investigated sounds of a vowel produced at comparable f_o levels that show vocal effort-related differences in the lower vowel spectrum, these differences were very pronounced: Estimated P_1/F_1 of the sounds produced with a high vocal effort surpassed P_1/F_1 of the sounds produced with a low vocal effort by 100 Hz or more, a frequency difference that approximates or equals F_1 differences of two adjacent vowel qualities as given in formant statistics. (For P_1/F_1 differences due to a combination of vocal effort and f_o variation that can equal F_1 differences of two non-adjacent vowel qualities, see Chapter 7.8.)

In these terms, the findings of the present experiment again point to the large range and the nonuniform character of variation of vowel-related spectral characteristics.

No clear indication was found for vocal effort-related variation of P_2/F_2 (increased second peak frequencies as a general result of increasing vocal effort).

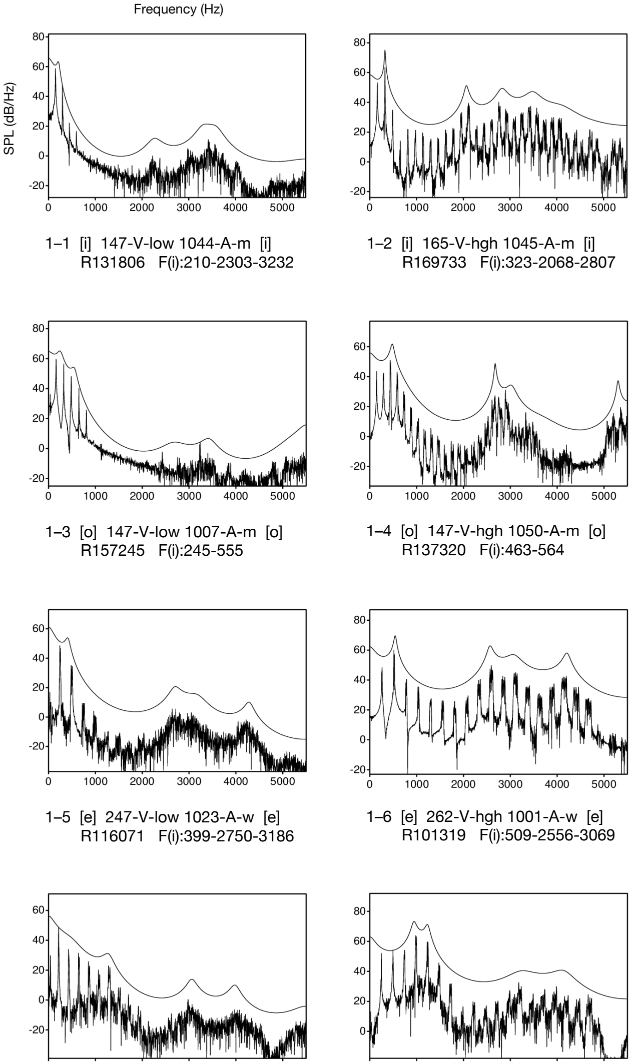
In contrast to some interpretations given in the literature, it is noteworthy that an increase in f_o is not directly linked to an increase in vocal effort: Vocal effort can be altered independently of f_o variation, and f_o levels of sounds with low vocal effort can be much higher than f_o levels of sounds with high vocal effort.

For references, extended background information, details of experimental design, method and results, an extended discussion (including some relativisations concerning the methodological substantiation of spectral peak and formant frequency estimation, vowel timbre variation and register changes) and documentation of results (tables including sound links), see the Materials, Chapter M5.4.

Figure 1: Spectral differences of vowel sounds related to vocal effort variation: Four pairs of sounds produced by adult speakers of the same gender with low and high vocal effort at f_o levels in their lower vocal range. Extract of Chapter M5.4, Table 1 (see Series 1, 6, 12 and 16 in this table). Illustration of increased SPL, alpha ratio, P_1/F_1 and L_{P2}/L_{F2} as a result of increased vocal effort. Sounds 1–4 = sound pairs of /i/ and /o/ produced by men at intended f_o in the range of 147–165 Hz. Sounds 5–8 = sound pairs of /e/ and /a/ produced by women at intended f_o in the range of 220–262 Hz.
[C-05-04-F01] ↗

Figure 2: Spectral differences of vowel sounds related to vocal effort variation: Three pairs of sounds produced by adult speakers of the same gender with low and high vocal effort (shouted) at f_o levels in their middle vocal range. Extract of Chapter M5.4, Table 2 (see Series 1, 16 and 14 in this table). Illustration of the role of f_o ranges and vowel openness for vocal effort-related differences in the lower vowel spectrum, of occurring spectral peak and formant frequency estimation problems and of occurring lack of validation of F-pattern estimation in Klatt resynthesis. Sounds 1 and 2 = sound pair of /i/ produced by two men at an intended f_o of 330 Hz. Comparing the two spectra, no vocal effort-related increase in P_1/F_1 is identifiable in contrast to the sounds of that vowel produced at f_o levels in the lower vocal range of the speakers (see Figure 1). Sounds 3 and 4 = sound pair of /a/ produced by two women at an intended f_o of 440 Hz. Methodological substantiation of F_1–F_2 estimation is weak for both sounds because of either a flat spectral envelope or only one peak < 1.5 kHz. However, vowel quality is maintained in Klatt resynthesis based on the estimated F-patterns and f_o levels of both natural reference sounds (author's estimate). Sounds 5 and 6 = sound pair of /o/ produced by two women at an intended f_o of 440 Hz. Methodological substantiation of F_1–F_2 estimation is again weak for both sounds due to the formant frequencies equalling harmonic frequencies. Moreover, contrary to the previous sound pair, the vowel quality of both natural sounds of this pair is not maintained in Klatt resynthesis based on the estimated F-patterns and f_o levels of the natural references (author's estimate).
[C-05-04-F02] ↗

Figure 1. Spectral differences of vowel sounds related to vocal effort variation: Four pairs of sounds produced by adult speakers of the same gender with low and high vocal effort at fo levels in their lower vocal range. [C-05-04-F01]



1–1 [i] 147-V-low 1044-A-m [i]
R131806 F(i):210-2303-3232

1–2 [i] 165-V-hgh 1045-A-m [i]
R169733 F(i):323-2068-2807

1–3 [o] 147-V-low 1007-A-m [o]
R157245 F(i):245-555

1–4 [o] 147-V-hgh 1050-A-m [o]
R137320 F(i):463-564

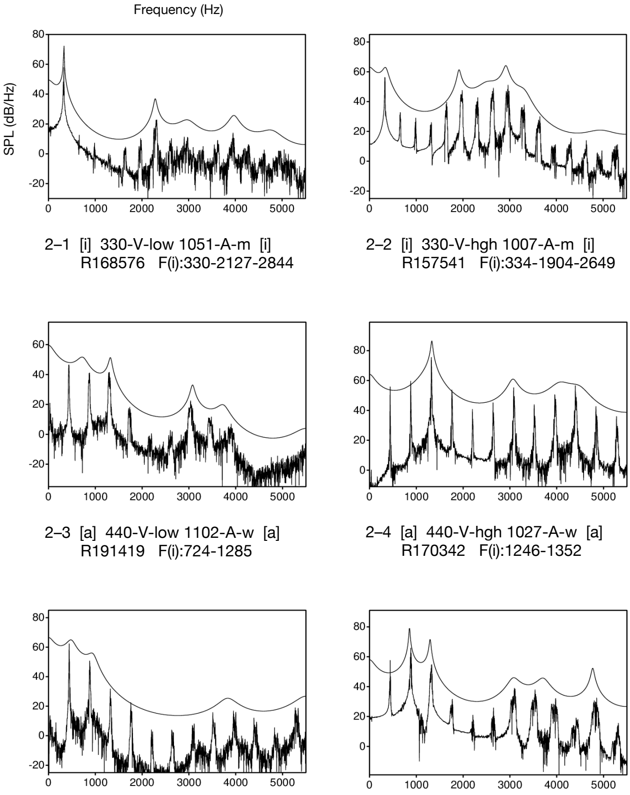
1–5 [e] 247-V-low 1023-A-w [e]
R116071 F(i):399-2750-3186

1–6 [e] 262-V-hgh 1001-A-w [e]
R101319 F(i):509-2556-3069

1–7 [a] 220-V-low 1006-A-w [a]
R114567 F(i):580-1284

1–8 [a] 262-V-hgh 1046-A-w [a]
R147546 F(i):937-1228

Figure 2. Spectral differences of vowel sounds related to vocal effort variation: Three pairs of sounds produced by adult speakers of the same gender with low and high vocal effort (shouted) at fo levels in their middle vocal range. [C-05-04-F02]



2–1 [i] 330-V-low 1051-A-m [i]
R168576 F(i):330-2127-2844

2–2 [i] 330-V-hgh 1007-A-m [i]
R157541 F(i):334-1904-2649

2–3 [a] 440-V-low 1102-A-w [a]
R191419 F(i):724-1285

2–4 [a] 440-V-hgh 1027-A-w [a]
R170342 F(i):1246-1352

2–5 [o] 440-V-low 1046-A-w [o]
R160043 F(i):478-945

2–6 [o] 440-V-hgh 1036-A-w [o]
R171024 F(i):832-1267

5.5 Conclusion

As explained, the re-examinations and sound examples presented in this main chapter primarily aim to extend the documentation and discussion of possible spectral variation for sounds of a given vowel. For sound configurations and comparisons similar to the experimental settings of the studies available in the literature, the general predictions of phonation- and articulation-related spectral differences were confirmed for the majority of the sounds documented here. The main limitations and relativisations of confirmation are discussed above in short terms and in the Materials in detail. In contrast, for sound configurations and comparisons with substantial f_o level variation for voiced sounds, the general predictions regarding the comparison of voiced and whispered sounds given in the literature were not confirmed for the sounds documented in Chapters 5.2 and 5.3. Moreover, the general predictions for vocal effort variation, as given in the literature, proved to be dependent on f_o ranges and vowel qualities. Here, these two findings are considered key to an advance in understanding the recognition and acoustic representation of vowel quality. Therefore, the focus of this conclusion lies in the discussion of these two findings.

As for the main observations on the comparison of voiced and whispered sounds, firstly, the phonation-related acoustic differences reported in the literature also occurred for the majority of the sound comparisons in our first re-examination (Chapter 5.1), in which voiced sounds were produced with a medium vocal effort at f_o levels as given in formant statistics. (However, note that some speaker- and vowel-related divergences from the predictions also occurred and that there is often only a weak methodological basis for spectral peak estimation of whispered sounds.) Yet, secondly, if the comparison of voiced and whispered sounds included voiced sounds with a variation of f_o levels, the differences in P_1 (or, more generally, the differences in the lower part of vowel spectrum of the assumed P_1/F_1 frequency range) were found to disappear or even be inverted (Chapter 5.2). Thirdly, the indication of voiced–whispered differences in the lower part of the vowel spectrum being dependent on the f_o of the voiced sounds of comparison was supported in the vowel synthesis experiment presented in Chapter 5.3: Formulated in general terms, if synthesis was based on whispered-related F-patterns combined with voiced source characteristics at lower f_o levels (comparable to average f_o levels as given in formant statistics), for part of the sounds, vowel confusion occurred. But if, in synthesis, the f_o level was increased stepwise up to a middle f_o level, vowel confusion gave way to identical vowel recognition for both whispered and voiced sounds. And if f_o was further increased, vowel confusion occurred again for part of the sounds.

It is well known that many listeners perceive whispered sounds as having a pitch, although no corresponding f_o level can be measured. (Note again that creaky sounds also are perceived as having a pitch, although the periodicity of the sounds is irregular.) Considering this pitch phenomenon for natural whispered vowel sounds in general as well as the findings regarding synthesising vowel sounds based on whispered-related *F*-patterns combined with voiced source characteristics and stepwise increased f_o levels in particular, we have conjectured that even if voiced and whispered sounds cannot be compared with regard to f_o, they may be comparable with regard to pitch: If natural voiced and whispered vowel sounds are set in the order of their pitch level, the level of whispered sounds being perceived as somewhat above 220 Hz, then vowel-related spectral peak frequency patterns of voiced and whispered vowel sounds may not manifest substantial differences. If this is the case, to repeat, the link between vowel-related acoustic characteristics of whispered and voiced vowel sounds – and possibly the link between vowel-related acoustic characteristics of all types of vowel sounds – is pitch, and pitch and f_o have to be separated for the acoustics of the vowel, a topic directly addressed in the following excursus and the sixth main chapter.

However, some relativisations have to be made regarding the findings presented: Above all, no details of the impact of a further extended variation of production parameters for voiced and whispered vowel sounds (including whisper substyles and intra- and inter-speaker whisper variation) were investigated. According to the findings reported up to here, it has to be expected that an extended variation of production parameters will have a substantial impact on the spectral characteristics in question.

As for the main observations on the comparison of sounds produced with very different vocal efforts, with two major exceptions, the phonation- and articulation-related acoustic differences reported in the literature also occurred for the majority of the sound comparisons in our last re-examination (see Chapter 5.4). The first major exception concerned the dependence of spectral differences concerning the lower part of the vowel spectrum: These differences were indicated to depend on f_o levels and vowel openness of the sounds compared. The second major exception concerned the supposed general association of increased vocal effort with increased f_o levels: f_o variation is not a general characteristic related to vocal effort variation since vocal effort can be varied independently of f_o.

The vocalises and the vowel sounds produced at high f_o levels discussed in the second main chapter demonstrated the large variability of occurring lower spectral energy for sounds of a vowel due to f_o variation, the documented spectral differences for sounds of a vowel far exceeding the differences reported in formant statistics for sounds of two adjacent vowels. Likewise, the sound comparisons of this fifth main chapter demonstrated another type of large variability of occurring lower spectral energy for sounds of a vowel due to phonation type or due to vocal effort variation, in its turn documenting numerous sounds of a single vowel quality for which their spectral differences exceed the differences reported in formant statistics for sounds of two adjacent vowels.

Excursus – Fundamental Frequency and Pitch

Introduction

In all of the discussions thus far – references made to the actual vocal range of recognisable vowel sounds and the documentation and discussion of occurring spectral variation due to variation of f_o (main Chapter 2), the evidence given for formant pattern and spectral shape ambiguity (main Chapter 3), the relativisation made for supposed age- and gender-related spectral characteristics (main Chapter 4) and the investigation and documentation of spectral characteristics of comparisons of vowel sounds produced with different phonation types or vocal effort (main Chapter 5) – the vowel spectrum was related to f_o without further questioning, with few exceptions. (The same held true for most of our earlier investigations on natural vowel sounds and related publications; see below.) This restriction may be adequate for a mere description and documentation of spectral characteristics of unmanipulated natural vowel sounds and of (re-)synthesised sounds with a uniquely defined periodicity as a source (harmonic structure of the sound spectrum, with f_o, H_1 and HCF according with each other). For a more profound investigation and understanding of the vowel sound and its acoustic representation and with regard to a future theory, however, it is not sufficient.

As indicated above, the question to be answered is: Why should a variation in fundamental frequency cause primary changes in the spectral representation of vowel quality?

With reference to our considerations in Chapter 5.3, in order to anticipate the answer and to introduce the following argument and subsequent investigation: The variation of fundamental frequency is not the cause of the observed phenomena reported in this treatise because it is not fundamental frequency but pitch (or a comparable perceptual referencing to a sound pattern repetition over time) to which the vowel sound and its spectral representation relate. More precisely: Within the vocal range of recognisable vowel sounds, besides occurring effects of source–filter coupling and ignoring methodological difficulties of vowel-related spectral analysis (*F*-pattern, spectral envelope), there is no physical or physiological reason for fundamental frequency to affect vowel quality and its acoustic representation in the way documented and discussed here. Presuming that evidence can be provided for this conclusion, the observed relation of the sound spectrum of a vowel to the fundamental frequency of unmanipulated natural vowel sounds or (re-)synthesised sounds with an unambiguously defined periodicity as a source turns out to be but an indication that vowel recognition relates to pitch or, formulated more cautiously, to a perceived sound characteristic comparable to pitch.

Terms and perspectives

Fundamental frequency is a term used to refer either to a source characteristic of sound production or to an acoustic measure of the radiated sound and its periodicity. These denotations pertain to acoustics (physics) and physiology. Pitch is a term used to refer to sound quality recognition. This denotation pertains to sound perception and recognition.

f_o measurement of a radiated sound depends on an algorithm being applied. Pitch level assessment depends on a recognition test involving listeners.

f_o is often understood as being directly related to a sound periodicity generally represented in the sound spectrum by the H_1 and by the HCF of the harmonics in the spectrum. Pitch is also often understood as being related to a sound periodicity. It is, however, long and well-known that this relation is not imperative: "[…] pitch is not correlated, in any simple way, with stimulus periodicity, spectral content, or wave-form fine structure." (Wightman, 1981)

Fundamental frequency, source and filter, acoustics, perception subordinated to physiology

From a purely acoustical (physical) perspective of the source–filter model of speech production, besides the effects of source–filter coupling, f_o variation does not affect the filter.

From a physiological perspective, however, the relation of source and filter in terms of phonation and articulation is not directly comparable to a purely physical model: Some changes in the vocal tract configuration in the production of sounds of a vowel occur not only for different phonation types, different vocal efforts and different production styles but also for register changes when f_o is substantially varied. (Besides, additional source–filter interactions also occur; see Titze, 2008; Titze and Palaparthi, 2016.) But put into simple terms – given voiced sounds of a close-mid vowel produced by a woman in isolation (V context) with medium vocal effort in nonstyle mode at an f_o of 200 Hz, and given a one-octave change in f_o from 200 Hz to 400 Hz, to the best of our knowledge and ignoring the role of self-perception for sound production, there is no physiological reason (related to vocal fold vibration and resonance characteristics of the vocal tract) for the observed change in the spectral envelope if the vowel quality of the sounds is maintained.

Concerning formant measurement of (quasi-)periodic sounds, spectral sampling of the supposed filter of sound production is related to f_o in such a way that, with increasing f_o levels, the frequency distance of the harmonics also increases, and the resonance frequencies of sound production may not be adequately indicated in the sound spectrum. However, this measurement problem does not concern all *R*-patterns and f_o level variations in general terms. It cannot explain the observed relation of the sound spectrum of a vowel to f_o. As shown in the model synthesis experiments described in Chapters 3.3 and 3.4, sounds can be produced based on interrelated resonance patterns and f_o levels in terms of all resonance frequencies of a pattern being multiples (in whole numbers) of f_o for two or three f_o levels, and in these "ideal" cases of resonances of production and harmonic spectra of the radiated sounds, the dominant harmonics always coincide with the resonances, and the only acoustic differences between the sounds are f_o (and pitch) and the frequency distances of harmonics. Thus, increasing f_o does not generally lead to undersampling of the resonance curve, but the undersampling is dependent on individual configurations of f_o and resonance frequencies. Yet, as shown for these "ideal" cases of interrelated resonance patterns and f_o levels, f_o variation will often cause a vowel quality shift.

In these terms, and as already indicated in Chapter 3.6, there is no identifiable acoustic cause that would explain the observed relation of the sound spectrum of a vowel to f_o, and the same holds true for a physiological perspective if the actual role and agency of perception for sound production is not taken into account.

Pitch, perception, physiology

This brings perception into focus. Suppose the cause for the observed relation of the sound spectrum of a vowel to f_o is a perceptual one, as brought up for discussion here. In that case, f_o is but an acoustic manifestation indicating that perception refers to pitch in vowel quality recognition or, if not to pitch, then another comparable perceived sound characteristic. The first option is addressed here; the second option will be discussed in the context of the experiments described in the following chapters.

In prevailing theory, pitch (aside from its possible contribution to sound timbre) is not considered a principal agent for shaping the vowel quality of vowel sounds. According to the prevailing source–filter model of speech production, f_0 and pitch (assumed to be associated) are source characteristics (or, more precisely, source characteristics associated with a perceptual quality), and vowel quality is a result of sound shaping caused by the filter, both production parts being quasi-independent. At the same time, it is an essential characteristic of vowel quality that it is recognisable as the same quality for all levels of f_0 and pitch within the vocal range of recognisable vowel sounds. Thus, the indication of the spectrum of natural vowel sounds being related to f_0 contrasts prevailing theory. Since there is no physical or physiological explanation for this, we conclude that the perceptual process of vowel quality recognition is linked to or even based on pitch recognition. We will later explain why this conclusion may have to be formulated in broader terms, considering possible differences between perception and recognition of vowel quality and pitch.

But how to obtain evidence for such a vowel–pitch relation?

When this question came up in the course of our investigation, we first used an exploratory and stepwise approach to create experimental designs to demonstrate sounds in which recognised vowel quality and recognised pitch level were related, but pitch either did not correlate with any harmonic structure of the sound spectrum or did not correlate with all three acoustic features of measured f_0, *H*1 and HCF. As a result, we conducted experiments in which vowel qualities and pitch levels or vowel quality shifts and pitch level differences were examined, but (i) the pitch level (if recognised) was not related to a sound spectrum with a harmonic structure (and thus it was not related to any of the three features of f_0, *H*1 and HCF; whispered vowel sounds, see Chapters 6.1–6.3), (ii) the pitch level was expected to relate to only HCF for sounds lacking *H*1 (suppressed fundamental, see Chapter 6.4), (iii) the pitch level was expected to relate to only *H*1 for sounds lacking HCF (sinewave vowel sounds without HCF, see Chapters 6.5 and 6.6), and (iv) the pitch level was expected to relate to HCF in contrast to *H*1 (three-sinusoid vowel sounds with HCF, see Chapter 6.7). In these experiments, we expected that pitch and its relation to vowel quality could be shown as standing in contrast to acoustic features generally assumed to be features of fundamental frequency. (For corresponding earlier indications given in the literature, see below.)

During the creation and in the process of the experiments, for some of the sounds examined, vowel quality and/or pitch proved to be ambiguous in that two qualities and/or two (or even more) pitch levels could be identified by the listeners in the listening tests. Therefore, in the context of a sinewave-like synthesis study based on two or three dominant or prominent harmonics extracted from natural sounds (the harmonics representing spectral peaks of these sounds), we began to address the matter of double-vowel and/or double-pitch recognition by not only testing dominant or prominent vowel qualities and pitch levels but also asking the listeners to label secondary qualities and/or secondary pitch levels (see Chapter 6.8). Further developing this experimental approach, we subsequently created experimental designs to demonstrate parallel shifts from one vowel quality and one pitch level to another quality and another pitch level, including a transitional phase with occurring double-vowel and/or double-pitch recognition (Chapters 6.9 and 6.10), motivated by the following idea: If cases of single sounds with simultaneous recognition of two vowels and two pitches can be demonstrated, pitch would neither relate to fundamental frequency nor to *H*1 or HCF because, for each of these three acoustic features, only one measure pertains to a single sound and its harmonic spectrum. Furthermore, if double-vowel and/or double-pitch recognition in general proves to be related to more open qualities for lower pitch levels and more close qualities for higher pitch levels, the systematic character of the vowel–pitch relation could be shown.

If all this can be demonstrated, the role of perception for vowel sound production – and thus the interplay of perception and physiology of the voice – would have to be fundamentally reconsidered. These reflections and this experimental development represent the background of the following main chapter.

Additional note on references

Pitch perception is a highly complex matter, and the literature proposes different theoretical models (for overviews, see Wightman, 1981; Houtsma, 1995; Plomp, 2002, Chapters 2 and 3; de Cheveigné, 2005; Fastl and Zwicker, 2006, Chapter 5; Yost, 2009; Gelfand, 2018, Chapter 12; Niebuhr et al., 2020). As a consequence, no simple and general reference to a relation of pitch and its acoustic correlates can be presumed here. As Niebuhr et al. (2020, p. 34) conclude, “[…] pitch perception in complex sound signals relies on multi-layer, signal-adaptive cognitive mechanisms in which f0 is neither required to be physically present nor directly translated into its psychoacoustic counterpart. Pitch is virtual […].”

Fundamental frequency is also no trivial matter. Above all, different algorithms can produce different measures, and the supposed parallelism between measured f_o, $H1$ and HCF is not imperative for all types of recognisable vowel sounds.

However, it is beyond the aim and scope of the present treatise to provide a correspondingly detailed discussion. Instead of adopting a general theoretical perspective on the matter of pitch, here, pitch is investigated in the simple terms of the recognition of a pitch level of a vowel sound or a level difference between two sounds compared (e.g. asking listeners whether, if listening to a sound, they are able to match the pitch level of that sound to a pitch level on a piano, or asking listeners whether, if comparing two sounds, they recognise pitch level differences). Concerning f_o, a measurement procedure comparable to a standard procedure often referred to in the literature was adopted (for details, see the standard procedure of the Zurich Corpus) with no further investigation and valuation of the procedure in terms of a comparison of measurements of different algorithms. Regarding $H1$ and HCF, an estimation was made on the basis of the sound spectra, which we considered unproblematic for the sounds investigated.

Hence, the description and discussion of the vowel–pitch experiments presented here are not embedded in the general debate on pitch recognition. This embedment will be a demand for future research. However, even a glance at the many topics of the general debate calls attention to general phenomena of pitch perception (most of them already addressed in the early period of modern pitch research), which are highly relevant for the present investigation of vowel sounds (only selected references are given here):

- Sounds with a “missing” fundamental, lower resolved harmonics versus higher unresolved harmonics (residue), and pitch (Seebeck, 1843; Fletcher, 1924; Schouten, 1938, 1940; Licklider, 1959; Schouten et al., 1962; Houtsma and Smurzynski, 1990; Shackleton and Carlyon, 1994; de Cheveigné, 2005, Chapter 10.4; Jackson and Moore, 2013)
- Sounds with more than one pitch (von Helmholtz, 1863, Chapter 4; Schouten, 1938; Jenkins, 1961; Schouten et al., 1962; Houtsma, 1995, pp. 281–282; de Cheveigné, 2005, Chapter 10.6)
- Sounds with inharmonic spectra and pitch (Jenkins, 1961; Schouten et al., 1962; Plomp, 1967)
- Noise and pitch (Small and Daniloe, 1967; Fastl, 1971; Fastl and Zwicker, 2006, pp. 125–129)

- Pitch recognition and particular sound timbre characteristics, attention and strategies of the listeners and design of listening tests (von Helmholtz, 1863, pp. 84–89; Jenkins, 1961; Smoorenburg, 1970; Ladd et al., 2013).

The experiments of the following sixth main chapter reflect these general aspects. The same holds true for the following additional notes.

Additional note on sound periodicity

The distinction between (quasi-)periodic and aperiodic sounds, e.g. voiced and whispered sounds, is one of the standards in acoustics. However, if pitch is the link between the acoustic characteristics of these two types of sounds, then the term periodicity may have to be reconsidered.

Additional notes on pitch recognition tests and the parallelism of vowel and pitch recognition

According to our understanding, and adding further aspects to the above indications given in the literature, the interpretation of the results of pitch recognition tests has to consider (i) the differences between sound types, (ii) the possibility of two (or more) concurring pitches, (iii) general listener-specific recognition strategies, (iv) listener-specific recognition consistencies or inconsistencies, and (v) the sound context. Furthermore, regarding the parallelism of vowel quality and pitch recognition, caution should be exercised to expect a uniform character of this parallelism.

Terminological correction

In our earlier descriptive studies (on natural vowel sounds mostly) and the related publications, we used phrases such as f_o-dependence of vowel-specific spectral characteristics (formants, spectral shape) or correlation between f_o and vowel-specific spectral characteristics. The above consideration indicates that these phrases have to be corrected. In precise terms, as will be concluded below, spectral representation of vowel quality relates to pitch (or to a comparable perceptual referencing to a sound pattern repetition over time, hereafter often abbreviated as “alternative” to the presumed vowel–pitch relation).

6 Vowel Sound, Vowel Spectrum and Pitch

6.1 Pitch Recognition Comparing Natural Whispered and Voiced Vowel Sounds for Utterances of Single Speakers

Many listeners perceive whispered speech, and with it whispered vowel sounds, as having pitch.

In the previous main chapter, it was demonstrated that the lower part of the spectral envelope of natural whispered vowel sounds tended to correspond to the lower part of the spectral envelope of voiced sounds if the voiced sounds were produced at intermediate levels of f_o of an adult speaker's vocal range. This finding was confirmed in vowel resynthesis since vowel quality was maintained for almost all sounds resynthesised based on *F*-patterns of natural whispered sounds and applying a voiced-like source with an f_o of 262 Hz, but vowel confusions occurred for both lower and higher f_o levels applied. Questioning this finding, adding it to the general observation of the vowel-related (lower) spectrum of voiced sounds being dependent on f_o and taking into consideration the pitch perception in whispered speech, in Chapters 5.3 and 5.5 and the excursus on fundamental frequency and pitch, we have presumed that perception and acoustic characteristics of whispered vowel sounds – and indeed of all vowel sounds – relate to pitch (or to a comparable perceived sound characteristic). In consequence, the present main chapter addresses this vowel–pitch relation thesis.

The first three experiments on the vowel–pitch relation continued to investigate whispered vowel sounds, that is, sounds with no measurable f_o. Posing the question of recognition and acoustic characteristics of whispered vowel sounds relating to pitch, various preliminary listening test trials with the standard listeners of the Zurich Corpus were first conducted. In the course of these trials, three experimental designs were created, addressing identifiable pitch level differences in (i) an intra-speaker comparison of natural whispered and voiced sounds of a given vowel based on the sound sample examined in Chapter 5.2, (ii) an intra-speaker comparison of resynthesised whispered- and voiced-like sounds of a given vowel based on the sound sample examined in Chapter 5.3, and (iii) an inter-speaker comparison (including different genders and ages of speakers) of natural whispered sounds of a given vowel and, in parallel, an inter-speaker comparison of resynthesised whispered-like sounds of a given vowel again relating to the sound samples examined in Chapters 5.2 and 5.3.

This first chapter on the matter of whisper and pitch describes and discusses the first experiment, the intra-speaker comparison of natural whispered and voiced sounds. The other two experiments are the subjects of the subsequent chapters.

Based on the sample of whispered and voiced sounds of a man, a woman and a child investigated in Chapter 5.2, for each of the three speakers and each single vowel quality, three sounds were selected: The whispered sound, the voiced sound produced at a low intended f_o level (131 Hz for the man, 220 Hz for the woman and 262 Hz for the child) and the voiced sound produced with medium vocal effort at the highest intended f_o level (see Table 1 in Chapter M5.2, sounds 1, 2 and 4 of a speaker and a vowel; for details of the sound sample investigated, see Chapter M6.1).

Pitch recognition of the compiled sounds was investigated in three speaker-blocked listening subtests according to the standard procedure of the Zurich Corpus and involving the five standard listeners of the corpus, with the following experiment-specific adaptation: Each single test item contained one whispered and one voiced sound (low or high f_o level) of the same vowel and produced by the same speaker (separated by an approximately 1 sec. pause), in AB (whispered–voiced) and BA (voiced–whispered) order. The listeners were asked to listen to the sounds and to identify the pitch level difference between the first and the second sound as falling, flat or rising, referring to dominant or prominent pitches.

The main pitch recognition results obtained were as follows. (Note that, in the following, f_o levels are given in terms of intended f_o, and pitch recognition results are given according to the labelling majority.)

For all sound comparisons of the man, the pitch level of the whispered sounds was recognised as above the level of the voiced sounds produced at a lower f_o of 131 Hz. Inversely, the pitch level of the whispered sounds was recognised as below the level of the voiced sounds produced at higher f_o, with a frequency range of these higher levels of 294–494 Hz, depending on vowel quality.

For the sound comparisons of the woman, the pitch level of the whispered sounds was recognised as either above the level of the voiced sounds produced at the lower f_o of 220 Hz or as equal to this level. Inversely, the pitch level of all whispered sounds was recognised as below the level of the voiced sounds produced at higher f_o, with a frequency range of these higher levels of 330–784 Hz.

For the sound comparisons of the child, the pitch level of the whispered sounds was again recognised as either above the level of the voiced sounds produced at a lower f_o of 262 Hz or equal to this level. Inversely, with two exceptions, the pitch level of the whispered sounds was recognised as below the level of the voiced sounds produced at higher f_o, with a frequency range of these higher levels of 440–659 Hz. The two remaining comparisons concerned whispered and voiced sounds of /e, ø/, with f_o levels of the voiced sounds in the frequency range of 330 Hz, surpassing 262 Hz only by four semitones. For these two comparisons, a weak labelling majority regarding flat pitch and a labelling minority regarding higher pitch levels for voiced than for whispered sounds were found.

Besides, the pitch recognition results for the AB order versus the BA order were somewhat inconsistent, with labelling inconsistency scattered among sound comparisons and listeners. However, no opposite low–high or high–low identifications occurred except for two sounds labelled by one of the listeners.

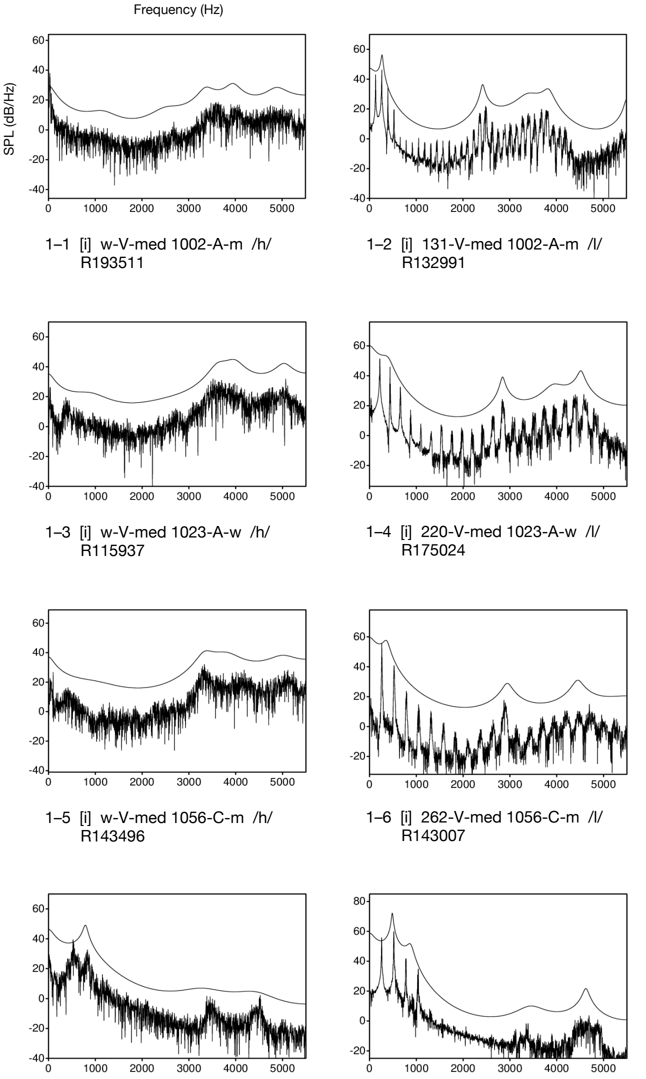
Notably, concerning voiced sounds produced in the lower vocal range of a speaker, the tendency for pitch levels to be identified as lower than the levels of the whispered sounds of comparison was more pronounced for the sounds of the man than of the women and the child. Considering the fact that the average f_o level of these voiced sounds of the man was nine semitones below the corresponding level of the woman and one octave below the corresponding level of the child, the results thus indicated that the pitch levels of the whispered sounds of the man were recognised closer to the levels of the whispered sounds of the woman and the child than to the average f_o level of 131 Hz of his voiced sounds. In contrast, no inter-speaker differences were found for the pitch levels of whispered sounds when compared with the levels of the voiced sounds produced at higher $f_o > 400$ Hz.

Thus, in sum, the pitch level of a voiced sound produced in the lower vocal range of a speaker was recognised as either lower (majority of cases) or equal to the level of the whispered sound of comparison, and the pitch level of a voiced sound produced in the middle and higher vocal range of a speaker was recognised as either equal to or higher than (majority of cases) the level of the whispered sound. Figure 1 illustrates this main finding. In terms of a first rough estimation, we conclude that the pitch level of the whispered vowel sounds investigated here fell somewhere in the frequency range of 200–400 Hz. Such an estimate is, at least for a substantial part, in line with the findings of the spectral comparison of natural whispered and voiced vowel sounds reported in Chapter 5.2 and the resynthesis of natural whispered vowel sounds reported in Chapter 5.3.

For references, extended background information, details of experimental design, method and results, an extended discussion and documentation of results (tables including sound links), see the Materials, Chapter M6.1. For the notation of pitch recognition results and pitch level differences, see the paragraph on figures and figure legends in the Introduction.

Figure 1. Comparison of phonation-related pitch levels of natural whispered and voiced vowel sounds produced by single speakers: Illustration of the main finding. Extract of Chapter M6.1, Table 1 (see Series 1, 9, 17, 23, 16, 2, 14 and 24 in this table). Sounds 1–6 = three sound pairs in terms of sound comparisons for which the pitch level of the whispered sound was recognised as above the level of the voiced sound produced at f_o levels in the lower part of the speaker's vocal range. Sounds 7–10 = two sound pairs in terms of sound comparisons for which the pitch level of the whispered sound was recognised as approximately equal to the level of the voiced sound, the voiced sound again produced at an f_o level in the lower part of the speaker's vocal range. Sounds 11–16 = three sound pairs in terms of sound comparisons for which the pitch level of the whispered sound was recognised as below the level of the voiced sound produced at an f_o level in the higher part of the speaker's vocal range. In the figure, intended f_o levels are given for the voiced sounds. For a visual estimate of *P*(i) and for calculated *F*-patterns, see the sounds online.
[C-06-01-F01]

Figure 1. Comparison of phonation-related pitch levels of natural whispered and voiced vowel sounds produced by single speakers: Illustration of the main finding. [C-06-01-F01]



1–1 [i] w-V-med 1002-A-m /h/
R193511

1–2 [i] 131-V-med 1002-A-m /l/
R132991

1–3 [i] w-V-med 1023-A-w /h/
R115937

1–4 [i] 220-V-med 1023-A-w /l/
R175024

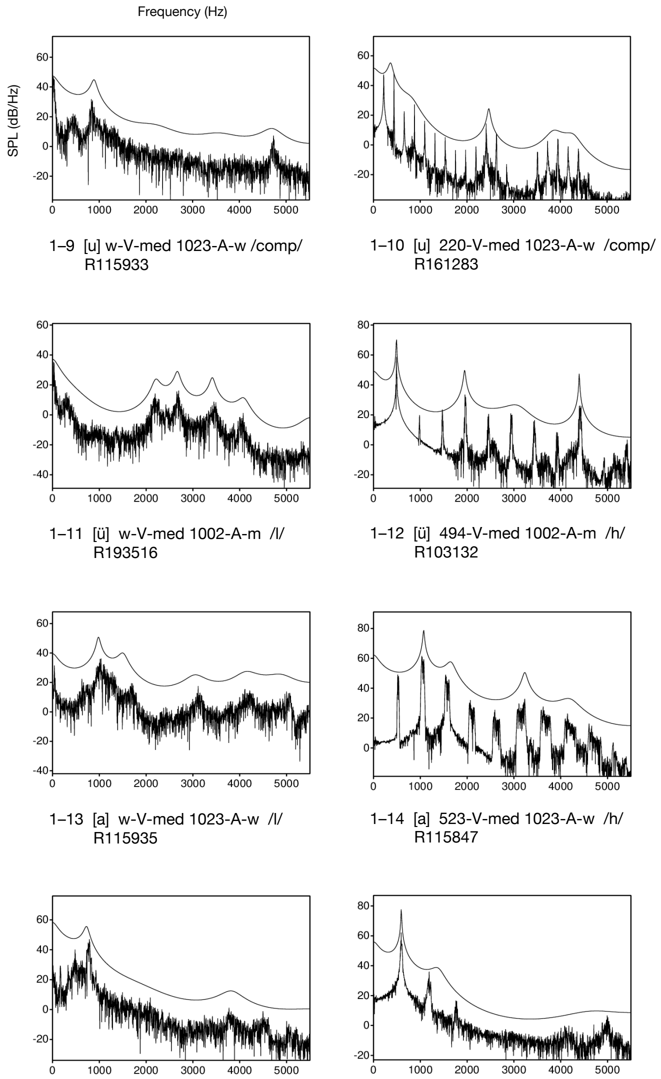
1–5 [i] w-V-med 1056-C-m /h/
R143496

1–6 [i] 262-V-med 1056-C-m /l/
R143007

1–7 [o] w-V-med 1056-C-m /comp/
R155294

1–8 [o] 262-V-med 1056-C-m /comp/
R142974

Figure 1 (continuation). [C-06-01-F01]



1–9 [u] w-V-med 1023-A-w /comp/
R115933

1–10 [u] 220-V-med 1023-A-w /comp/
R161283

1–11 [ü] w-V-med 1002-A-m /l/
R193516

1–12 [ü] 494-V-med 1002-A-m /h/
R103132

1–13 [a] w-V-med 1023-A-w /l/
R115935

1–14 [a] 523-V-med 1023-A-w /h/
R115847

1–15 [u] w-V-med 1056-C-m /l/
R155293

1–16 [u] 587-V-med 1056-C-m /h/
R155505

6.2 Pitch Recognition Comparing Synthesised Whispered-Like and Voiced-Like Vowel Sounds Related to Natural Whispered Utterances of Single Speakers

In a second experiment, pitch level differences for synthesised whispered-like and voiced-like vowel sounds were investigated based on an extract of the sound sample described in Chapter 5.3 (for the use of the term synthesis here, see also this chapter): For each of the eight long Standard German vowels and both speakers (man and woman), four synthesised replicas based on estimated *F*-patterns of natural whispered reference sounds were selected from that sample, related to four source characteristics in synthesis: Noise as a whispered-like source and a voiced-like source with f_o of 131–262–393 Hz (sounds of the man) or 165–262–440 Hz (sounds of the woman; note the selection of the lowest f_o level of 165 Hz investigated in Chapter 5.3 in order to reduce the low f_o level frequency difference of the voiced-like sounds of the woman and the man).

Pitch recognition of the compiled sounds was investigated in two speaker-blocked listening subtests according to the procedure described in the previous chapter, with the following experiment-specific adaptation: For each speaker and each vowel, the whispered-like sound was compared with each of the three voiced-like sounds at lower, medium and higher f_o levels, in AB and BA order.

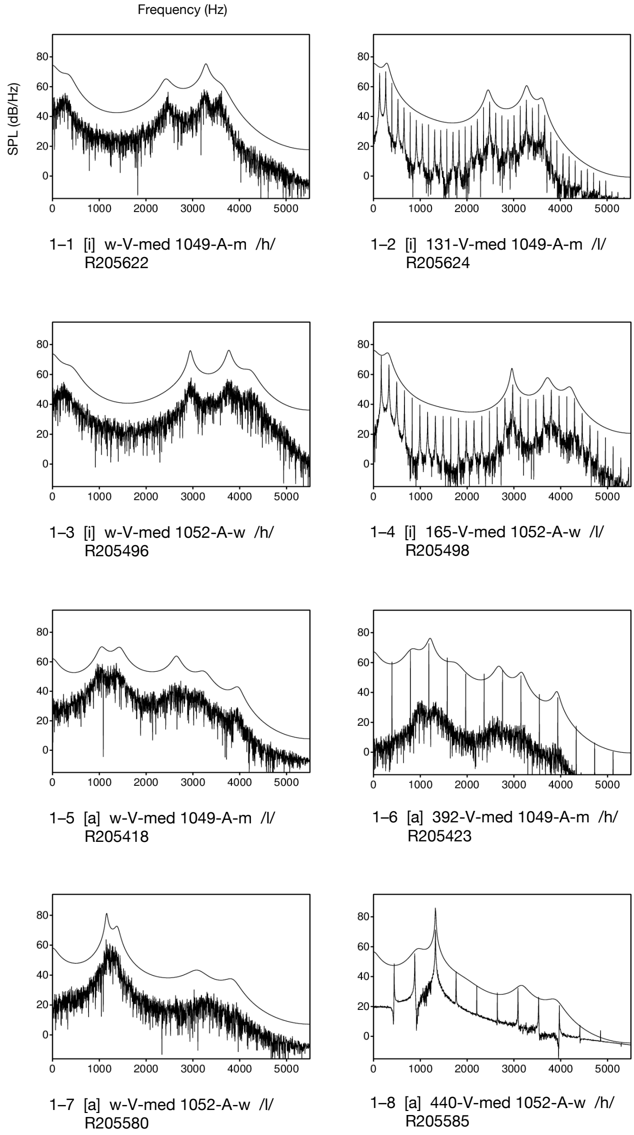
The main pitch recognition results obtained were as follows (again, f_o levels are given in terms of intended f_o, and pitch recognition results are given according to the labelling majority): For the sounds of the man, the pitch levels of the whispered-like sounds were recognised as higher than the levels of the voiced-like sounds synthesised at a low f_o of 131 Hz for six sound pairs and as higher or equal (no majority of labelling for a single option) for the remaining two pairs. Conversely, for all sound pairs where the voiced-like sounds were synthesised at a high f_o of 392 Hz, the pitch levels of the voiced-like sounds were recognised as higher than those of the whispered-like sounds. Finally, for the sound pairs where the voiced-like sounds were synthesised at a middle f_o of 262 Hz, the recognition results were somewhat mixed, with a tendency towards either equal pitch levels or somewhat balanced contradicting identifications for voiced-like and whispered-like sounds. Comparable results were also found for the sounds of the woman. (Note that low and high f_o levels of the voiced sounds of the woman were 165 Hz and 440 Hz.) Figure 1 illustrates this main finding. Besides, as was the case for the previous experiment, the pitch recognition results for the AB versus BA order were somewhat inconsistent between vowel qualities and listeners.

The results of this second experiment comparing synthesised sounds were in line with the results of the previous experiment comparing natural sounds: As a general tendency, the pitch levels for whispered-like sounds were perceived as higher in comparison to voiced-like sounds at f_o ≤ 165 Hz and lower in comparison to voiced-like sounds at f_o ≥ 393 Hz. Furthermore, when comparing whispered-like sounds with voiced-like sounds at f_o = 262 Hz, no general tendency of marked and consistent pitch level differences was found, and numerous cases of contradicting identifications occurred. (Notably, no such contradictions occurred for the comparisons with voiced-like replicas at f_o ≤ 165 Hz and ≥ 393 Hz.) Thus, our above estimate that the pitch level of the investigated natural whispered vowel sounds may be assessed as very often lying above c. 200 Hz and below c. 400 Hz was again supported on the bases of the investigated synthesised replicas.

For references, details of experimental design, method and results and their documentation (tables including sound links), see the Materials, Chapter M6.2.

Figure 1: Comparison of pitch levels of synthesised whispered-like and voiced-like vowel sounds, synthesis based on *F*-patterns of natural whispered reference utterances of single speakers: Illustration of the main finding. Extract of Chapter M6.2, Table 1 (see Series 1, 9, 6, 14, 4 and 12 in this table). Sounds 1–4 = two sound comparisons illustrating higher pitch levels for the whispered-like sound than for the voiced-like sound if f_o of the voiced-like sound in synthesis was ≤ 165 Hz. Sounds 5–8 = two sound comparisons illustrating lower pitch levels for the whispered-like sound than for the voiced-like sound if f_o of the voiced-like sound in synthesis was ≥ 393 Hz. Sounds 9–12 = two sound comparisons illustrating comparable pitch levels for the whispered-like and the voiced-like sounds if f_o of the voiced-like sound in synthesis was 262 Hz. In the figure, intended f_o levels are given for the voiced sounds. Vowel-related *F*-patterns of synthesis are given according to Chapter M5.3, Table 1 (see whispered sounds). For the notation of recognised pitch level in slashes (labelling majority), see the paragraph on figures and figure legends in the Introduction. For a visual estimate of *P*(i) and for calculated *F*-patterns, see the sounds online.
[C-06-02-F01] ↗

Figure 1. Comparison of phonation-related pitch levels of synthesised whispered-like and voiced-like vowel sounds, synthesis based on F-patterns of natural whispered reference utterances of single speakers: Illustration of the main finding. [C-06-02-F01]



1–1 [i] w-V-med 1049-A-m /h/
R205622

1–2 [i] 131-V-med 1049-A-m /l/
R205624

1–3 [i] w-V-med 1052-A-w /h/
R205496

1–4 [i] 165-V-med 1052-A-w /l/
R205498

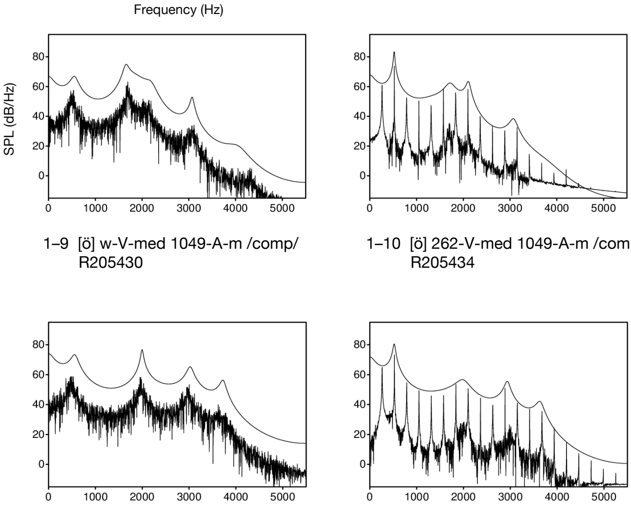
1–5 [a] w-V-med 1049-A-m /l/
R205418

1–6 [a] 392-V-med 1049-A-m /h/
R205423

1–7 [a] w-V-med 1052-A-w /l/
R205580

1–8 [a] 440-V-med 1052-A-w /h/
R205585

Figure 1 (continuation). [C-06-02-F01]



1–9 [ö] w-V-med 1049-A-m /comp/ R205430

1–10 [ö] 262-V-med 1049-A-m /comp/ R205434

1–11 [ö] w-V-med 1052-A-w /comp/ R205502

1–12 [ö] 262-V-med 1052-A-w /comp/ R205506

6.3 Pitch Recognition Comparing Either Natural Whispered or Synthesised Whispered-Like Vowel Sounds Related to Utterances of Speakers Different in Age and Gender

In the previous two experiments, the pitch of whispered and whispered-like vowel sounds was investigated by means of a comparison with voiced and voiced-like vowel sounds, all comparisons being related to utterances of single speakers. In two further experiments, using the same two sound samples of the previous chapters, pitch level differences were investigated by means of a comparison of whispered or whispered-like sounds (either natural or synthesised replicas) produced by different speakers of different ages or gender. The initial goal was to investigate whether pitch assessment is subject to age- and gender-related differences.

The first experiment related to the sample of all 24 natural whispered sounds of the eight long Standard German vowels produced by the man, the woman and the child described in Chapter 6.1. For each vowel, each of the three related sounds was compared with one of its opposing sounds, in AB and BA order. The recognition of pitch level differences between the compared sounds was investigated in a listening test according to the procedure also described in Chapter 6.1: Each test item contained two whispered sounds of the same vowel produced by two different speakers (separated by an approximately 0.5 sec. pause), and the listeners were asked to identify the pitch level difference between the first and the second sound as falling, flat or rising, referring to dominant or prominent pitches.

According to the recognition results (labelling majority), as a general tendency, the pitch level of the sounds of the man was identified as either lower than or equal to the pitch of the sounds of the women and the child, and the same held true for the comparison of the sounds of the women and the child. Figure 1 illustrates this finding. However, the results were not uniform among all vowel qualities, and labelling consistency was somewhat limited (see Chapter M6.3).

The second experiment related to the sample of the 16 synthesised whispered-like replicas of the sounds of the eight long Standard German vowels produced by the man and the woman described in Chapter 6.2. For each vowel, the two sounds of the two speakers were compared with each other, in AB and BA order. The recognition of pitch level differences between the sounds compared was investigated by means of a listening test according to the above procedure.

For the synthesised sounds of front vowels and of /a/ related to the natural reference sounds produced by the man, according to the recognition results, the pitch level was unanimously identified as lower than the level of the sounds of the woman. In contrast, for the synthesised replicas related to the natural reference sounds of /o/ produced by the man, the pitch level was either identified as lower than or equal to the level of the sounds of the woman. For the synthesised replicas of /u/, no inter-speaker differences were identified for the pitch levels. Figure 2 illustrates this second finding.

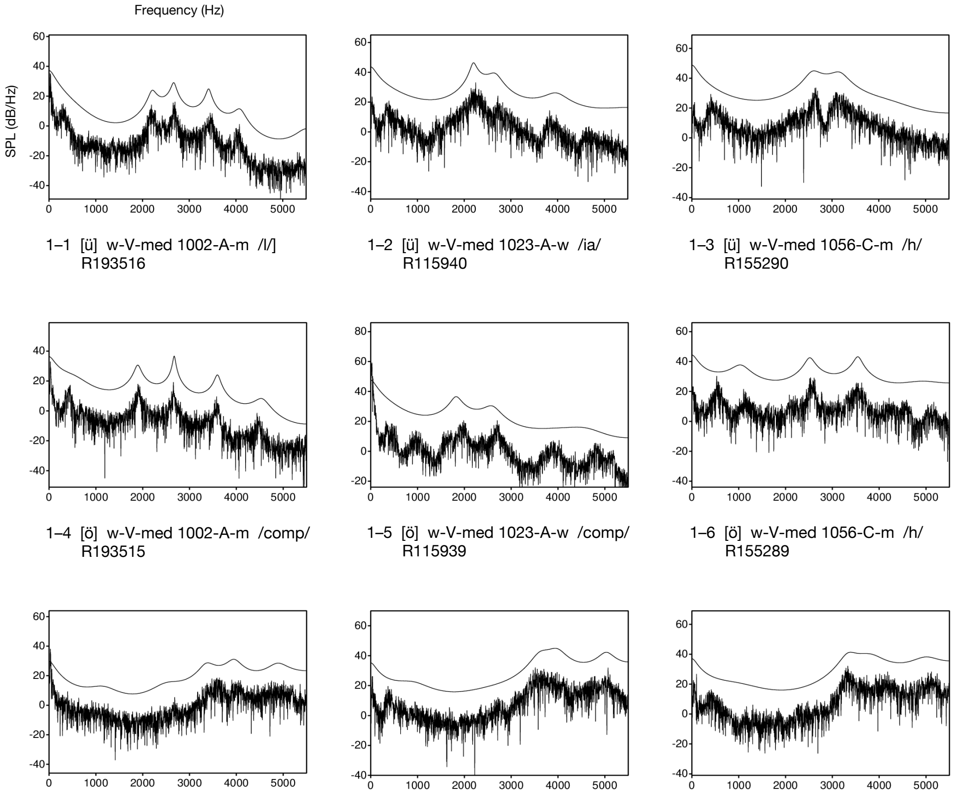
In sum, the recognition results for natural whispered sounds indicated lower or equal pitch levels for the sounds of the man compared to the sounds of the woman and the child and lower or equal levels for the sounds of the woman compared to the sounds of the child. Hence, for the investigated sound sample, a somewhat limited and inconsistent tendency towards age- and gender-related pitch level recognition was found. Correspondingly, labelling consistency did not surpass 74%. The results of the listening test for synthesised sounds revealed much more pronounced gender-related differences for the adults, with high labelling consistency, although these differences were limited to sounds of front vowels and of /a/. In these terms, for the sounds investigated, the tendency towards age- and gender-related pitch recognition proved to be dependent on sound types and vowel qualities. Besides, the finding that pitch differences were easier to spot for synthesised whispered sounds as opposed to natural whispered sounds was unexpected and needs to be addressed in future research.

For details of experimental design, method and results, an extended discussion and documentation of results (tables including sound links), see the Materials, Chapter M6.3.

Figure 1. Recognised pitch level differences for the comparison of natural whispered vowel sounds produced by a man, a woman and a child: Illustration of the main finding of experiment 1. Extract of Chapter M6.3, Table 1 (see Series 2, 4 and 1 in this table). Sounds 1–3 = by tendency lower–intermediate–higher pitch levels for the sounds of the man, the woman and the child. Sounds 4–6 = by tendency comparable pitch levels for the sounds of the adults and higher level for the sound of the child. Sounds 7–9 = by tendency comparable levels for the sounds of all three speakers. For the notation of recognised pitch level in slashes (labelling majority), see the paragraph on figures and figure legends in the Introduction.
[C-06-03-F01] ↗

Figure 2. Recognised pitch level differences for the comparison of synthesised whispered-like vowel sounds related to natural whispered reference sounds produced by a man and a woman: Illustration of the main finding of experiment 2. Extract of Chapter M6.3, Table 2 (see Series 1, 3, 5, 7 and 8 in this table). Sounds 1–6 = sound pairs of /i, e, ɛ/, with lower pitch levels for the sound of the man compared with the sound of the woman. Sounds 7 and 8 = sound pair of /o/ with lower or comparable (l–comp) pitch levels for the sound of the man compared with the sound of the woman (no labelling majority in the listening test). Sounds 9 and 10 = sound pair of /u/ with comparable pitch levels for the sounds of both speakers. For a visual estimate of *P*(i) and for calculated *F*-patterns, see the sounds online.
[C-06-03-F02] ↗

Figure 1. Recognised pitch level differences for the comparison of natural whispered vowel sounds produced by a man, a woman and a child. Illustration of the main finding of experiment 1. [C-06-03-F01]



1–1 [ü] w-V-med 1002-A-m /l/]
R193516

1–2 [ü] w-V-med 1023-A-w /ia/
R115940

1–3 [ü] w-V-med 1056-C-m /h/
R155290

1–4 [ö] w-V-med 1002-A-m /comp/
R193515

1–5 [ö] w-V-med 1023-A-w /comp/
R115939

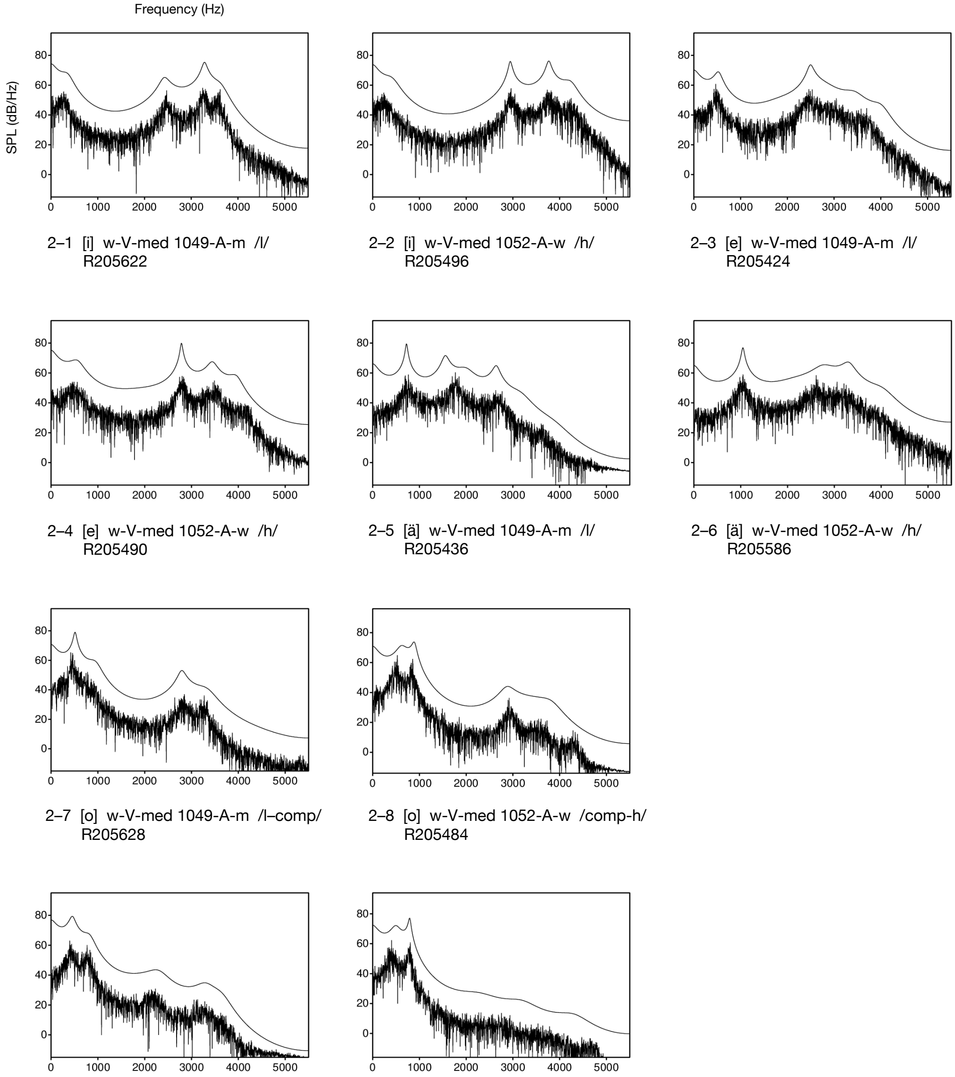
1–6 [ö] w-V-med 1056-C-m /h/
R155289

1–7 [i] w-V-med 1002-A-m /comp/
R193511

1–8 [i] w-V-med 1023-A-w /comp/
R115937

1–9 [i] w-V-med 1056-C-m /comp/
R143496

Figure 2. Recognised pitch level differences for the comparison of synthesised whispered-like vowel sounds related to natural whispered reference sounds produced by a man and a woman. Illustration of the main finding of experiment 2. [C-06-03-F02]



2–1 [i] w-V-med 1049-A-m /l/
R205622

2–2 [i] w-V-med 1052-A-w /h/
R205496

2–3 [e] w-V-med 1049-A-m /l/
R205424

2–4 [e] w-V-med 1052-A-w /h/
R205490

2–5 [ä] w-V-med 1049-A-m /l/
R205436

2–6 [ä] w-V-med 1052-A-w /h/
R205586

2–7 [o] w-V-med 1049-A-m /l–comp/
R205628

2–8 [o] w-V-med 1052-A-w /comp-h/
R205484

2–9 [u] w-V-med 1049-A-m /comp/
R205412

2–10 [u] w-V-med 1052-A-w /comp/
R205478

6.4 Natural Vowel Sounds With a Suppressed Fundamental

The pitch of a periodic sound can be perceived independently of whether or not the first harmonic, *H*1, commonly termed the fundamental, is present in the spectrum (“missing fundamental” phenomenon). For example, a sound with a series of harmonics as integer multiples of 200 Hz remains at a 200 Hz pitch level even if the first harmonic is (or also some of the lower harmonics are) removed. Similarly, speech remains intelligible even if frequencies below c. 300 Hz are filtered, including the original pitch contour (consider e.g. fixed telephone line transmission with a band-pass filter of c. 300–3400 Hz).

The previous chapters demonstrated vowel sounds that have a pitch but lack measurable f_o. Vowel sounds with a “missing fundamental” represent a second phenomenon of sounds for which f_o-related acoustic characteristics, vowel quality and pitch are at issue. Therefore, with the aim of documenting the “missing fundamental” phenomenon in the context of the debate about the role of f_o, *H*1, HCF, sound periodicity and pitch for vowel recognition, a corresponding experiment was conducted: Based on the Zurich Corpus, for each of the eight long Standard German vowels and each of the speaker groups of men, women and children, three sounds produced by three speakers in nonstyle mode, V context and with a medium vocal effort at intended f_o of 131 Hz (men), 220 Hz (women) and 262 Hz (children) were selected. According to the standard listening test results when creating the corpus, all sounds were fully recognised (100% vowel recognition rate matching vowel intention). In these terms, a sample of 72 natural reference sounds produced by different speakers was investigated. The sounds of this sample were HP filtered two times, firstly suppressing *H*1 and secondly suppressing *H*1–*H*2, and vowel recognition was tested for all HP-filtered sounds according to the standard procedure of the corpus and involving the five standard listeners.

According to the labelling majority of the listening test, with one exception, all sounds with suppressed *H*1 were recognised as either matching vowel intention of the speakers or an adjacent vowel quality or as a vowel boundary or area of intended and adjacent qualities. The same held true for c. 80% of the sounds with suppressed *H*1–*H*2. For the remaining sounds, vowel confusion involving more than one adjacent vowel occurred.

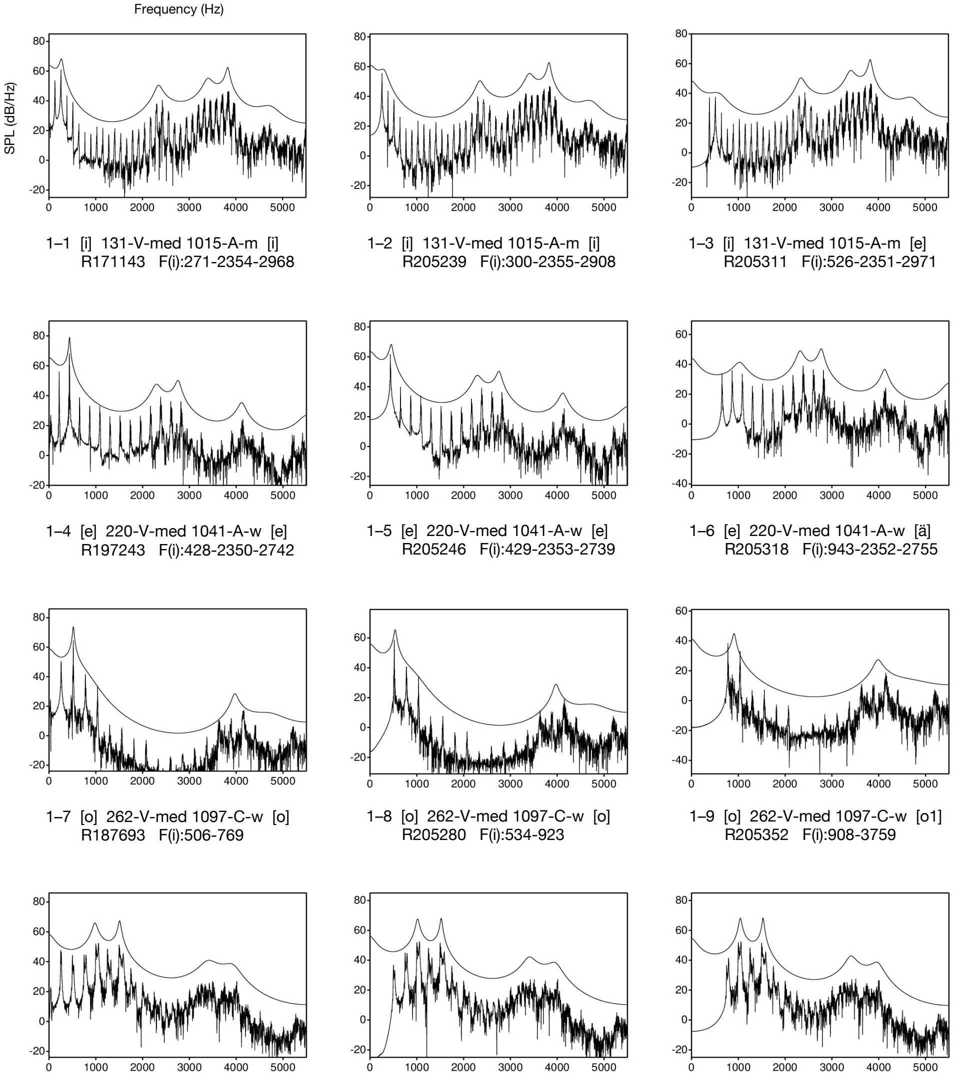
It is noteworthy that vowel quality shifts or vowel confusions triggered by suppressed *H*1 occurred almost exclusively for sounds of close vowels produced by women and children, and shifts or confusions triggered by suppressed *H*1–*H*2 occurred only for sounds of close vowels produced by men and for sounds of close and close-mid vowels produced by women and children. Further, the shift direction from the vowel quality intended by the speaker to an adjacent or non-adjacent vowel quality was found to generally be close–open. Finally, with few exceptions, measured f_o for the sounds with suppressed *H*1 or *H*1–*H*2 corresponded to measured f_o of the unfiltered reference sounds.

Pitch recognition was not investigated in this experiment. However, it is assumed here that no pitch variation resulted from the HP filtering applied. (For verification, see Chapter M6.4, Table 1, sound links.) Accordingly, if vowel quality shifts occurred, they concerned either (initial) shifts in a close–open direction or vowel confusions, in contrast to the shifts found for increasing f_o in sound synthesis, keeping the spectral envelope unchanged, which resulted in shifts in an open–close direction. Thus, if suppression of *H*1 or *H*1–*H*2 affected vowel quality recognition, these shifts are assumed here as unrelated to pitch. At the same time and in these terms, vowel recognition does not rely on the frequency level of *H*1, and sounds with suppressed *H*1 or *H*1–*H*2 represent cases for which f_o and the HCF cannot be equated with *H*1. Figure 1 shows corresponding sound examples.

For references, extended background information, details of experimental design, method and results, an extended discussion and documentation of results (tables including sound links), see the Materials, Chapter M6.4.

Figure 1. Vowel recognition of natural vowel sounds with suppressed *H*1 or *H*1–*H*2: Sound examples. Four triplets of vowel sounds are shown, each triplet consisting of an unfiltered reference sound, the related sound with suppressed H1 and the related sound with suppressed *H*1–*H*2. Extract of Chapter M6.4, Table 1 (see Series 1, 12, 22 and 24 in this table). The first three triplets illustrate cases of maintained vowel qualities for *H*1 suppression and close–open vowel quality shifts for *H*1–*H*2 suppression (/i/ to /e/ for the first triplet, /e/ to /ɛ/ for the second triplet, and /o/ to /ɔ/ for the third triplet). The fourth triplet illustrates a case of maintained vowel quality for both *H*1 and *H*1–*H*2 suppression. For all these sounds, the pitch level is unaffected by the suppression of the first two harmonics, which can be examined by listening to the sounds.
[C-06-04-F01] ↗

Figure 1. Vowel recognition of natural vowel sounds with suppressed H1 or H1–H2: Sound examples. [C-06-04-F01]



1–1 [i] 131-V-med 1015-A-m [i]
R171143 F(i):271-2354-2968

1–2 [i] 131-V-med 1015-A-m [i]
R205239 F(i):300-2355-2908

1–3 [i] 131-V-med 1015-A-m [e]
R205311 F(i):526-2351-2971

1–4 [e] 220-V-med 1041-A-w [e]
R197243 F(i):428-2350-2742

1–5 [e] 220-V-med 1041-A-w [e]
R205246 F(i):429-2353-2739

1–6 [e] 220-V-med 1041-A-w [ä]
R205318 F(i):943-2352-2755

1–7 [o] 262-V-med 1097-C-w [o]
R187693 F(i):506-769

1–8 [o] 262-V-med 1097-C-w [o]
R205280 F(i):534-923

1–9 [o] 262-V-med 1097-C-w [o1]
R205352 F(i):908-3759

1–10 [a] 262-V-med 1098-C-w [a]
R187747 F(i):999-1515

1–11 [a] 262-V-med 1098-C-w [a]
R205284 F(i):1026-1531

1–12 [a] 262-V-med 1098-C-w [a]
R205356 F(i):1050-1541

6.5 Sinewave Vowel Sounds I – Replicas Related to Statistical Formant Patterns

To some degree, sinewave speech – synthesised replicas of utterances based on time-varying sinusoidal patterns following the changing formant centre frequencies of the natural sounds – is intelligible. In this context, of particular interest are single sinewave vowel sounds that are synthesised based on a small number of sinusoids related to statistical average *F*-patterns of natural sounds produced in V context or the context of minimal pairs: Sounds of this type can be synthesised with sinusoid frequency configurations that are directly related to these *F*-patterns but lacking both a fundamental *H*1 as well as HCF comparable to a harmonic spectrum of a natural vowel sound. Thus, the question of the relation between f_o measure, pitch and vowel recognition is posed from a broader perspective.

In the literature, the recognition of sinewave vowel sounds replicating *F*-patterns of natural sounds produced in citation-form words is reported as impaired when compared with the recognition of natural vowel sounds. However, and most importantly, vowel confusion for synthesised sounds is also indicated to relate to vowel openness, with markedly better vowel recognition for sounds of close vowels than for open-mid and open vowels. When considering this recognition difference in relation to vowel openness, attention should be given to the fact that S_1, the frequency level of the first sinusoid representing F_1 of the natural sounds, is substantially lower for *S*-patterns related to sounds of close vowels than for *S*-patterns related to the sounds of all other vowel qualities, with the highest S_1 for open-mid and open vowels. Since some studies of pitch recognition in sinusoidal sentences indicated that pitch approximately relates to the frequency of the first sinusoid, and assuming that vowel recognition in its turn relates to pitch, vowel quality shifts or confusions can be expected to occur with increasing differences between the pitch level of the natural reference sounds and the pitch level of the related sinewave replicas, at least for some of the vowel qualities. This would explain the indication of the recognition of sinewave vowel sounds being related to vowel openness.

With regard to a corresponding experiment investigating vowel and pitch recognition based on sinewave vowel sounds, we have further taken into account a major aspect of the formant pattern and spectral shape ambiguity phenomenon: Given single spectral envelopes of sounds of close-mid vowels produced at f_o of 200–250 Hz, a one-octave increase in f_o in synthesis, keeping the spectral envelope unchanged, was shown to result in pronounced close-mid–close vowel recognition shifts (see Chapter 3), in contrast to sounds of these vowels produced at f_o of 100–125 Hz and a one-octave increase in f_o in synthesis. We, therefore, assumed that statistical average *F*-patterns of vowel sounds produced by women might represent a promising outset for the exploration of the matter: Statistical average *F*-patterns of women are generally reported for sounds produced at f_o of 200–250 Hz, and a one-octave f_o variation appertains to the everyday speech of women.

In view of this, vowel and pitch recognition of sinusoid S_1–S_2–S_3 sounds replicating statistical F_1–F_2–F_3 patterns of the eight Standard German vowels for women, as reported by Pätzold and Simpson (1997; see Chapter 3.1), were investigated according to the following main idea and assumption: If vowel and pitch recognition interrelate, and if pitch recognition of sounds in three-sinusoid synthesis relates to S_1, then lower pitch levels and less vowel confusion can be expected for sounds replicating the *F*-patterns of close vowels, above all when compared with the levels of sounds replicating the *F*-patterns of close-mid vowels. However, because of the nonuniform relation of the vowel spectrum to f_o, no assumption was made for sounds of the open-mid and open vowels. Hence, sinewave vowel sounds were produced in terms of synthesised static sounds based on S_1–S_2–S_3 configurations related to the above F_1–F_2–F_3 patterns, with the sinewave levels set to 100–80–80 dB for the sounds of front vowels and to 100–90–70 dB for the sounds of back vowels. The sound duration was 1 sec. including a 0.05 sec. fade in/fade out. As a result, a sample of eight sinewave vowel sounds was created.

Vowel recognition of the synthesised sounds was then investigated in a listening test according to the standard procedure of the corpus involving the five standard listeners, with some experiment-specific adaptations (for details, see Chapter M6.5). Separately, pitch recognition of the synthesised sounds was investigated in two subtests involving the same listeners. In subtest 1, the sounds related to the *F*-patterns of close-mid and close and of close-mid, open-mid and open vowels were compared as follows: A close-mid versus a close vowel sound, a close-mid versus an open-mid vowel sound and a close-mid versus an open vowel sound, in AB and BA order. Thus, each test item consisted of two vowel sounds (separated by a 0.5 sec. pause). The listeners were asked to identify whether the pitch level of the second sound when compared with the pitch level of the first sound was falling, flat or rising, referring to dominant or prominent levels. In subtest 2, pitch recognition was investigated by comparing all sinewave vowel sounds with two single sinusoids of 349 Hz and 440 Hz separately: One test item consisted of either the 349 Hz or the 440 Hz sinusoid followed by a sinewave vowel sound (separated by a 0.5 sec. pause). Listeners were again asked to identify the pitch level difference.

According to the vowel recognition results (labelling majority), all three sounds with *S*-patterns related to *F*-patterns of close vowels were recognised according to vowel intention, with a recognition rate of ≥ 80%. On the contrary, all three sounds related to *F*-patterns of close-mid vowels were confused and were recognised as close vowels, again with a recognition rate of ≥ 80%. The sound related to the *F*-pattern of /ɛ/ was confused with /ø/, and the sound related to the *F*-pattern of /a/ was mostly identified as /a/ or /ɔ/.

Uniform pitch recognition results were obtained across listeners, tests and order of sound presentation: The pitch level of all sounds related to the *F*-patterns of close vowels was recognised as being lower than the pitch level of all sounds related to the *F*-patterns of close-mid vowels (pitch recognition subtest 1), and it was identified as lower than or equal to the 349 Hz sinusoid (pitch recognition subtest 2). The pitch level of all sounds related to the *F*-patterns of the close-mid vowels was recognised as being lower than the pitch level of all sounds related to the *F*-patterns of the open-mid and open vowels, and it was identified as being above the 349 Hz sinusoid and lower than or equal to the 440 Hz sinusoid. Finally, the pitch level of all sounds related to the *F*-patterns of the open-mid and the open vowels was recognised as being higher than the 440 Hz sinusoid.

In sum, the vowel qualities of the sounds related to *F*-patterns of close vowels were matched successfully to the intended quality (majority of labelling), with the pitch of the sounds being recognised at a lower level than the level of the sounds of the other vowels. In contrast, the vowel qualities of the sounds related to *F*-patterns of close-mid and open-mid vowels were confused, and the pitch of these sounds was recognised at middle or higher levels. Besides, the fact that the vowel quality of sounds of /a/ was recognised in the /a-ɔ/ range accorded with the observation of a weak or absent relation of the *F*-patterns or spectral envelopes to f_o for sounds of that vowel, as indicated in the earlier experiments presented in this treatise.

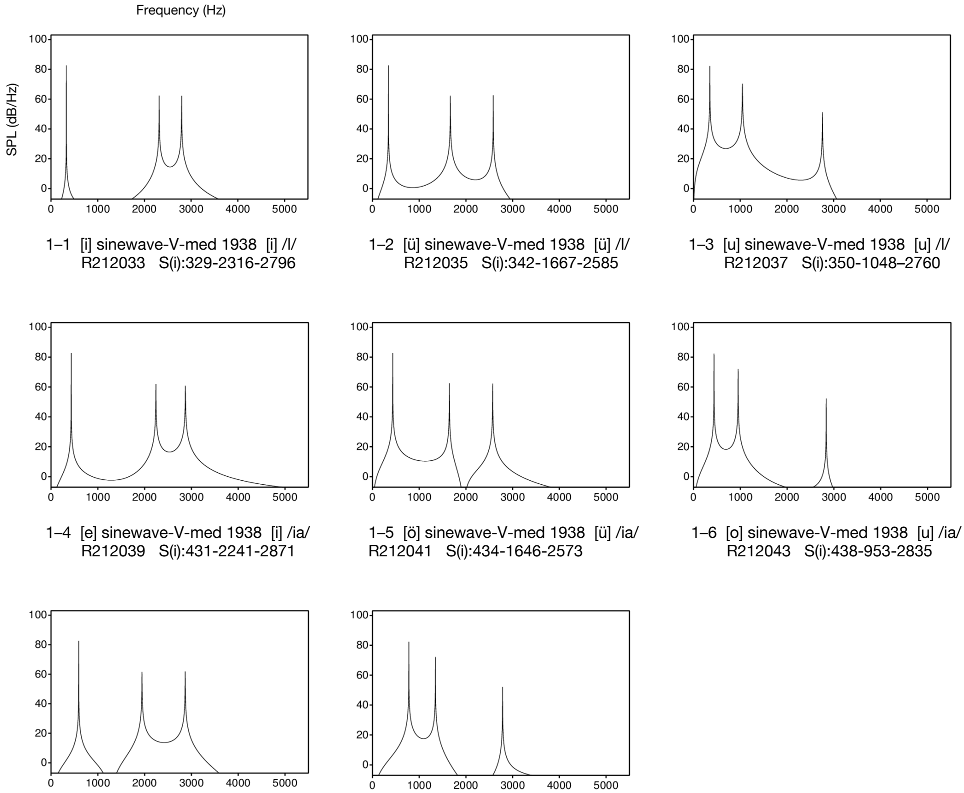
In these terms, given a frequency range of f_o of c. 200–250 Hz of the natural reference sounds for which the *F*-patterns were replicated in three-sinusoid synthesis, (i) sinewave vowel sounds could be identified as having specific vowel qualities and (ii) as having an identifiable pitch level, although the HCF comparable to a harmonic sound spectrum was lacking (in the present experiment the pitch level being comparable to S_1), whereby (iii) the pitch interacted with vowel recognition. Thus, spectral peak frequencies and estimated *F*-patterns *per se* were again indicated as not representing vowel qualities *in general* but as being related to pitch in the perceptual process of vowel recognition. Figure 1 illustrates this main indication, presenting all eight sounds investigated.

Remarkably enough, the confusions above all for the sounds of close-mid vowels found in the present experiment paralleled vowel quality shifts for natural sounds of these vowels as demonstrated in the chapters on formant pattern and spectral shape ambiguity for an increase of f_o (and pitch) of one octave from c. 200–250 Hz to c. 400–500 Hz: Accordingly, in the present experiment, the sounds with *S*-patterns related to the *F*-patterns of /e/, /ø/ and /o/ were mostly recognised as /i/, /y/ and /u/, respectively.

For references, extended background information, details of experimental design, method and results, an extended discussion and documentation of results (tables including sound links), see the Materials, Chapter M6.5.

Figure 1. Vowel and pitch recognition of synthesised three-sinusoid sounds based on statistical *F*-patterns of sounds of the long Standard German vowels produced by women: Illustration of the main finding. Extract of Chapter M6.5, Table 1. Eight synthesised sounds corresponding to statistical *F*-patterns of /i, y, e, ø, ε, a, o, u/ are shown (for reference, see text). Sounds 1–3 = synthesised sounds related to *F*-patterns of close vowels, with recognised close vowel qualities and lower pitch levels. Sounds 4–6 = synthesised sounds related to *F*-patterns of close-mid vowels, with recognised close vowel qualities and intermediate pitch levels. Sounds 7 and 8 = synthesised sounds related to *F*-patterns of /ε/ and /a/, with recognised vowel qualities of /ø/ and /a–ɔ/ and higher pitch levels. For each sound, the *S*-pattern used for synthesis is given. For the notation of recognised pitch level in slashes (labelling majority), see the paragraph on figures and figure legends in the Introduction.
[C-06-05-F01] ↗

Figure 1. Vowel and pitch recognition of synthesised three-sinusoid sounds based on statistical F-patterns of sounds of the long Standard German vowels produced by women: Illustration of the main finding. [C-06-05-F01]



1–1 [i] sinewave-V-med 1938 [i] /l/
R212033 S(i):329-2316-2796

1–2 [ü] sinewave-V-med 1938 [ü] /l/
R212035 S(i):342-1667-2585

1–3 [u] sinewave-V-med 1938 [u] /l/
R212037 S(i):350-1048–2760

1–4 [e] sinewave-V-med 1938 [i] /ia/
R212039 S(i):431-2241-2871

1–5 [ö] sinewave-V-med 1938 [ü] /ia/
R212041 S(i):434-1646-2573

1–6 [o] sinewave-V-med 1938 [u] /ia/
R212043 S(i):438-953-2835

1–7 [ä] sinewave-V-med 1938 [ö] /h/
R212045 S(i):592-1944-2867

1–8 [a] sinewave-V-med 1938 [a–o1] /h/
R212047 S(i):779-1347-2785

6.6 Sinewave Vowel Sounds II – Replicas Related to Estimated Formant Patterns of Single Natural Vowel Sounds

Having obtained the results of the first sinewave vowel sound experiment conducted, the experiment was replicated based on a sample of single natural sounds produced by women at an intended f_o of 220 Hz and the respective estimated *F*-patterns of these sounds. The aim of this second sinewave experiment was to create a basis for and documentation of vowel quality recognition in sinewave synthesis that allows for a direct relation between single natural reference sounds and synthesised replicas. This direct relation strengthens the reliability of the relation between *F*-patterns, *S*-patterns and vowel and pitch recognition for the sounds investigated.

Based on sounds of the Zurich Corpus produced by women in non-style mode with a medium vocal effort and in V context, for each of the eight long Standard German vowels, a sound produced at an intended f_o of 220 Hz was selected by the author (for further selection criteria, see Chapter M6.6). The vowel recognition rate of all sounds was 100% (matching vowel intention) according to the standard listening test conducted when creating the corpus.

Based on S_1–S_2–S_3 patterns related to the estimated F_1–F_2–F_3 patterns of the eight natural reference vowel sounds, static three-sinusoid sounds were synthesised, with sinewave levels, sound duration and on- and offsets according to the previous experiment of Chapter 6.5. As a result, a sample of eight sinewave vowel sounds was created. Vowel and pitch recognition of the synthesised sounds were then investigated in listening subtests according to the procedure also described in the previous chapter, with the second pitch recognition subtest related to sinusoid frequencies of 330 Hz (above the F_1/S_1 of the investigated sounds of the close vowels) and 440 Hz (below the F_1/S_1 of the investigated sounds of the open-mid and open vowels, and also corresponding to the F_1/S_1 of the investigated sounds of the close-mid vowels).

In general, the results obtained for vowel and pitch recognition strongly supported the findings of the previous experiments, with only marginal differences. According to the vowel recognition results (labelling majority), all three sounds with *S*-patterns related to *F*-patterns of close vowels were recognised according to vowel intention, with a recognition rate of 100%. On the contrary, all three sounds related to *F*-patterns of close-mid vowels were confused and were identified as close vowels with a recognition rate of ≥ 70%. The sound related to the

F-pattern of /ɛ/ was mostly confused with /e/, and the sound related to the *F*-pattern of /a/ was mostly identified as /a/ or /ɔ/. Concerning pitch recognition, uniform results were again obtained across listeners, tests and sound presentation order, with recognised pitch levels below 330 Hz for all sounds related to *F*-patterns of close vowels, above 330 Hz and lower than or equal to 440 Hz for all sounds of close-mid vowels, and higher than 440 Hz for the two sounds of /ɜ/ and /a/. Notably, again, pitch recognition accorded to the frequency ranges of S_1.

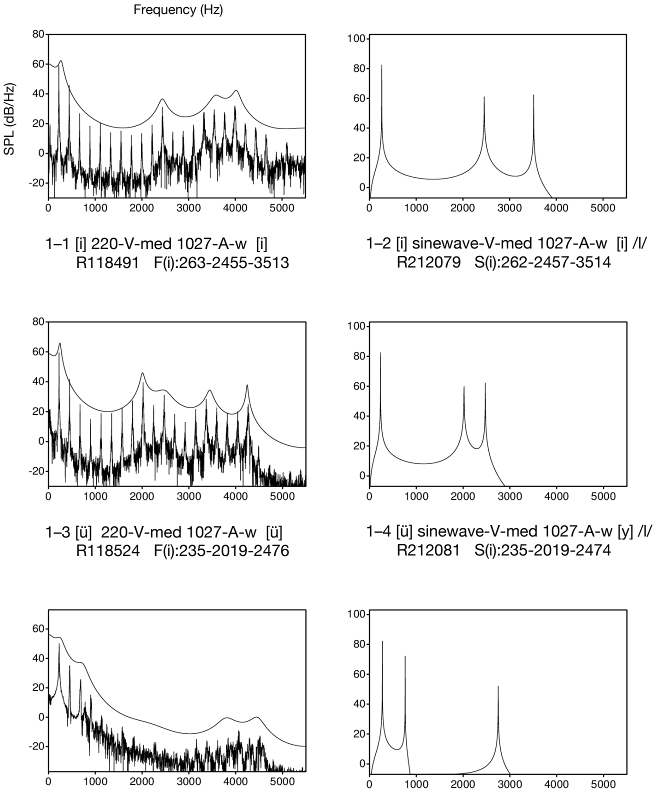
In sum, apart from minor differences, the vowel and pitch recognition of the synthesised sounds of the present experiment with *S*-patterns relating to *F*-patterns of single natural sounds produced by women corresponded to the recognition of synthesised sounds with *S*-patterns relating to statistical average *F*-patterns. Above all, the results again strongly supported the two notions of recognsied vowel quality being related to pitch and this relation being nonuniform: Firstly, vowel confusion occurred for sounds of close-mid and open-mid vowels with higher pitches than the sounds of close vowels, which were recognised according to vowel intention. Figures 1 and 2 illustrate this main finding for the natural reference sounds of close and close-mid vowels and their sinewave replicas. Secondly, the effect of high pitch levels on vowel recognition was somewhat limited for the sounds of intended /a/, the most open vowel quality (note that the difference between /a/ and /ɔ/ is not a difference of two long vowel qualities in Standard German). As mentioned, this accorded with the observation of a weak or absent relation of the *F*-patterns or spectral envelopes to f_o for sounds of that vowel (see Chapters 2 and 3).

For details of the method and results, an extended discussion and documentation of results (tables including sound links), see the Materials, Chapter M6.6. For some occurring (marginal) differences between estimated *F*- and *S*-patterns when conducting the experiment and *F*-patterns as given in the Zurich Corpus, see the Introduction (differences in the figures are < 5 Hz).

Figure 1. Vowel and pitch recognition of synthesised three-sinusoid sounds based on estimated *F*-patterns of single natural reference sounds of Standard German close vowels produced by women: Illustration of maintained vowel quality and lower pitch levels. Extract of Chapter M6.6, Table 1 (see Series 1 in this table). For each of the three close vowels /i/, /y/ and /u/, a natural reference sound and the respective synthesised three-sinusoid sound based on the *F*-pattern of the reference sound are shown. All synthesised sounds were recognised as close vowels, and their pitch level was assessed as lying in the lower frequency range below 330 Hz. For the notation of recognised pitch level in slashes (labelling majority), see the paragraph on figures and figure legends in the Introduction.
[C-06-06-F01] ↗

Figure 2. Vowel and pitch recognition of synthesised three-sinusoid sounds based on estimated *F*-patterns of single natural reference sounds of Standard German close-mid vowels produced by women: Illustration of vowel confusion and middle pitch levels. Extract of Chapter M6.6, Table 1 (see Series 2 in this table). For each of the three close-mid vowels /e/, /ø/ and /o/, a natural reference sound and the respective synthesised three-sinusoid sound based on the *F*-pattern of the reference sound are shown. All synthesised sounds were recognised as close vowels, and their pitch level was assessed as lying in an intermediate frequency range of 330–440 Hz.
[C-06-06-F02] ↗

Figure 1. Vowel and pitch recognition of synthesised three-sinusoid sounds based on estimated F-patterns of single natural reference sounds of Standard German close vowels produced by women: Illustration of maintained vowel quality and lower pitch levels. [C-06-06-F01]



1–1 [i] 220-V-med 1027-A-w [i]
R118491 F(i):263-2455-3513

1–2 [i] sinewave-V-med 1027-A-w [i] /I/
R212079 S(i):262-2457-3514

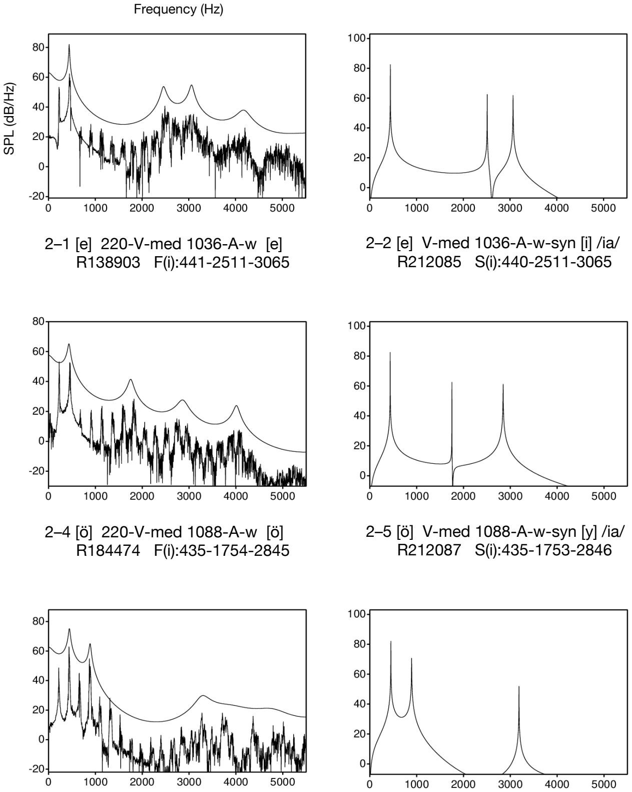
1–3 [ü] 220-V-med 1027-A-w [ü]
R118524 F(i):235-2019-2476

1–4 [ü] sinewave-V-med 1027-A-w [y] /I/
R212081 S(i):235-2019-2474

1–5 [u] 220-V-med 1027-A-w [u]
R118445 F(i):273-760

1–6 [u] sinewave-V-med 1027-A-w [u] /I/
R212083 S(i):274-762-2750

Figure 2. Vowel and pitch recognition of synthesised three-sinusoid sounds based on estimated F-patterns of single natural reference sounds of Standard German close-mid vowels produced by women: Illustration of vowel confusion and middle pitch levels. [C-06-06-F02]



2–1 [e] 220-V-med 1036-A-w [e]
R138903 F(i):441-2511-3065

2–2 [e] V-med 1036-A-w-syn [i] /ia/
R212085 S(i):440-2511-3065

2–4 [ö] 220-V-med 1088-A-w [ö]
R184474 F(i):435-1754-2845

2–5 [ö] V-med 1088-A-w-syn [y] /ia/
R212087 S(i):435-1753-2846

2–7 [o] 220-V-med 1036-A-w [o]
R170730 F(i):449-893

2–8 [o] V-med 1036-A-w-syn [u] /ia/
R212089 S(i):449-893-3185

6.7 Harmonic Synthesis I – Changing Vowel Quality by Changing Either the Lower Spectral Energy Maximum or the Highest Common Factor of Sinusoid Configurations

As shown, natural voiced vowel sounds can be recognised with suppressed first or suppressed first and second harmonic(s). For these cases, the HCF of the HP-filtered spectrum is not affected and can be assumed to generally represent the sound periodicity to which perception and recognition relate. However, in contrast, synthesised vowel sounds with only a few partials in their spectrum can be recognised, too, independently of whether or not the partials are in a harmonic relation, and pitch recognition tests indicated that, for these sounds, the frequency of the lowest partial often represents the sound periodicity which perception and recognition relate to. In consequence, concerning vowel recognition of voiced-like sounds, fundamental frequency – and pitch – do not simply relate to the first harmonic and its multiples of the vowel spectrum, that is, *H*1 and HCF (see also the excursus preceding this main chapter). From this perspective, in two further sinewave vowel synthesis experiments, attempts were made to trigger a vowel quality shift by a change in either the HCF or S_1 (frequency of the lowest sinusoid used in synthesis).

In the first experiment, based on sounds produced with three sinusoids in a harmonic relation, the HCF was altered in terms of changing either S_2–S_3 distance (sounds expected to be recognised as front vowels) or S_1–S_2 distance (sounds expected to be recognised as back vowels): Pairs of S_1–S_2–S_3 configurations were compiled with fixed S_1 and S_3 and varying S_2 only. All three sinusoids of both configurations of a pair were in a harmonic relation, with frequency levels of the HCF being either 0.5×S_1 (configuration a) or equal to S_1 (configuration b), varying HCF by one octave. For sounds of front vowels, S_1 was set below 500 Hz, and S_2 and S_3 were set ≥ 1.2 kHz. For sounds of back vowels, S_1 and S_2 were set to ≤ 1.2 kHz and S_3 was set to 2.8 kHz. Because a smaller frequency change for lower harmonics is related to a larger change in the higher harmonics, which might affect vowel recognition, S_1–S_2–S_3 configurations related to one-octave HCF variations of 200–400 Hz, 210–420 Hz and 220–440 Hz were investigated for front vowels; however, only S_1–S_2–S_3 configurations related to an HCF variation of 200–400 Hz were investigated for back vowels. (For exemplary illustration, see Figure 1; for details of the method, see Chapter M6.7.) The range of HCF was chosen based on the previous experiences regarding sinewave experiments (see Chapters 6.5 and 6.6), regarding formant pattern and spectral shape ambiguity for sounds of adjacent vowels (documented in Chapter 3) and regarding the perceptual effect observed by the author when creating the experiment. As a result, 15 pairs of S_1–S_2–S_3 configurations were created in total. Based on these *S*-patterns, static sounds of 1 sec. (including a 0.1 sec. fade in/out) were synthesised using the SinSyn tool, with the sinewave levels set to 100–80–80 dB for sounds of front vowels and 100–90–70 dB for sounds of back vowels.

Vowel recognition of the synthesised sounds was tested in a listening test according to the standard procedure of the Zurich Corpus and involving the five standard listeners of the corpus, with some test specifications adopted (see Chapter M6.7; note that each sound was presented twice in the test). Pitch recognition of the sounds was tested in an experiment-specific listening test, again involving the five standard listeners of the corpus: Single sounds were presented, and the listeners were asked to label the pitch level they recognised using a prepared paper form and an online electronic piano keyboard (assignment of the dominant or prominent pitch level only, forced choice; the listeners wrote down levels as musical notes within the C-major scale).

When analysing the results of the entire sound sample, vowel and pitch recognition results were found to vary markedly among the different *S*-patterns of the 15 sound pairs investigated. Above all, some pairs showed a marked vowel quality shift in an open–close direction related to both an increase in HCF and an indication of an increase in the recognised pitch level, while for other pairs, no indication of a vowel quality shift or only a weak one was observed. However, for an increase of HCF, neither inverse vowel quality shifts in a close–open direction nor inverse decreasing pitch levels occurred. On this basis, for this treatise, we decided to select four sound pairs related to four recognised close-mid–close vowel quality shifts /e–i/, /e–y/, /ø–y/ and /o–u/ in terms of “best cases” (i.e., highest recognition rates for vowel quality shifts and associated pitch level shifts in the above recognition tests) in order to demonstrate and document cases for which this type of a change of HCF could trigger a parallel vowel quality and pitch level shift.

For all four sound pairs presented, according to the labelling majority, increasing HCF resulted in a close-mid–close vowel quality shift: For the sound pairs recognised as front vowels, an increase in HCF and a related close-mid–close vowel quality shift resulted from a decrease in S_2. Notably, therefore, the sound of /i/ was associated with a lower S_2 than the sound of /e/, in opposition to statistical F_2 generally reported as being higher for sounds of /i/ than of /e/. The same held true for sounds of /y/ and /e/ and of /y/ and /ø/. For the sound pair recognised as back vowels, an increase in HCF and a related close-mid–close vowel quality shift resulted from an increase in S_2. Notably again, there is no general indication given in the literature on formant statistics that a change of only F_2 is related to an /o/–/u/ change in recognised vowel quality. Besides these general results, between-listener differences and recognition inconsistencies (within-listener recognition differences for the two equal sounds presented in the test) were observed. Concerning pitch recognition, according to the majority of labelling, a one-octave upward shift was indicated by the listening test results for three of the four sound pairs. However, the indication was only pronounced for the sounds recognised as back vowels, and marked between-listener differences occurred (see Chapter M6.7). Figure 1 illustrates this main finding of the first experiment.

Because of the somewhat limited indication of parallelism of vowel quality and pitch level shifts and because of the marked listener-specific recognition differences, the results were further analysed concerning the individual listener recognition profiles and the possible combinations of vowel quality and pitch level shifts and shift directions. This analysis showed that vowel quality shifts were associated with either pitch shifts from lower to higher levels or constant higher pitch levels for both sounds of a sound pair. It is noteworthy that no close-mid–close vowel quality shift associated with a downward pitch level shift or lower pitch levels for both sounds of a sound pair occurred. This finding can be understood as supporting the indication of associated vowel quality and pitch level shifts and shift directions, above all when considering the “borderline” character of this type of sound in general and taking into account the very unnatural and sharp sound timbre of synthesised sounds used for this experiment in particular, possibly affecting the results of the vowel and pitch recognition tasks.

In the second experiment, based on sounds with a single lower sinusoid < 1 kHz combined with equal-amplitude sinusoid series in frequency ranges > 1 kHz, all sinusoids in a harmonic relation, periodicity variation was caused by changing either the frequency of the low harmonic < 1 kHz (change of a relative spectral maximum) or the frequency distance of the higher harmonics > 1 kHz (change of HCF) with the frequency range of the higher harmonics kept unchanged. The main idea was to create an experimental design in which two different spectral variations were directly opposed, possibly triggering the same change in the recognised quality of front vowels but with only one type of spectral variation affecting the sound periodicity. At the same time, a spectral peak structure of harmonics > 1 kHz was avoided to create equal bands of higher spectral energy for the configurations compared with each other. In consequence, no filter configuration corresponded to the higher spectrum > 1 kHz.

Accordingly, and on the basis of extensive acoustic analyses of natural front vowel sounds with flat spectral envelopes or flat envelope parts (vowel-related frequency ranges with consecutive harmonics equal in amplitude, see below, Chapter 7.3) and on the basis of a broader investigation of synthesised sounds related to consecutive harmonics equal in amplitude (see below, Chapter 7.4), eight series of sinewave configurations with a single lower harmonic < 1 kHz and with a sequence of consecutive equal-amplitude harmonics in a frequency band > 1 kHz were compiled: Series 1–3 consisted of *S*-pattern triplets, with the first two *S*-patterns differing in S_1 only (220 Hz and 440 Hz), the higher harmonics and HCF being identical (multiples of 220 Hz), and with the second and third *S*-patterns differing in the frequency distance and in the HCF of the higher harmonics (multiples of 220 and 440 Hz), the single low harmonic (440 Hz) and the frequency band of the higher harmonics being identical (for exemplary illustration, see Figure 2). Likewise, but with increased HCF variation, Series 4–6 consisted of *S*-pattern triplets, with the first two patterns differing in S_1 only (300 Hz and 450 Hz), the higher harmonics and HCF being identical (multiples of 150 Hz), and with the second and third pattern differing in the frequency distance and in the HCF of the higher harmonics (multiples of 150 and 450 Hz), the single low harmonic (450 Hz) and the frequency band of the higher harmonics being identical. Also likewise, but with decreased frequency distances of the higher harmonics for all sounds compared, Series 7 consisted of an *S*-pattern triplet, with the first two patterns differing in S_1 only (450 Hz and 600 Hz), the higher harmonics and HCF being identical (multiples of 150 Hz), and with the second and third pattern differing in the frequency distance and in the HCF of the higher harmonics (multiples of 150 and 300 Hz), the single low harmonic (600 Hz) and the frequency band of the higher harmonics being identical. Series 8 consisted of six *S*-patterns, with an increased range of S_1 variation (300–450–600 Hz) for the comparison of S_1 differences and increased HCF variation (150–300–600 Hz) for the comparison of HCF differences (for illustration, see Figure 3). Based on these *S*-patterns, static sounds of 1.2 sec. (including a 0.1 sec. fade in/out) were synthesised using the SinSyn tool, with all sinewave levels set to 100 dB.

Vowel recognition of the synthesised sounds was tested in two listening tests according to the standard procedure of the Zurich Corpus and involving the five standard listeners of the corpus. Labelling was restricted to long Standard German vowel qualities (forced choice, no vowel boundaries). In the first test, the test items consisted of single sounds, which the listeners were asked to assign to a dominant or prominent vowel quality. In the second test, for each series of sounds compared, the test item consisted of two of these sounds presented one after the other (presented in AB and BA order). The listeners were asked to assign the dominant or prominent vowel quality of the second sound only. Sound presentation and test procedure of the pitch recognition test accorded with the second vowel recognition test. The listeners were asked to label whether the pitch level of the second sound when compared to the level of the first sound was falling, flat or rising, referring to dominant or prominent levels.

In general, according to the labelling majority in the vowel and pitch recognition tests, decreasing the frequency level of one single low harmonic but keeping HCF unchanged resulted in a change of vowel quality in an open–close direction: If unrounded–rounded variants were disregarded, the recognition rates for the close-mid–close shifts (all series) and for the open-mid–close-mid shift (Series 8) were ≥ 92%. Likewise, increasing the frequency of HCF but keeping the single low harmonic on an equal frequency level and also keeping the frequency band of the higher harmonics equal resulted, in its turn, in a change of vowel quality in an open–close direction: If unrounded–rounded variants were again disregarded, the recognition rates for the close-mid–close shifts (all series) and the open-mid–close-mid shifts (Series 8) were ≥ 84%. In parallel to these vowel quality shifts, except for three single labellings, all listeners recognised associated upward pitch level shifts. Figures 2 and 3 illustrate this main finding of the second experiment.

In sum, the results of the first experiment showed that, due to the variation of a single intermediate sinusoid frequency of a three-sinusoid sound, an HCF increase by one octave could trigger both a vowel quality shift in an open–close direction and a parallel one-octave upward pitch level shift. However, the demonstration of this double effect was somewhat limited and depended on the specific configuration of the *S*-patterns. The results of the second experiment showed that a vowel quality shift in an open–close direction could be triggered by a downward shift of a single low harmonic only, with the remaining spectral configuration kept unchanged. But, in line with the results of the first experiment, the same shift could also be triggered by changing HCF in terms of changing the frequency distance of the higher harmonics, both the low harmonic and the higher frequency range of prominent spectral energy kept unchanged in synthesis. Thereby, parallel shifts of vowel quality and pitch levels for sounds with HCF variation were very pronounced for sounds of this type.

In conclusion, vowel quality shifts could be triggered by either an energy maximum change in the lower vowel spectrum, with HCF and recognised pitch level kept unchanged, or by a change in HCF and recognised pitch level, with the lower prominent spectral energy as well as the frequency range of the higher prominent spectral energy in the vowel spectrum kept unchanged. Thereby, consistent shift directions were found: open–close vowel quality shifts associated with an increase in the pitch level.

Although not explicitly discussed for the synthesis experiments presented in the two preceding chapters as well as in the present chapter, the listeners involved in the recognition tasks of these experiments constantly reported cases of sounds for which, when giving very specific attention to the sound characteristics during the tests, they could recognise two vowel qualities and/or two (or even more) pitch levels (see also the excursus preceding this sixth main chapter; see also the corresponding note in Chapter M6.1). The experiments discussed in the following chapters were designed and conducted to integrate this double-vowel and/or double-pitch phenomenon into the general investigation of the vowel–pitch relation (or its alternative).

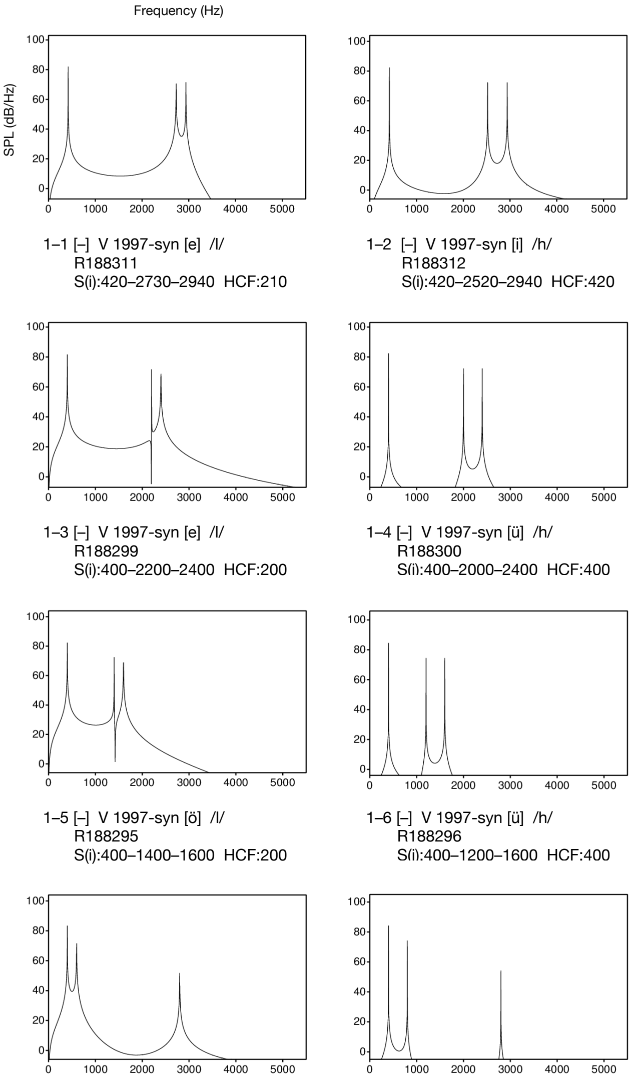
For references, details of the method and results, an extended discussion and documentation of results (tables including sound links), see the Materials, Chapter M6.7.

Figure 1. Vowel and pitch recognition for sounds produced with harmonic sinewave synthesis, synthesis based on S_1–S_2–S_3 patterns with fixed S_1 and S_3 but varying S_2 (experiment 1): Illustration of the main finding. Extract of Chapter M6.7, Table 1. Four pairs of synthesised sounds are shown: Sounds 1 and 2 = sounds with HCF variation of 210–420 Hz recognised as /e/ and /i/ weakly associated with an upward shift of the pitch level. Sounds 3 and 4 = sounds with HCF variation of 200–400 Hz recognised as /e/ and /y/ associated with an upward shift of the pitch level. Sounds 5 and 6 = sounds with HCF variation of 200–400 Hz recognised as /ø/ and /y/ associated with an upward shift of the pitch level. Sounds 7 and 8 = sounds with HCF variation of 200–400 Hz recognised as /o/ and /u/ associated with an upward shift of the pitch level. For the notation of recognised pitch level in slashes (labelling majority), see the paragraph on figures and figure legends in the Introduction. *S*-patterns in all figures of this chapter are given in Hz.
[C-06-07-F01] ↗

Figure 2. Vowel and pitch recognition for sounds produced with harmonic sinewave synthesis, synthesis based on *S*-patterns with a single lower harmonic < 1 kHz and a series of equal-amplitude harmonics > 1 kHz, and including a variation of either the lower sinusoid or HCF (experiment 2): Illustration of the main finding for the triplets of *S*-patterns investigated. Extract of Chapter M6.7, Table 3 (see Series 1, 2 and 7 in this table). Three triplets of synthesised sounds are shown: Sounds 1–3 = sounds with either S_1 variation of 220–440 Hz (sounds 1 and 2) recognised as /i/ and /e/ unassociated with pitch level differences, or HCF variation of 220–440 Hz (sounds 2 and 3) recognised as /e/ and /i/ associated with an upward shift of the pitch level. Sounds 4–6 = sounds with either S_1 variation of 220–440 Hz (sounds 1 and 2) recognised as /y/ and /ø/ unassociated with pitch level differences, or HCF variation of 220–440 Hz (sounds 2 and 3) recognised as /ø/ and /y/ associated with an upward shift of the pitch level. Sounds 7–9 = sounds with either S_1 variation of 450–600 Hz (sounds 1 and 2) recognised as /e/ and /ɛ/ unassociated with pitch level differences, or HCF variation of 150–450 Hz (sounds 2 and 3) recognised as /ɛ/ and /e/ associated with an upward shift of the pitch level.
[C-06-07-F02] ↗

Figure 3. Vowel and pitch recognition for sounds produced with harmonic sinewave synthesis, synthesis based on *S*-patterns with a single lower harmonic < 1 kHz and a series of equal-amplitude harmonics > 1 kHz, and including a variation of either the lower sinusoid or HCF (experiment 2): Illustration of the main finding for the sextuplet of *S*-patterns investigated. Extract of Chapter M6.7, Table 4 (see sounds 1, 2, 4, 5, 6 in this table). Five synthesised sounds are shown: Sounds with either S_1 variation of 300–450–600 Hz (sounds 1–3 in the figure) recognised as /y/ and /e/ and /ɛ/ unassociated with pitch level differences, or HCF variation of 150–300–600 Hz (sounds 3–5 in the figure) recognised as /ɛ/ and /e/ and /y/ associated with an upward shift of the pitch level. Note that the vowel quality shifts exceed adjacent qualities, and the pitch level shifts exceed one octave.
[C-06-07-F03] ↗

Figure 1. Vowel and pitch recognition for sounds produced with harmonic sinewave synthesis, synthesis based on S1–S2–S3 patterns with fixed S1 and S3 but varying S2 (experiment 1): Illustration of the main finding. [C-06-07-F01]



1–1 [–] V 1997-syn [e] /l/
R188311
S(i):420–2730–2940 HCF:210

1–2 [–] V 1997-syn [i] /h/
R188312
S(i):420–2520–2940 HCF:420

1–3 [–] V 1997-syn [e] /l/
R188299
S(i):400–2200–2400 HCF:200

1–4 [–] V 1997-syn [ü] /h/
R188300
S(i):400–2000–2400 HCF:400

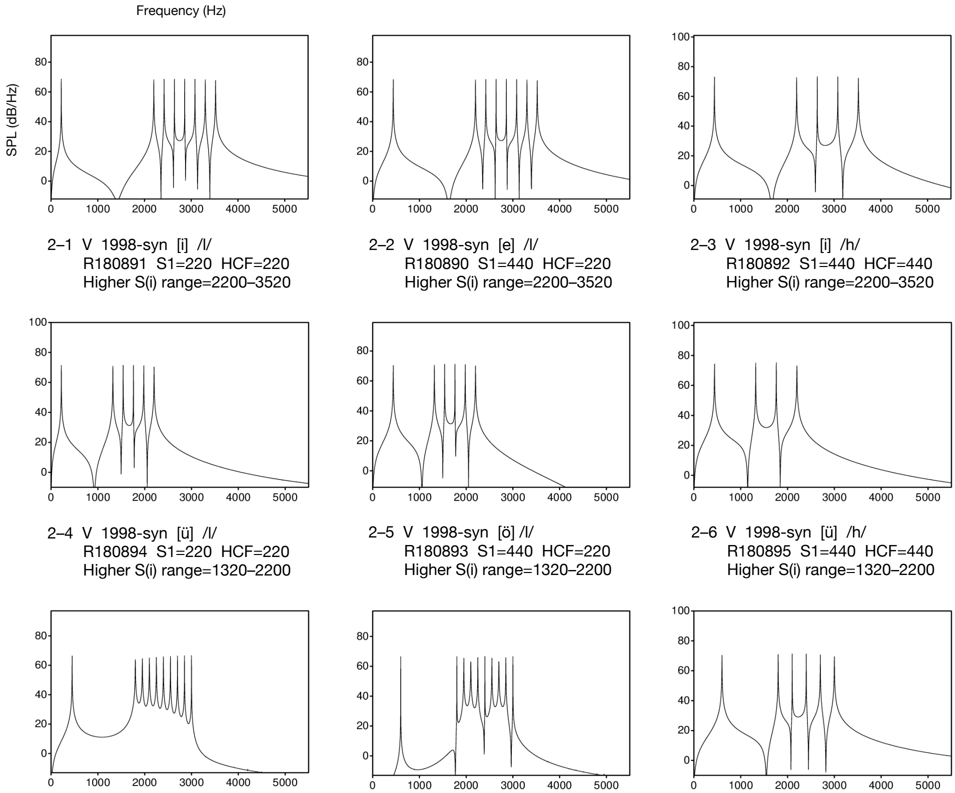
1–5 [–] V 1997-syn [ö] /l/
R188295
S(i):400–1400–1600 HCF:200

1–6 [–] V 1997-syn [ü] /h/
R188296
S(i):400–1200–1600 HCF:400

1–7 [–] V 1997-syn [o] /l/
R188283
S(i):400–600–2800 HCF:200

1–8 [–] V 1997-syn [u] /h/
R188284
S(i):400–800–2800 HCF:400

Figure 2. Vowel and pitch recognition for sounds produced with harmonic sinewave synthesis, synthesis based on S-patterns with a single lower harmonic < 1 kHz and a series of equal amplitude harmonics > 1 kHz, and including a variation of either the lower sinusoid or HCF (experiment 2): Illustration of the main finding for the triplets of S-patterns investigated. [C-06-07-F02]



2–1 V 1998-syn [i] /l/
R180891 S1=220 HCF=220
Higher S(i) range=2200–3520

2–2 V 1998-syn [e] /l/
R180890 S1=440 HCF=220
Higher S(i) range=2200–3520

2–3 V 1998-syn [i] /h/
R180892 S1=440 HCF=440
Higher S(i) range=2200–3520

2–4 V 1998-syn [ü] /l/
R180894 S1=220 HCF=220
Higher S(i) range=1320–2200

2–5 V 1998-syn [ö] /l/
R180893 S1=440 HCF=220
Higher S(i) range=1320–2200

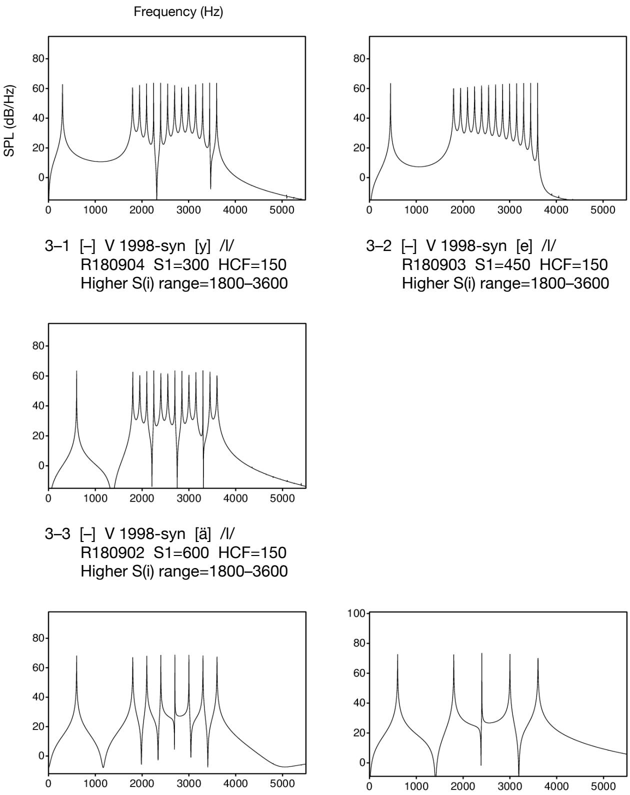
2–6 V 1998-syn [ü] /h/
R180895 S1=440 HCF=440
Higher S(i) range=1320–2200

2–7 V 1998-syn [e] /l/
R180888 S1=450 HCF=150
Higher S(i) range=1800–3000

2–8 V 1998-syn [ä] /l/
R180887 S1=600 HCF=150
Higher S(i) range=1800–3000

2–9 V 1998-syn [e] /h/
R180889 S1=600 HCF=300
Higher S(i) range=1800–3000

Figure 3. Vowel and pitch recognition for sounds produced with harmonic sinewave synthesis, synthesis based on S-patterns with a single lower harmonic < 1 kHz and a series of equal amplitude harmonics > 1 kHz, and including a variation of either the lower sinusoid or HCF (experiment 2): Illustration of the main finding for the sextuplet of S-patterns investigated. [C-06-07-F03]



3–1 [–] V 1998-syn [y] /l/
R180904 S1=300 HCF=150
Higher S(i) range=1800–3600

3–2 [–] V 1998-syn [e] /l/
R180903 S1=450 HCF=150
Higher S(i) range=1800–3600

3–3 [–] V 1998-syn [ä] /l/
R180902 S1=600 HCF=150
Higher S(i) range=1800–3600

3–4 [–] V 1998-syn [e] /ia/
R180905 S1=600 HCF=300
Higher S(i) range=1800–360

3–5 [–] V 1998-syn [y] /h/
R180907 S1=600 HCF=600
Higher S(i) range=1800–360

6.8 Harmonic Synthesis II – Sinewave-Like Replicas Related to Harmonics at or Near Spectral Peaks of Single Natural Vowel Sounds

Monotonous sounds produced with sinewave synthesis for the above experiments were of very artificial sound quality, and vowel and pitch recognition may have been affected as a result. This experimental condition is a major drawback. However, as a first approach to sounds of this type of spectral characteristics, we wanted to exclude all dynamic characteristics.

The results of the first two experiments on sinewave vowel sounds described in Chapters 6.5 and 6.6 indicated that, for static three-sinusoid sound synthesis with *S*-patterns relating to estimated *F*-patterns of natural vowel sounds, vowel recognition was related to pitch, although in a nonuniform manner. The results of the experiments described in the previous chapter supported this indication for synthesised vowel sounds based on sinusoids (with an HCF) which were only indirectly related to natural sounds. Furthermore, although only indicated and not explicitly discussed, the vowel–pitch relation (or its alternative) seemed to be associated with the phenomenon of sounds for which listeners could recognise two vowel qualities and/or two (or even more) pitch levels.

In order to (i) further evaluate vowel and pitch recognition for synthesised sinewave-like vowel sounds related to two or three harmonics (dominant harmonics in the spectrum of a natural reference sound), but using sounds with a more natural sound quality than was the case in the previous experiments, (ii) address at the same time the question of different configurations of spectral peak frequencies for sounds of a given vowel and (iii) extend the investigation in terms of including the question of double-vowel and/or double-pitch recognition, a further sinewave-like experiment on the matter of the vowel–pitch relation was conducted based on extracted harmonics (their dynamic course) of natural vowel sounds at or near the first three estimated peaks in their spectrum: On the basis of the Zurich Corpus, for each of the eight long Standard German vowels and each of the three age- and gender-related speaker groups of men, women and children and related intended levels of f_o of 131 Hz, 220 Hz and 262 Hz, respectively, three natural vowel sounds produced with voiced phonation in nonstyle mode and V context were selected which manifested a spectral peak structure that allowed for the assignment of single harmonics as their representation. Vocal effort was disregarded. The vowel recognition rate was 100% (matching vowel intention) for all selected sounds according to the standard listening test conducted when creating the corpus. Concerning the selection criteria, sounds of the back vowels and /a/ were selected for which two peaks were manifest in the spectrum below 2 kHz; however, a few examples of sounds with only one peak in this frequency range were also included. Sounds of the front vowels were selected for which three or more peaks were manifest in the entire spectrum. In the selection process, the inclusion of some spectral variation for sounds of a single vowel was also attempted: If possible, for a given vowel and a given f_o level, sounds with one or two different spectral peak frequencies and/or peak levels resulting in different patterns of harmonic frequencies and/or levels of dominant harmonics used for synthesis were selected. As a result, a sample of 72 natural reference sounds was created.

For the selected sounds of back vowels and /a/, *D*1–*D*2 patterns in terms of the two dominant or prominent harmonics that corresponded to either the two lower spectral peaks or the estimated spectral envelope of a sound were assigned. For the selected sounds of front vowels, *D*1–*D*2–*D*3 patterns in terms of one dominant or one of two prominent harmonics that corresponded to one of the first three peaks of a sound spectrum were assigned. Note that the selected dominant or prominent harmonics are abbreviated here as *D*1–*D*2 or *D*1–*D*2–*D*3 (including their levels) or D_1–D_2 or D_1–D_2–D_3 (frequencies only), and their patterns are termed *D*-patterns in order not to confuse the number of a dominant or prominent harmonic and the number of any harmonic *H*(i) in the original spectrum of the natural sound (see the Introduction). Subsequently, using the HarmSyn tool, the dynamic harmonic spectra of the natural reference sounds were analysed and, based on this analysis and on the selected two or three harmonics assigned in a *D*-pattern, sounds were synthesised. As a result, a sample of 72 synthesised sounds was created.

Vowel and pitch recognition of the synthesised sounds was investigated in two experiment-specific tests involving the five standard listeners of the Zurich Corpus. In the first test, the listeners were asked to simultaneously label the dominant or prominent vowel quality and the dominant or prominent pitch level. For pitch level recognition, the listeners used an online electronic piano keyboard. They selected a note with a level (according to the musical C-major scale) comparable to the level of the synthesised sound in question. In the second test, the listeners were asked to state whether they heard a second non-dominant or non-prominent vowel quality and/or a second non-dominant or non-prominent pitch level. If this was the case, they were asked to label the respective vowel quality and/or pitch level according to the procedure of the first test.

According to the vowel and pitch recognition results of the first listening test (dominant or prominent vowel qualities and pitch levels, labelling majority), all synthesised replicas of natural close vowel sounds were recognised as close vowels, that is, recognised openness of the synthesised sounds matched with intended openness. In parallel, except for one synthesised sound, the pitch levels were matched to a narrow lower frequency band of 196–262 Hz.

For the synthesised replicas of natural close-mid vowel sounds, the results were somewhat vowel-specific. Concerning /e/, all replicas were confused in terms of a close-mid–close shift when compared with vowel intention. With few exceptions, the recognised pitch levels were assigned to the middle frequency band of 392–440 Hz for the replicas of the adults and equal to or above 523 Hz for the replicas of the children, that is, approximately one octave or more above the replicas of close vowels. Concerning /ø/, seven of nine replicas were confused in terms of a close-mid–close shift when compared with vowel intention. In parallel, except for one sound, the recognised pitch levels mostly corresponded to the levels found for sounds of /e/. Concerning /o/, four of nine replicas were confused in terms of a close-mid–close shift when compared with vowel intention, four replicas were recognised according to vowel intention, and one replica was recognised in the vowel boundary of /o–u/. In parallel, recognised pitch levels were somewhat scattered. However, vowel confusion was markedly higher for sounds with higher pitch levels.

All synthesised replicas of the open-mid vowel were confused (open-mid–close-mid or open-mid–close shifts when compared with vowel intention). In parallel, the recognised pitch levels were somewhat scattered but with a marked tendency towards middle and higher levels up to 659 Hz and above. Concerning /a/, five replicas were mostly recognised according to vowel intention, and four replicas were recognised as /a/ or /ɔ/ or in the /a–ɔ/ boundary. The recognised pitch levels were again scattered, with a tendency towards the middle and higher levels up to 659 Hz and above.

When analysing the relation of the frequency levels of recognised pitch, D_1 and HCF, for the replicas of close, close-mid and open-mid vowel sounds, recognised pitch levels related either to D_1 only or to D_1 and HCF with their frequency levels being equal, with few exceptions. However, the relation was mixed for the sounds of /a/, with the pitch relating to either *D*1 and/or HCF or neither of them.

According to the vowel and pitch recognition results of the second listening test (secondary vowel qualities and pitch levels), for 36 of the 72 synthesised sounds, some listeners recognised secondary vowel qualities and pitch levels. Thus, numerous cases of double-vowel and/or double-pitch recognition occurred. Notably, (i) this kind of double-recognition proved to be dependent on vowel qualities, on the speaker group and/or on the f_0 of the reference sounds, and also on the individual harmonic configurations for sounds of a given vowel, (ii) double-pitch recognition without double-vowel recognition was more frequent than parallel double-recognition of vowel and pitch, (iii) double-vowel recognition without double-pitch recognition was very rare, (iv) parallel double-recognition of vowel and pitch and, to a lesser degree, also double-pitch without double-vowel recognition proved to be listener-specific, and (v) with only a few exceptions, the secondary pitch level was generally recognised as higher than the prominent one and mostly exceeded 523 Hz.

As a first main finding, the results again strongly supported the thesis that vowel quality is related to pitch and that this relation is nonuniform: Marked vowel confusion in terms of marked vowel quality shifts in an open–close direction did not occur for synthesised replicas of natural close vowel sounds and, in parallel, the recognised dominant or prominent pitch level of the synthesised replicas of these vowels did not surpass 262 Hz (labelling majority), in strong contrast to the replicas of the sounds of the other vowels. Conversely, vowel confusion generally occurred for replicas of natural close-mid and open-mid vowel sounds, and, as said, the dominant or prominent pitch level of the replicas of these vowels was mostly recognised as largely above the corresponding levels of the replicas of natural close vowel sounds. It is noteworthy that the occurring vowel quality shifts in an open–close direction related to the shifts of the pitch levels were comparable to the findings reported for *F*-pattern and spectral shape ambiguity and the previous synthesis experiments. The same held true for the nonuniform character of this relation since the effect of a high pitch on vowel recognition was rather limited for sounds of /a/. Figures 1–3 illustrate this first main finding.

Besides these general recognition tendencies, some variation in recognised vowel qualities and pitch levels was also found among speaker groups and individual sounds. This finding indicated a possible impact of the individual harmonic configuration and related HCF a sound synthesis was based on. Figure 4 illustrates this variation.

As a second main finding, most importantly, numerous synthesised replicas were recognised as having two vowel qualities and two pitch levels or two pitch levels only (double-vowel-only recognition was very rare). Figure 5 illustrates this second main finding. However, double-vowel recognition and, to a lesser degree, double-pitch recognition strongly depended on the listeners. Thus, the perception and recognition of vowel quality and pitch level seem to relate to a referencing operation which – at least for vowel sounds of the type investigated here – is to some degree an individual operation of a listener: Depending on an individual listener's attention span or listening focus, simultaneous recognitions of vowel qualities and/or of pitch levels may or may not occur.

As an additional but important finding, according to the listeners' comments, the sound quality of many of the replicas synthesised with the HarmSyn tool was much more natural-like than the synthesised sounds of the previous experiments based on sinusoids. Indeed, the fact that highly recognisable vowel sounds with a natural-like quality can be synthesised based on only two or three extracted harmonics and their dynamic course, as was obtained for the present sample, is remarkable and represents a major gain for future experimental designs. (For sound quality examination, see Figures 1–4 and the corresponding sound links.) Notably, synthesised sounds of this type provide evidence that the spectral shape in terms of an estimated spectral envelope, including the fine structure of the vowel spectrum, is by no means a better acoustic representation of vowel quality than the respective *F*-pattern: There is no spectral fine structure in synthesised sounds based on two or three extracted harmonics of a natural reference sound, and no common spectral shape concept accounts for the documented synthesised sounds.

Sounds for which two vowel qualities and two pitch levels are recognised offer paradigmatic cases for experimental exploration of the vowel–pitch relation thesis. In this context, experiments addressing a transition from one to another recognised vowel quality and, in parallel, from one to another pitch level may provide more experimental evidence on the matter. The experiments reported in the following two chapters address this question.

For details of the method and results, an extended discussion and documentation of results (tables including sound links), see the Materials, Chapter M6.8.

Figure 1. Vowel and pitch recognition for sounds produced with harmonic synthesis based on *D*1–*D*2–*D*3 or *D*1–*D*2 patterns of natural voiced reference sounds: Illustration of the first main finding concerning sounds of close vowels. Extract of Chapter M6.8, Tables 1 and 2 (see sounds 1, 5 and 7 for /i/ and sounds 2, 4 and 7 for /u/ in these tables). Three pairs of a natural reference sound and the respective synthesised *D*1–*D*2–*D*3 replica are shown for the close front vowel /i/, and three corresponding pairs with *D*1–*D*2 replicas are shown for the close back vowel /u/. According to the vowel and pitch recognition results (labelling majority), the dominant or prominent vowel quality of all six replicas matched the intended vowel quality of the natural reference sounds, and their dominant or prominent pitch levels were recognised within a comparatively lower frequency range of 196–262 Hz (/l/). For the notation of recognised pitch level in slashes (labelling majority), see the paragraph on figures and figure legends in the Introduction.
[C-06-08-F01]

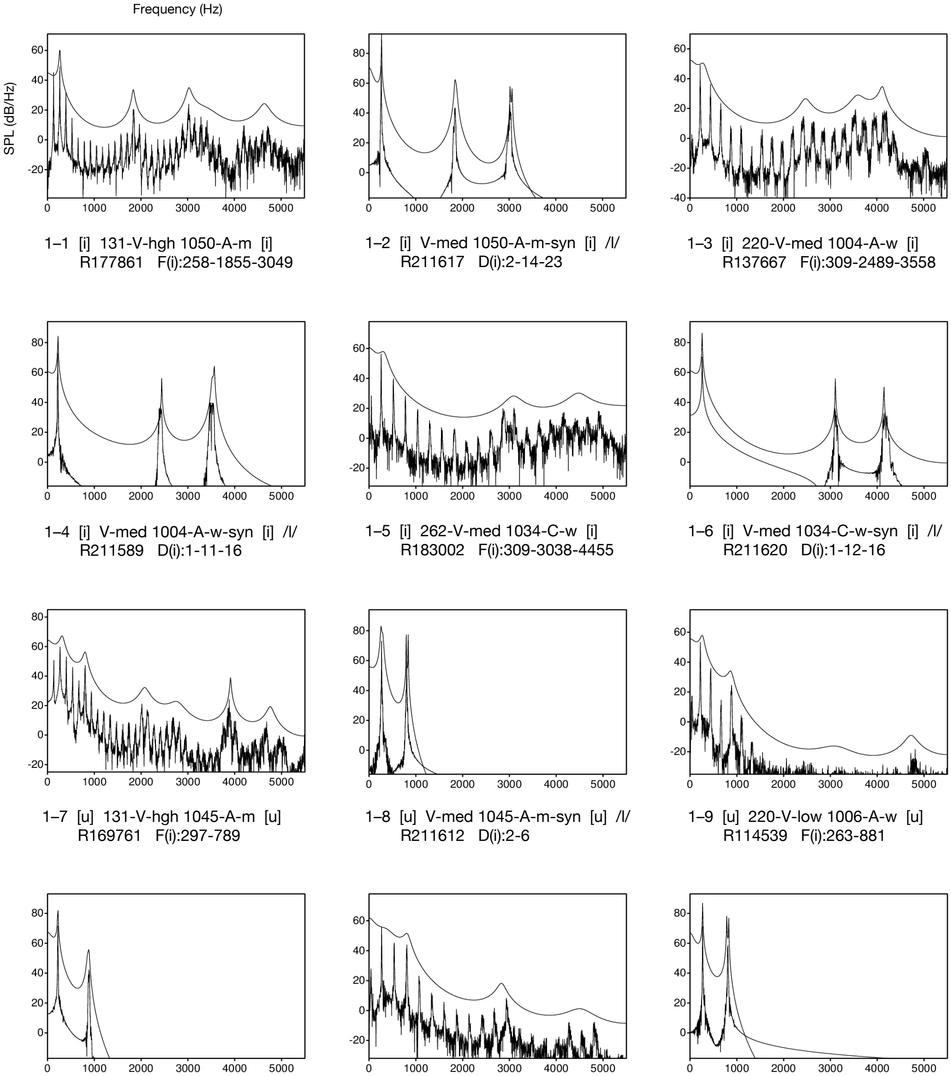
Figure 2. Vowel and pitch recognition for sounds produced with harmonic synthesis based on *D*1–*D*2–*D*3 or *D*1–*D*2 patterns of natural voiced reference sounds: Illustration of the first main finding concerning sounds of close-mid vowels. Extract of Chapter M6.8, Tables 1 and 2 (see sounds 2, 4 and 7 for /e/ and sounds 2, 5 and 9 for /o/ in these tables). Three pairs of a natural reference sound and the respective synthesised *D*1–*D*2–*D*3 replica are shown for the close-mid front vowel /e/, and three corresponding pairs with *D*1–*D*2 replicas are shown for the close-mid back vowel /o/. According to the vowel and pitch recognition results (labelling majority), for all six replicas, close vowels were recognised as dominant or prominent. For the sounds of the adults, the dominant or prominent pitch levels of the replicas were recognised within an intermediate frequency range of 392–440 Hz (/ia/), and for the sounds of the children, the dominant or prominent pitch levels of the replicas were recognised as ≥ 523 Hz (/h/).
[C-06-08-F02]

Figure 3. Vowel and pitch recognition for sounds produced with harmonic synthesis based on *D*1–*D*2–*D*3 or *D*1–*D*2 patterns of natural voiced reference sounds: Illustration of the first main finding concerning sounds of open-mid and open vowels. Extract of Chapter M6.8, Tables 1 and 2 (see sounds 2, 4 and 9 for /ɛ/ and sounds 1, 4 and 8 for /a/ in these tables). Three pairs of a natural reference sound and the respective synthesised *D*1–*D*2–*D*3 replica are shown for the open-mid front vowel /ɛ/, and three corresponding pairs with *D*1–*D*2 replicas are shown for the vowel /a/. According to the vowel and pitch recognition results (labelling majority), close or close-mid vowels were recognised as dominant or prominent for the synthesised replicas of the natural sounds of /ɛ/, and the dominant or prominent pitch levels of the replicas were recognised as ≥ 392 Hz (/ia–h/). For the synthesised replicas of the natural sounds of /a/, recognised dominant or prominent vowel quality was either /a/ or within the /a–ɔ/ range despite recognised dominant or prominent pitch levels of ≥ 392 Hz (/ia–h/).
[C-06-08-F03]

Figure 4. Vowel and pitch recognition for sounds produced with harmonic synthesis based on *D*1–*D*2–*D*3 or *D*1–*D*2 patterns of natural voiced reference sounds: Illustration of some variation in recognised vowel qualities and pitch levels among speaker groups and individual sounds. Extract of Chapter M6.8, Tables 1 and 2 (see sounds 1 and 2 for /ø/ and sounds 6 and 9 for /o/ in these tables). Two natural reference sounds of /ø/ and their synthesised *D*1–*D*2–*D*3 replicas and two reference sounds of /o/ with their *D*1–*D*2 replicas are shown. The synthesised replicas related to *D*-patterns of natural sounds of /ø/ illustrate different pitch level recognition due to different configurations of the frequencies of the dominant harmonics: For the first replica, the dominant or prominent pitch level was recognised within a lower frequency range of 196–262 Hz (/l/), but for the second replica, the recognised level was in an intermediate range of 392–440 Hz (/ia/), both replicas being recognised as /y/ (results according to the labelling majority). The synthesised replicas related to *D*-patterns of natural sounds of /o/ illustrate different vowel quality and pitch level recognition due to *D*-configuration differences: For the first replica, the intended vowel quality was maintained in synthesis and the dominant or prominent pitch level was recognised within the frequency range of 196–440 Hz (/l–ia/), but for the second replica, recognised dominant or prominent vowel quality shifted to /u/ and the pitch level was recognised as equal to or above 523 Hz (/h/).
[C-06-08-F04] ↗

Figure 5. Vowel and pitch recognition for sounds produced with harmonic synthesis based on *D*1–*D*2–*D*3 or *D*1–*D*2 patterns of natural voiced reference sounds: Illustration the second main finding of occurring double-pitch or double-vowel and double-pitch recognition. Extract of Chapter M6.8, Tables 1 to 4 (see sounds 1 and 6 for /y/, sound 3 for /u/, sound 2 for /ø/, sound 9 for /ɛ/, sound 8 for /a/, sound 1 for /e/, sound 8 for /ø/, sounds 3 and 6 for /o/, sound 5 for /ɛ/ and sound 1 for /a/ in these tables, in this order according to the presentation in the figure). Sounds 1–6 = six synthesised replicas of natural vowel sounds are shown for which some listeners recognised two pitch levels. Sounds 7–12 = six replicas are shown for which some listeners recognised two pitch levels and two vowel qualities.
[C-06-08-F05] ↗

Figure 1. Vowel and pitch recognition for sounds produced with harmonic synthesis based on D1–D2–D3 or D1–D2 patterns of natural voiced reference sounds: Illustration of the first main finding concerning sounds of close vowels. [C-06-08-F01]



1–1 [i] 131-V-hgh 1050-A-m [i]
R177861 F(i):258-1855-3049

1–2 [i] V-med 1050-A-m-syn [i] /I/
R211617 D(i):2-14-23

1–3 [i] 220-V-med 1004-A-w [i]
R137667 F(i):309-2489-3558

1–4 [i] V-med 1004-A-w-syn [i] /I/
R211589 D(i):1-11-16

1–5 [i] 262-V-med 1034-C-w [i]
R183002 F(i):309-3038-4455

1–6 [i] V-med 1034-C-w-syn [i] /I/
R211620 D(i):1-12-16

1–7 [u] 131-V-hgh 1045-A-m [u]
R169761 F(i):297-789

1–8 [u] V-med 1045-A-m-syn [u] /I/
R211612 D(i):2-6

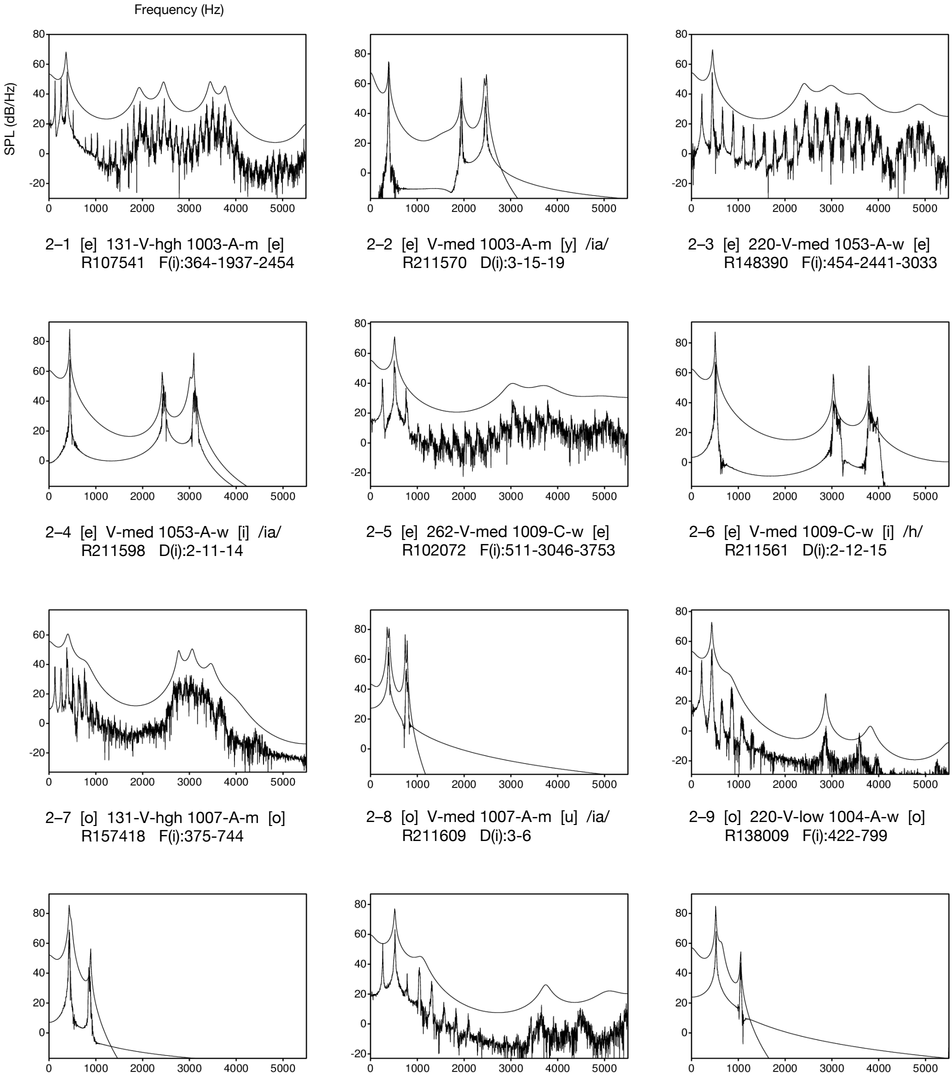
1–9 [u] 220-V-low 1006-A-w [u]
R114539 F(i):263-881

1–10 [u] V-med 1006-A-w-syn [u] /I/
R211573 D(i):1-4

1–11 [u] 262-V-hgh 1034-C-w [u]
R120870 F(i):355-818

1–12 [u] V-med 1034-C-w-syn [u] /I/
R211579 D(i):1-3

Figure 2. Vowel and pitch recognition for sounds produced with harmonic synthesis based on D1–D2–D3 or D1–D2 patterns of natural voiced reference sounds: Illustration of the first main finding concerning sounds of close-mid vowels. [C-06-08-F02]



2–1 [e] 131-V-hgh 1003-A-m [e]
R107541 F(i):364-1937-2454

2–2 [e] V-med 1003-A-m [y] /ia/
R211570 D(i):3-15-19

2–3 [e] 220-V-med 1053-A-w [e]
R148390 F(i):454-2441-3033

2–4 [e] V-med 1053-A-w [i] /ia/
R211598 D(i):2-11-14

2–5 [e] 262-V-med 1009-C-w [e]
R102072 F(i):511-3046-3753

2–6 [e] V-med 1009-C-w [i] /h/
R211561 D(i):2-12-15

2–7 [o] 131-V-hgh 1007-A-m [o]
R157418 F(i):375-744

2–8 [o] V-med 1007-A-m [u] /ia/
R211609 D(i):3-6

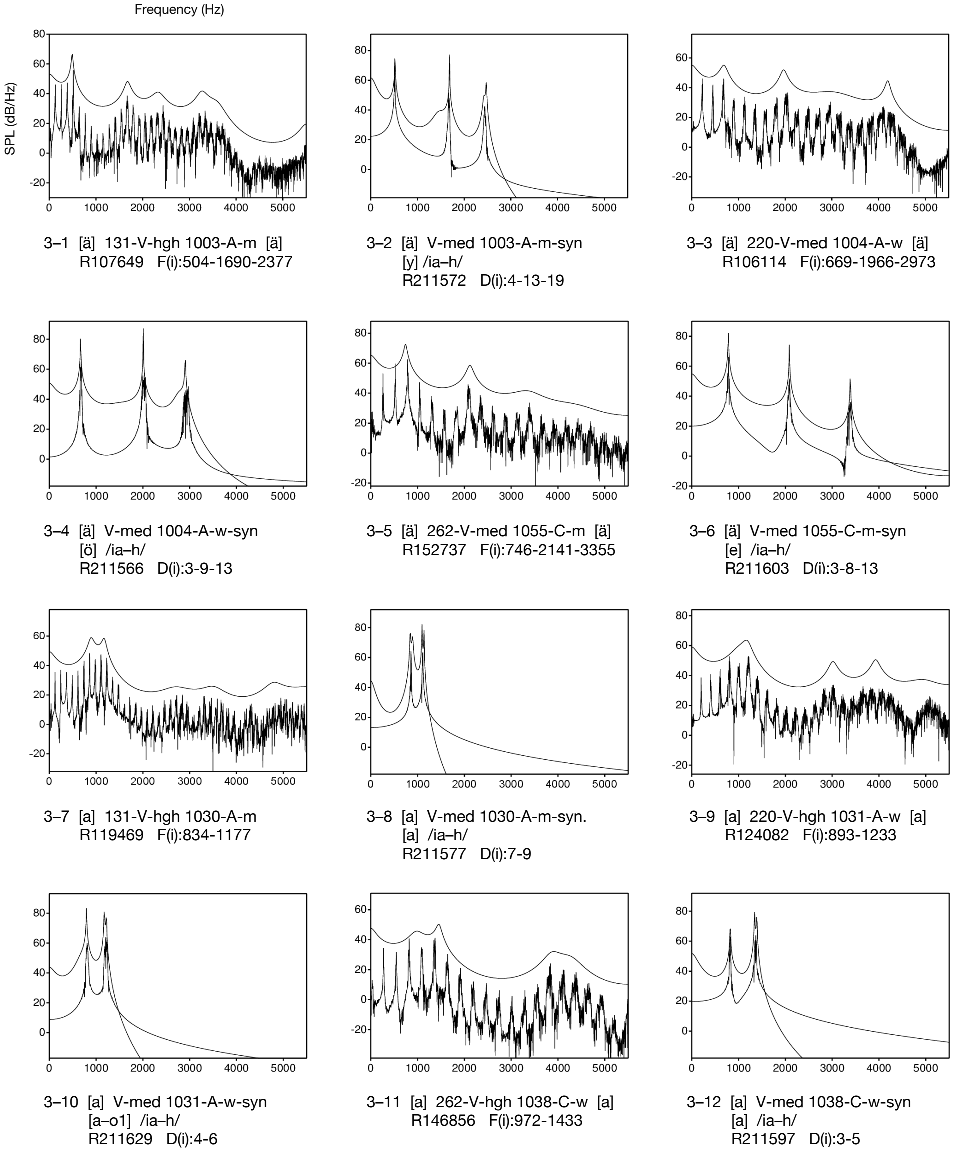
2–9 [o] 220-V-low 1004-A-w [o]
R138009 F(i):422-799

2–10 [o] V-med 1004-A-w [u] /ia/
R211591 D(i):2-4

2–11 [o] 262-V-med 1057-C-m [o]
R143592 F(i):522-1128

2–12 [o] V-med 1057-C-m [u] /h/
R211595 D(i):2-4

Figure 3. Vowel and pitch recognition for sounds produced with harmonic synthesis based on D1–D2–D3 or D1–D2 patterns of natural voiced reference sounds: Illustration of the first main finding concerning sounds of open-mid and open vowels. [C-06-08-F03]



3–1 [ä] 131-V-hgh 1003-A-m [ä]
R107649 F(i):504-1690-2377

3–2 [ä] V-med 1003-A-m-syn
[y] /ia–h/
R211572 D(i):4-13-19

3–3 [ä] 220-V-med 1004-A-w [ä]
R106114 F(i):669-1966-2973

3–4 [ä] V-med 1004-A-w-syn
[ö] /ia–h/
R211566 D(i):3-9-13

3–5 [ä] 262-V-med 1055-C-m [ä]
R152737 F(i):746-2141-3355

3–6 [ä] V-med 1055-C-m-syn
[e] /ia–h/
R211603 D(i):3-8-13

3–7 [a] 131-V-hgh 1030-A-m
R119469 F(i):834-1177

3–8 [a] V-med 1030-A-m-syn.
[a] /ia–h/
R211577 D(i):7-9

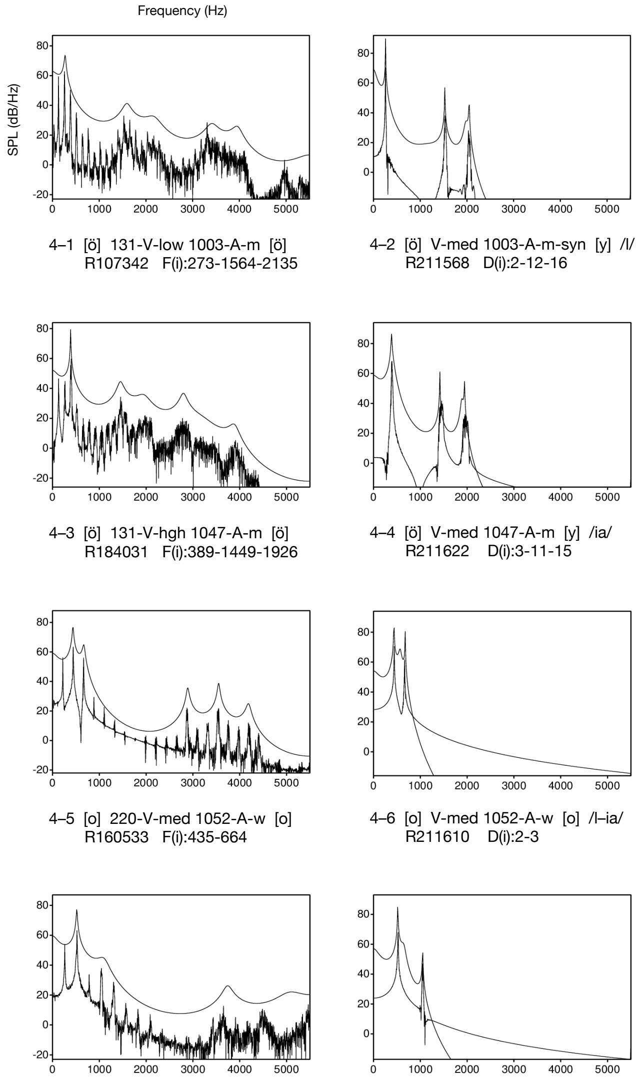
3–9 [a] 220-V-hgh 1031-A-w [a]
R124082 F(i):893-1233

3–10 [a] V-med 1031-A-w-syn
[a–o1] /ia–h/
R211629 D(i):4-6

3–11 [a] 262-V-hgh 1038-C-w [a]
R146856 F(i):972-1433

3–12 [a] V-med 1038-C-w-syn
[a] /ia–h/
R211597 D(i):3-5

Figure 4. Vowel and pitch recognition for sounds produced with harmonic synthesis based on D1–D2–D3 or D1–D2 patterns of natural voiced reference sounds: Illustration of some variation in recognised vowel qualities and pitch levels among speaker groups and individual sounds. [C-06-08-F04]



4–1 [ö] 131-V-low 1003-A-m [ö]
R107342 F(i):273-1564-2135

4–2 [ö] V-med 1003-A-m-syn [y] /I/
R211568 D(i):2-12-16

4–3 [ö] 131-V-hgh 1047-A-m [ö]
R184031 F(i):389-1449-1926

4–4 [ö] V-med 1047-A-m [y] /ia/
R211622 D(i):3-11-15

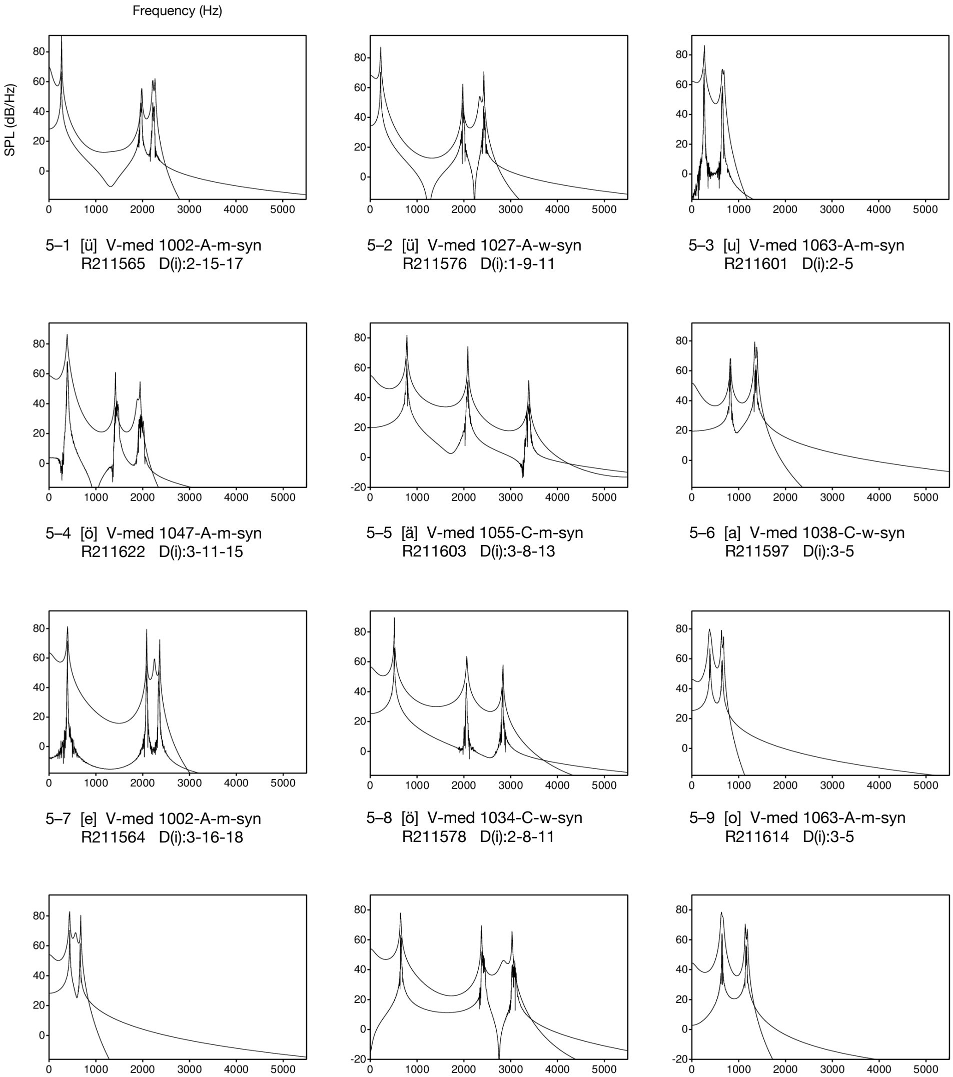
4–5 [o] 220-V-med 1052-A-w [o]
R160533 F(i):435-664

4–6 [o] V-med 1052-A-w [o] /I–ia/
R211610 D(i):2-3

4–7 [o] 262-V-med 1057-C-m [o]
R143592 F(i):522-1128

4–8 [o] V-med 1057-C-m [u] /h/
R211595 D(i):2-4

Figure 5. Vowel and pitch recognition for sounds produced with harmonic synthesis based on D1–D2–D3 or D1–D2 patterns of natural voiced reference sounds: Illustration of the second main finding of occurring double-pitch or double-vowel and double–pitch recognition. [C-06-08-F05]



5–1 [ü] V-med 1002-A-m-syn
R211565 D(i):2-15-17

5–2 [ü] V-med 1027-A-w-syn
R211576 D(i):1-9-11

5–3 [u] V-med 1063-A-m-syn
R211601 D(i):2-5

5–4 [ö] V-med 1047-A-m-syn
R211622 D(i):3-11-15

5–5 [ä] V-med 1055-C-m-syn
R211603 D(i):3-8-13

5–6 [a] V-med 1038-C-w-syn
R211597 D(i):3-5

5–7 [e] V-med 1002-A-m-syn
R211564 D(i):3-16-18

5–8 [ö] V-med 1034-C-w-syn
R211578 D(i):2-8-11

5–9 [o] V-med 1063-A-m-syn
R211614 D(i):3-5

5–10 [o] V-med 1052-A-w-syn
R211610 D(i):2-3

5–11 [ä] V-med 1023-A-w-syn
R211616 D(i):3-11-14

5–12 [a] V-med 1003-A-m-syn
R211569 D(i):5-9

6.9 Harmonic Synthesis III – Sinewave-Like Replicas Related to Non-Dominant *H*1 and to Harmonics at or Near Spectral Peaks of Single Natural Vowel Sounds, With Gradual Attenuation of *H*1 Causing Double-Vowel and Double-Pitch Recognition

In the synthesis experiments based on a few harmonics discussed in the preceding Chapters 6.5 to 6.8, the same tendency of nonuniform open–close vowel quality shifts with increasing pitch was observed, as was demonstrated in the chapters on formant pattern and spectral shape ambiguity. At the same time, for many cases of these synthesised sounds, some listeners reported that they recognised either two vowel qualities and two pitch levels or two (or even more) pitch levels. In rare cases, they recognised two vowel qualities only.

As indicated in the previous chapter, sounds for which some listeners may recognise two vowel qualities and two pitch levels offer paradigmatic cases for experimental exploration of the vowel–pitch relation thesis. Above all, experiments addressing a transition from one to another recognised vowel quality and, in parallel, from one to another pitch level may provide crucial experimental evidence on the matter. The experiments reported in this and the next chapter address this matter. Based on the above experiences and reflection, the further developed experimental approach focused on the investigation of sound series with transitions from one vowel quality associated with a lower pitch level to another vowel quality associated with a higher pitch level: Starting from one voiced sound with mostly unambiguous single vowel and pitch recognition, is it possible to create a series of sounds by stepwise lowering the level of a specific harmonic or the levels of a series of harmonics to, firstly, create sounds with two competing HCFs (HCF of all harmonics and HCF of the harmonics with unaltered levels) and two recognised pitch levels and possibly also two recognised vowel qualities until, secondly, the vowel quality and the pitch level fully shift to a second sound with unambiguous recognition?

In the first study discussed in the present chapter, this question was investigated with regard to synthesised sounds based on non-dominant *H*1 and the two or three vowel-related dominant or prominent harmonics *D*1–*D*2 (back vowels) or *D*1–*D*2–*D*3 (front vowels) in the spectrum of a natural voiced reference sound, the HCF of the dominant or prominent harmonics being higher than the HCF of the entire *H*1–*D*1–*D*2 or *H*1–*D*1–*D*2–*D*3 pattern: Based on the Zurich Corpus, for each of the three close-mid Standard German vowels /e, ø, o/, two sounds produced by women with medium vocal effort in V context and nonstyle mode at calculated f_o in the range of 200–250 Hz were selected, fulfilling the following conditions: (i) The spectral peaks generally assumed to relate to vowel quality were associated with single dominant harmonics near the frequencies of estimated formants, formant estimation being methodologically substantiated. For sounds of /o/, however, a prominent spectral energy above a single lower spectral peak and a correspondingly estimated second formant related to only a prominent harmonic were also accepted; (ii) *H*2 represented the first dominant harmonic *D*1; (iii) the higher dominant (or prominent) harmonic(s) were integer multiples of *D*1. As a result, a sample of six natural reference sounds was created. (Sound duration was disregarded.) For all six sounds, the dominant harmonics *D*1–*D*2–*D*3 (sounds of /e, ø/) and the dominant or prominent harmonics *D*1–*D*2 (sounds of /o/) were assigned, and *H*1 was added to a *D*-configuration to create the harmonic patterns for the subsequent sound synthesis. For details of the method, see Chapter M6.9 in the Materials.

For each single natural reference sound, the dynamic course of its harmonic spectrum for the entire sound duration was analysed using the analysis function of the HarmSyn tool (default parameter setting). Subsequently, based on this analysis and the selected harmonics *H*1–*D*1–*D*2–*D*3 (sounds of the front vowels /e, ø/) or *H*1–*D*1–*D*2 (sounds of the back vowel /o/), eight replicas applying eight attenuation levels for *H*1 of 0/-5/-10/-15/-20/-30/-50/-100 dB were created using the harmonic synthesis function of the HarmSyn tool. As a result, six series of eight replicas with a one-octave transition of dominant HCF from a lower to a higher level were created, and a total of 48 synthesised sounds were investigated in the listening tests. (For an illustration of the experimental design, see Figures 1 and 2.)

Vowel and pitch recognition of the opposing replicas of a series with unchanged *H*1 and with fully deleted *H*1 were tested in four subtests involving the five standard listeners of the Zurich Corpus and applying the standard procedure of the corpus, with the following additional specifications: In the first subtest, S1, each test item consisted of a single replica and the listeners were asked to assign the dominant or prominent vowel quality (forced choice, all long Standard German vowels and schwa, no vowel boundaries). In the subtests S2–S4, each test item consisted of the two opposing replicas of a sound series (separated by a 1 sec. pause), the replica with unchanged *H1* versus the replica with fully deleted *H1*, or vice versa (sound pairs tested in AB and BA order). In subtest S2, the listeners were asked to assign a vowel quality to the second sound presented (forced choice; for vowel qualities, see above). In subtest S3, the listeners were asked to compare the pitch levels of the first and second sounds presented and to label the corresponding level difference as falling, rising or flat. In subtest S4, the listeners were asked to compare both vowels and both pitch levels of the two sounds presented and to assign the two recognised (dominant or prominent) vowel qualities of the two sounds as well as the recognised (dominant or prominent) pitch level difference according to the order of sound presentation.

Vowel and pitch recognition of the replicas of a series with attenuated *H*1 (hereafter transitional sounds) was tested in three subsequent subtests, S5–S7, involving the five standard listeners of the Zurich Corpus and applying the standard procedure of the corpus, with the following additional specifications: In subtest S5, each test item consisted of a single synthesised sound with attenuated *H*1 and the listeners were asked to assign the dominant or prominent vowel quality (forced choice; for vowel qualities, see above). In subtests S6 and S7, each test item consisted of two synthesised replicas of a series (separated by a 1 sec. pause), the replica with unchanged *H*1 versus a second replica with attenuated or deleted *H*1 or, inversely, the replica with deleted *H*1 versus a second replica with unchanged or attenuated *H*1. In subtest S6, the listeners were asked to assign a vowel quality to the second sound presented (forced choice; for vowel qualities, see above). In subtest S7, the listeners were asked to compare the pitch levels of the two sounds presented and to label the corresponding pitch difference as falling, rising or flat.

Double-vowel and double-pitch recognition of all replicas of a series were tested in two further subtests, S8 and S9, involving the same listeners of the Zurich Corpus and applying the standard procedure of the corpus, with the following additional test specifications: In subtest S8, each test item consisted of a single replica presented and the listeners were asked whether they recognised one or two vowel qualities. (No details of vowel qualities were labelled.) Likewise, in subtest S9, single replicas were presented, and the listeners were asked whether they recognised one or two pitch levels.

According to the results of vowel and pitch recognition subtests S1–S4 for the opposing replicas of a *D*-pattern (without and with full attenuation of *H*1; labelling majority), the synthesised sounds based on the *H*1–*D*1–*D*2–*D*3 or *H*1–*D*1–*D*2 patterns with unattenuated *H*1 were associated with a lower pitch level and a close-mid vowel quality, and the synthesised sounds based on the *D*1–*D*2–*D*3 or *D*1–*D*2 patterns only – full level attenuation of *H*1 in synthesis – were associated with a higher pitch level and a close vowel quality. The recognition rate of vowel openness was 80–100%, and pitch recognition was uniform among all listeners.

The results of the vowel and pitch recognition subtests S5–S7 for the transitional replicas of a *D*-pattern with stepwise attenuation of *H*1 from -5 to -50 dB and their comparison with the results of the opposing replicas showed that, as a tendency, recognition of close-mid vowel qualities and lower pitch levels was maintained mostly for weak *H*1 attenuation up to -10, -15 or -20 dB, and it markedly shifted to close qualities and higher pitch levels for *H*1 attenuation of -30 or -50 dB. At the same time, in contrast to the results for the opposing sounds of a series, marked between-listener recognition differences occurred in the transition from close-mid vowels and lower pitch levels (no attenuation of *H*1) to close vowels and higher pitch levels (full attenuation of *H*1): Testing vowel and pitch recognition of the sound series investigated resulted in distinct recognition profiles of single listeners, with the labelling of the listeners markedly differing in the attenuation levels of *H*1 associated with a shift in vowel quality and/or pitch level as well as in the labelling consistency.

The results of double-vowel and double-pitch recognition subtests S8 and S9 showed that (i) for each of the sound series investigated, sounds occurred for which two vowel qualities and/or two pitch levels were recognised, (ii) double-vowel and double-pitch recognition was most pronounced for *H*1 attenuation in the range of -15 to -30 dB, (iii) double-pitch recognition occurred more often than double-vowel recognition, (iv) double-vowel recognition unparalleled by double-pitch recognition was rare, and (v) between-listener recognition differences occurred again.

In sum, once again, the results strongly supported the vowel–pitch relation hypothesis, here in terms of a pronounced general tendency for a close-mid–close vowel quality shift associated with an increase in recognised pitch level. Further, and most importantly, for numerous cases of transitional replicas with stepwise attenuation of *H*1, two vowels and/or two pitches were recognised, a finding that strongly underpins the vowel–pitch relation hypothesis. Figure 3 illustrates these two main findings.

Comparable to the experiment described in Chapter 6.7, a change in HCF triggered the pitch level shift. However, in the present experiment, the lower HCF level of a sound series was created by using (non-dominant) *H*1 for synthesis. This experimental condition may explain why, in contrast to the experiment of Chapter 6.7 using intermediate harmonics to alter HCF, listeners unanimously recognised a pitch level difference for the opposing sounds without and with full attenuation of *H*1.

It is noteworthy that within the transitional phase between recognised low-pitched sounds of close-mid vowels and high-pitched sounds of close vowels, vowel quality and pitch level shifts did not obey strict parallelism: Concerning single identifications of single listeners and in relation to attenuation levels of *H*1, pitch level shifts occurred that were either unassociated with vowel quality shifts or preceding or succeeding them (if looked at from the perspective of increasing *H*1 level attenuation within a sound series). Vice versa, but only rarely, some vowel quality shifts occurred without a simultaneous pitch level shift.

The finding that marked between-listener recognition differences occurred for the transitional sounds underpinned the indication of a vowel–pitch relation (or its alternative): It pointed to a corresponding perceptual referencing which, in our view, is expected to be listener-specific for the created sound transitions (see also the next chapter).

Impressively, the synthesised sounds presented here highlight anew the fact that neither the *F*-pattern nor the spectral envelope *per se* acoustically represents vowel quality: Spectral maxima and estimated *F*-patterns did not alter for the sounds investigated in the above series despite the occurrence of vowel quality shifts, and current concepts of spectral envelope estimation do not account for the harmonic configurations of the vowel spectra of these sounds, as was the case for the sounds of the previous experiments. As a consequence, the spectrograms of the two opposing sounds of a *D*-pattern without and with full attenuation of *H*1, recognised as two different vowel qualities, do not reflect this quality difference in terms of pronounced differences of spectral energy maxima (for illustration, see Figure 4).

For references, details of experimental design, method and results, an extended discussion of several aspects such as listener-specific recognition profiles, the lack of imperative parallelism between double-vowel and double-pitch recognition, the comparison of calculated f_o, HCF and recognised pitch level for sounds with stepwise attenuation of *H*1, some relativisations that have to be made when interpreting the results of the present experiment and possible improvements to the experimental design with regard to future research, and for documentation of the sound sample and the results of the investigation (tables including sound links), see the Materials, Chapter M6.9.

For an indication of synthesis and vowel and pitch level recognition as given in the below figures, including double-vowel and double-pitch recognition, see the Introduction. Note also that, in these figures, the recognition results are given according to Table 3 of Chapter M6.9 (see coloured results in this table). The attenuation of the level of *H*1 is abbreviated as AH1.

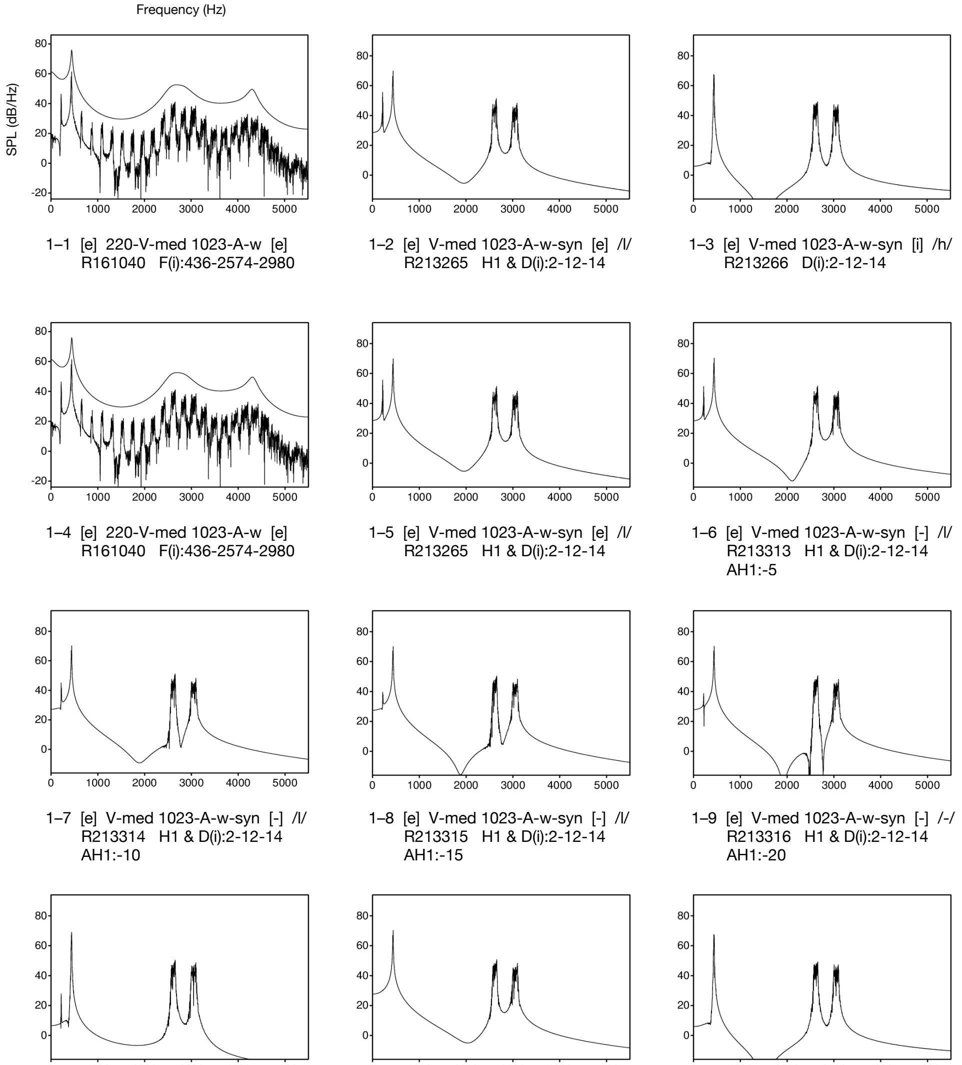
Figure 1. Harmonic synthesis based on non-dominant *H*1 and *D*-patterns of natural voiced reference sounds, including stepwise *H*1 attenuation: Illustration of the experimental design for sounds of front vowels. Extract of Chapter M6.9, Tables 1 to 3 (see Series 1 in these tables). Sounds 1–3 = a natural reference sound of the close-mid vowel /e/ produced by a woman at an intended f_o of 220 Hz, and the two opposing synthesised replicas based on *H*1 and the three dominant harmonics *D*1–*D*2–*D*3 of the reference sound, the first replica with unchanged *H*1 and the second replica with fully deleted *H*1. Sounds 4–12 = the natural reference sound and the eight synthesised replicas with stepwise attenuation of the *H*1 level (AH1).
[C-06-09-F01] ↗

Figure 2. Harmonic synthesis based on non-dominant *H*1 and *D*-patterns of natural voiced reference sounds, including stepwise *H*1 attenuation: Illustration of the experimental design for sounds of back vowels. Extract of Chapter M6.9, Tables 1 to 3 (see Series 5 in these tables). Sounds 1–3 = a natural reference sound of the close-mid vowel /o/ produced by a woman at an intended f_o of 220 Hz, and the two opposing synthesised replicas based on *H*1 and the two dominant or prominent harmonics *D*1–*D*2 of the reference sound, the first replica with unchanged *H*1 and the second replica with fully deleted *H*1. Sounds 4–12 = the natural reference sound and the eight synthesised replicas with stepwise attenuation of the *H*1 level.
[C-06-09-F02] ↗

Figure 3. Harmonic synthesis based on non-dominant *H*1 and *D*-patterns of natural voiced reference sounds, including stepwise *H*1 attenuation: Three examples of vowel quality and pitch level shifts and intermediate double-vowel and double-pitch recognition. Extract of Chapter M6.9, Table 3 (see synthesised replicas 1, 4 and 8 of Series 1 for the /e/–/i/ transition, replicas 1, 6 and 8 of Series 3 for the /ø/–/y/ transition, and replicas 1, 6 and 8 of Series 6 for the /o/–/u/ transition). For each of the three examples, the sound with *H*1 unattenuated, the sound with *H*1 attenuated and the sound with *H*1 deleted are shown. Sounds 1–3: According to the vowel and pitch recognition results (labelling majority), sound 1 was recognised as /e/ at a comparably lower pitch level, sound 2 was recognised as /e/ and /i/ and as associated with a lower and a higher pitch level (occurring double-vowel and double-pitch recognition, [d] and /d/), and sound 3 was recognised as /i/ at a higher pitch level. Sounds 4–6: Sound 4 was recognised as /ø/ at a lower pitch level, sound 5 was recognised as /ø/ and /y/ and as associated with a lower and a higher pitch level, and sound 6 was recognised as /y/ at a higher pitch level. Sounds 7–9: Sound 7 was recognised as /o/ at a lower pitch level, sound 8 was recognised as /o/ and /u/ and as associated with a lower and a higher pitch level, and sound 9 was recognised as /u/ at a higher pitch level.
[C-06-09-F03] ↗

Figure 4. Harmonic synthesis based on non-dominant *H*1 and *D*-patterns of natural voiced reference sounds, including stepwise *H*1 attenuation: Illustration of undetectable vowel quality differences in the spectrographic analysis. The spectrograms of sounds 1–3 in Figure 3 are shown. According to the rules of estimating the vowel quality's spectrographic characteristics, the recognised vowel quality shifts are not reflected in the spectrograms for the three sounds shown.
[C-06-09-F04] ↗

Figure 1. Harmonic synthesis based on nondominant H1 and D-patterns of natural voiced reference sounds, including stepwise H1 attenuation: Illustration of the experimental design for sounds of front vowels. [C-06-09-F01]



1–1 [e] 220-V-med 1023-A-w [e]
R161040 F(i):436-2574-2980

1–2 [e] V-med 1023-A-w-syn [e] /l/
R213265 H1 & D(i):2-12-14

1–3 [e] V-med 1023-A-w-syn [i] /h/
R213266 D(i):2-12-14

1–4 [e] 220-V-med 1023-A-w [e]
R161040 F(i):436-2574-2980

1–5 [e] V-med 1023-A-w-syn [e] /l/
R213265 H1 & D(i):2-12-14

1–6 [e] V-med 1023-A-w-syn [-] /l/
R213313 H1 & D(i):2-12-14
AH1:-5

1–7 [e] V-med 1023-A-w-syn [-] /l/
R213314 H1 & D(i):2-12-14
AH1:-10

1–8 [e] V-med 1023-A-w-syn [-] /l/
R213315 H1 & D(i):2-12-14
AH1:-15

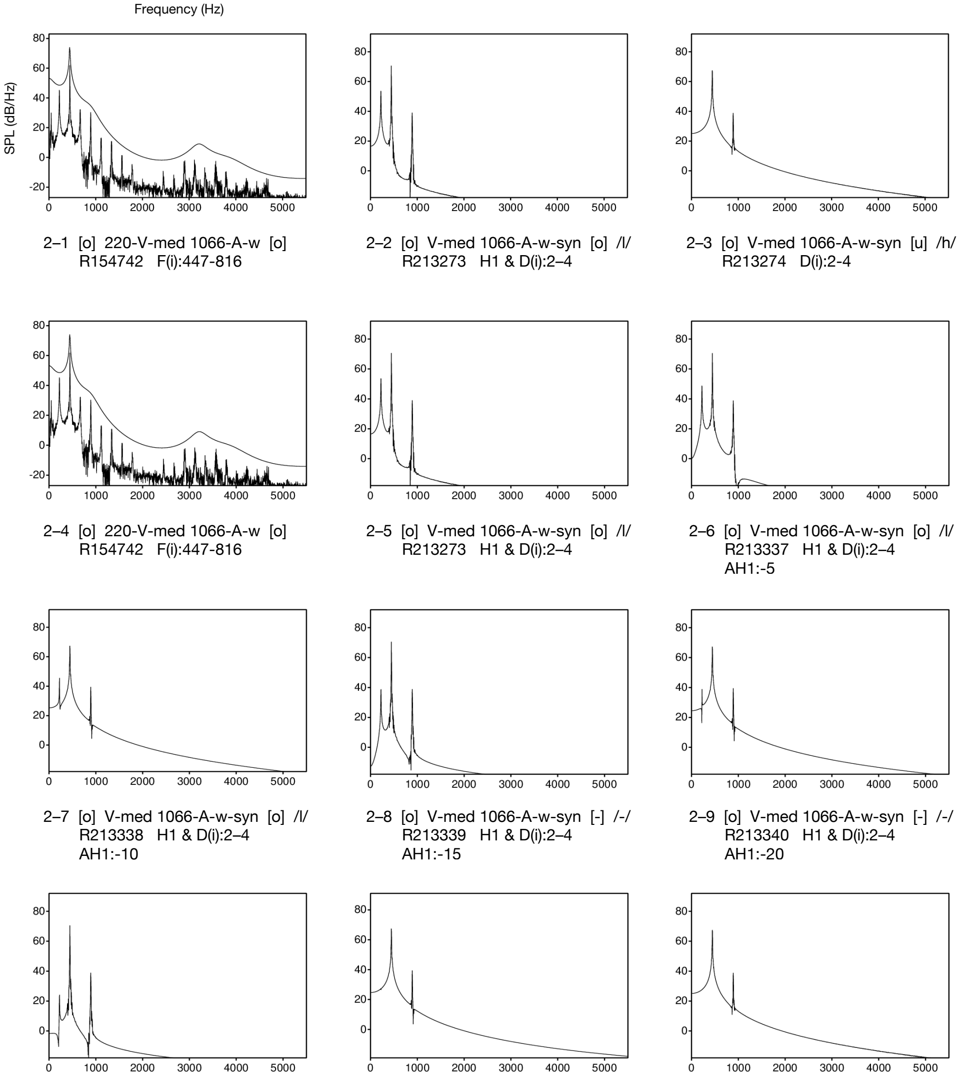
1–9 [e] V-med 1023-A-w-syn [-] /-/
R213316 H1 & D(i):2-12-14
AH1:-20

1–10 [e] V-med 1023-A-w-syn [i] /h/
R213317 H1 & D(i):2-12-14
AH1:-30

1–11 [e] V-med 1023-A-w-syn [i] /h/
R213318 H1 & D(i):2-12-14
AH1:-50

1–12 [e] V-med 1023-A-w-syn [i] /h/
R213266 D(i):2-12-14

Figure 2. Harmonic synthesis based on nondominant H1 and D-patterns of natural voiced reference sounds, including stepwise H1 attenuation: Illustration of the experimental design for sounds of back vowels. [C-06-09-F02]



2–1 [o] 220-V-med 1066-A-w [o]
R154742 F(i):447-816

2–2 [o] V-med 1066-A-w-syn [o] /l/
R213273 H1 & D(i):2–4

2–3 [o] V-med 1066-A-w-syn [u] /h/
R213274 D(i):2-4

2–4 [o] 220-V-med 1066-A-w [o]
R154742 F(i):447-816

2–5 [o] V-med 1066-A-w-syn [o] /l/
R213273 H1 & D(i):2–4

2–6 [o] V-med 1066-A-w-syn [o] /l/
R213337 H1 & D(i):2–4
AH1:-5

2–7 [o] V-med 1066-A-w-syn [o] /l/
R213338 H1 & D(i):2–4
AH1:-10

2–8 [o] V-med 1066-A-w-syn [-] /-/
R213339 H1 & D(i):2–4
AH1:-15

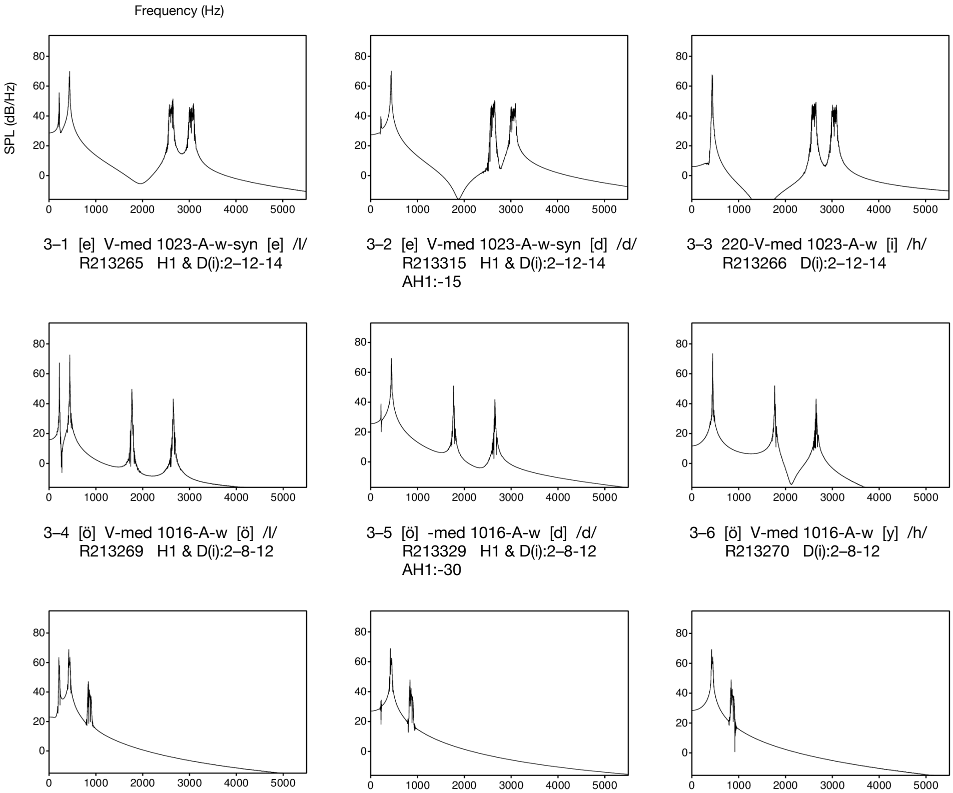
2–9 [o] V-med 1066-A-w-syn [-] /-/
R213340 H1 & D(i):2–4
AH1:-20

2–10 [o] V-med 1066-A-w-syn [-] /-/
R213341 H1 & D(i):2–4
AH1:-30

2–11 [o] V-med 1066-A-w-syn [u] /h/
R213342 H1 & D(i):2–4
AH1:-50

2–12 [o] V-med 1066-A-w-syn [u] /h/
R213274 D(i):2–4

Figure 3. Harmonic synthesis based on nondominant H1 and D-patterns of natural voiced reference sounds, including stepwise H1 attenuation: Three examples of vowel quality and pitch level shifts and intermediate double-vowel and double-pitch recognition. [C-06-09-F03]



3–1 [e] V-med 1023-A-w-syn [e] /l/
R213265 H1 & D(i):2–12-14

3–2 [e] V-med 1023-A-w-syn [d] /d/
R213315 H1 & D(i):2–12-14
AH1:-15

3–3 220-V-med 1023-A-w [i] /h/
R213266 D(i):2–12-14

3–4 [ö] V-med 1016-A-w [ö] /l/
R213269 H1 & D(i):2–8-12

3–5 [ö] -med 1016-A-w [d] /d/
R213329 H1 & D(i):2–8-12
AH1:-30

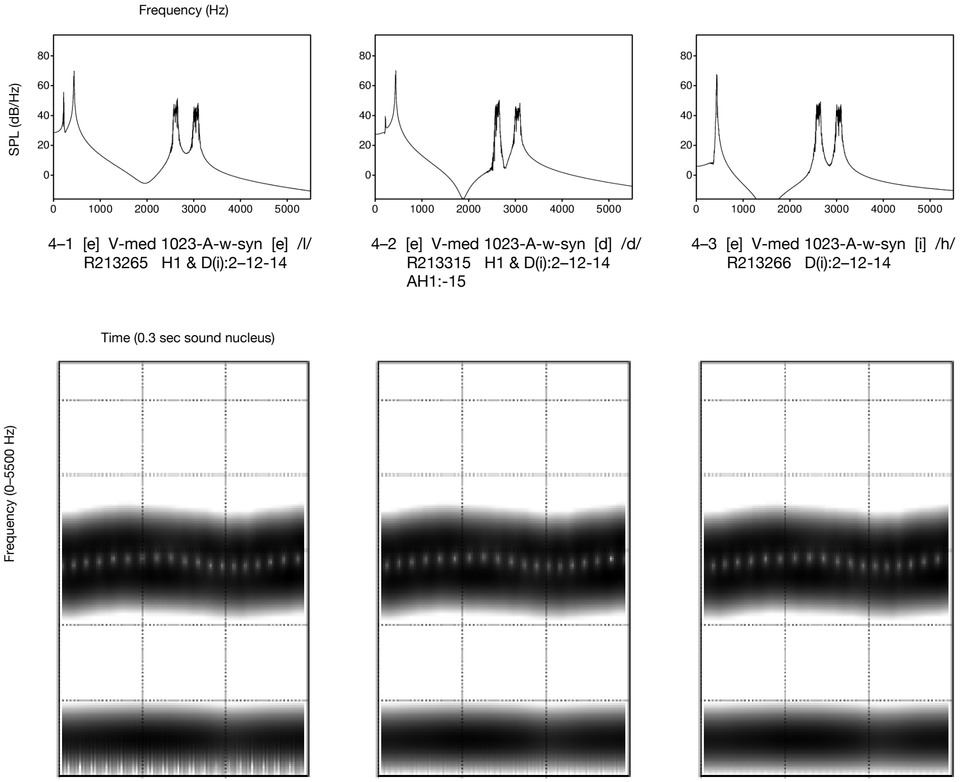
3–6 [ö] V-med 1016-A-w [y] /h/
R213270 D(i):2–8-12

3–7 [o] V-med 1005-A-w [o] /l/
R213275 H1 & D(i):2–4

3–8 [o] V-med 1005-A-w [d] /d/
R213347 H1 & D(i):2–4
AH1:-30

3–9 [o] V-med 1005-A-w [u] /h/
R213276 D(i):2–4

Figure 4. Harmonic synthesis based on nondominant H1 and D-patterns of natural voiced reference sounds, including stepwise H1 attenuation: Illustration of undetectable vowel quality differences in the spectrographic analysis. [C-06-09-F04]



4–1 [e] V-med 1023-A-w-syn [e] /l/ R213265 H1 & D(i):2–12-14

4–2 [e] V-med 1023-A-w-syn [d] /d/ R213315 H1 & D(i):2–12-14 AH1:-15

4–3 [e] V-med 1023-A-w-syn [i] /h/ R213266 D(i):2–12-14

4–4 Spectrogram of the first sound, with H1 & D(i):2–12-14

4–5 Spectrogram of the second, with H1 & D(i):2–12-14 and AH1:-15

4–6 Spectrogram of the third sound, with D(i):2–12-14

6.10 Harmonic Synthesis IV – Replicas Related to Harmonics at or Near Spectral Peaks of Single Natural Vowel Sounds, With Gradual Attenuation of Selected Intermediate Harmonics Causing Double-Vowel and Double-Pitch Recognition

Pursuing the investigation of synthesised sound series with vowel quality shifts related to pitch level shifts, and pursuing harmonic synthesis related to natural reference sounds keeping vowel-related spectral maxima unchanged, we further developed the experimental design of the previous experiment. In a new study, sounds with the same spectral peak structure as described in the previous chapter were investigated, but in contrast to the previous experiment, the entire harmonic spectrum of natural reference sounds was manipulated in order to create sound transitions: The harmonic analysis and subsequent synthesis were conducted producing, firstly, a synthesised replica related to the entire calculated harmonic spectrum of a selected natural reference sound (harmonic resynthesis), secondly, a series of synthesised sounds with stepwise attenuation of the harmonics lying in between the multiple integers of the first dominant harmonic *D*1 and, thirdly, a synthesised sound based on only the harmonics as multiple integers of *D*1. As a further part of the development of the experimental design, speaker groups, vowel qualities and the f_o range of the reference sounds were extended, *D*1 was not limited to *H*2, editing of the calculated harmonic spectra of the natural reference sounds was applied, and one of the listening tests described in the previous chapter was also extended. (For details of the method, see Chapter M6.10 in the Materials.)

According to this developed approach and based on the Zurich Corpus, for each of the five Standard German vowels /e, ø, o/ and /ɛ, ɔ/ and each of the speaker groups of men, women and children, a natural reference sound produced in V context and nonstyle mode at calculated f_o in the range of c. 100–300 Hz was selected fulfilling similar conditions as was the case in the previous experiment: Sounds of the three front vowels were selected for which (i) the three spectral peaks generally assumed to relate to vowel quality were associated with single dominant harmonics *D*1–*D*2–*D*3 near the frequencies of estimated formants, formant estimation being methodologically substantiated, (ii) the first dominant harmonic *D*1 was above the fundamental *H*1, that is, it was *H*2 or higher, (iii) the higher dominant harmonics *D*2 and *D*3 were integer multiples of *D*1. (Vocal effort and sound duration were disregarded.) The same conditions were applied for the selection of sounds of the back vowels, with two modifications: Only the first two spectral peaks generally assumed to relate to vowel quality had to be associated with single dominant harmonics *D*1–*D*2 near the frequencies of estimated formants, and if the second estimated formant related to only a prominent but not a dominant harmonic *D*2 or if the spectral envelope of the lower harmonics showed only one peak associated with a single harmonic *D*1 near the estimated first formant but the course of the subsequent harmonic envelope was continuously sloping, this was also accepted.

As a new aspect of the method, if necessary, single harmonic levels of a few of the selected natural reference sounds were adjusted for synthesis in terms of assigning dB values for attenuating or amplifying the levels of these harmonics in synthesis (spectral editing), with the aim of creating sounds for which the harmonic spectra represented exemplary cases for the experimental conditions in question (for details, see Table 1 in Chapter M6.10).

Although the vowel /ɔ/ is short in Standard German, for three reasons, sounds of that vowel were included in the investigation: (i) /ɔ/ is an open-mid vowel, and the experiment aimed at also producing open-mid–close vowel quality shifts, that is, shifts exceeding adjacent qualities; (ii) the vowel quality distance of the long Standard German vowels /o/ and /a/ is pronounced and when creating the Zurich Corpus, therefore, we also recorded sustained sounds of /ɔ/ produced by some of the speakers investigated (see Chapter 1.1); (iii) the recognition of /ɔ/ as a quality in between /o/ and /a/ was not difficult to develop for the standard listeners of the Zurich Corpus even though the sound duration was long and corresponded to the duration of long vowels.

As a result of the selection process, a sample of 15 natural reference sounds was created, and their *D*1–*D*2–*D*3 or *D*1–*D*2 patterns (frequencies and levels) and, if necessary, the dB values for the adjustments of single levels of the dominant harmonics were assigned. For each single natural reference sound, the dynamic course of its harmonic spectrum for the entire sound duration was analysed using the analysis function of the HarmSyn tool (default parameter setting). Subsequently, based on this analysis and the assigned *D*-pattern and, if necessary, including level adjustments of single harmonics, five replicas applying five attenuation levels of 0/-12/-24/-36/-100 dB for the harmonics that are not integer multiples of the *D*1 frequency were created using the harmonic synthesis function of the HarmSyn tool. Harmonics that are integer multiples of *D*1 frequency were kept unchanged. Thus, for each natural reference sound, a series of synthesised replicas was created, with the first and last sound representing two opposing sounds with two different H_1 and HCF (unchanged harmonics of the reference sound versus multiples of *D*1 only) and with three transitional sounds in between (harmonics as integer multiples of *D*1 unattenuated, all other harmonics stepwise attenuated). In these terms, 15 series of five sounds each were created, and a total of 75 synthesised sounds were investigated in the listening tests. For an illustration of the experimental design, see Figure 1.

The same nine subtests, S1–S9, described in the previous chapter were also conducted for the sounds of the present experiment: Vowel and pitch recognition of the opposing sounds of a series, presented as single sounds or as sound pairs (subtests 1–4), vowel and pitch recognition of all sounds of a series, presented as single sounds or as sound pairs (subtests 5–7 combined with the results of subtests 1–3), and double-vowel and double-pitch recognition of all sounds of a series, presented as single sounds (subtests 8 and 9). (Note that the first subtest was further developed in that additional natural sounds were added to the synthesised sounds of this listening subtest to balance vowel qualities and pitch ranges of the sounds in the task; for details, see Chapter M6.10.)

According to the results of separate vowel and pitch recognition tests for the opposing sounds of a series (subtests 1–4; labelling majority), with few exceptions of single vowel quality identifications of single listeners, no attenuation of the harmonics in between the integer multiples of *D*1 was associated with a lower pitch level and with an open-mid or close-mid vowel quality in correspondence to the natural reference sound, and full attenuation of the harmonics in between the integer multiples of *D*1 was associated with a higher pitch level and a vowel quality shift in an open–close direction when compared with the natural reference sound. Ignoring unrounded–rounded differences and based on subtest 4 (sounds presented as sound pairs), the recognition rate for vowel quality shifts in an open–close direction was 100% for the replicas of all natural close-mid reference sounds and for four of the six natural open-mid reference sounds (including Series 13, with a shift from /ɔ–o/ to /u/). For the remaining two natural open-mid reference sounds, the rate was ≥ 80%. The recognition of upward shifts of pitch levels was uniform among all sounds and listeners.

For the replicas of the natural reference sounds of close-mid vowels /e, ø, o/, the results of separate vowel and pitch recognition tests with stepwise attenuation of the levels of the harmonics not being integer multiples of the *D*1 frequency (subtests 1–3 combined with subtests

5–7) showed consistent transition patterns from close-mid to close vowel qualities and from lower to higher pitch levels. Comparable results were also found for the sounds of the open-mid vowels /ɛ, ɔ/, which showed transitions from open or open-mid to close-mid or close vowel qualities and from lower to higher pitch levels.

For all sounds, the results of testing double-vowel and double-pitch recognition (subtests 8 and 9) showed comparable but more pronounced results than those found for the previous experiment discussed in Chapter 6.9: (i) For each of the sound series investigated, sounds occurred for which two vowel qualities and/or two pitch levels were recognised. Notably, a labelling majority for simultaneous double-vowel and double-pitch recognition for at least one of the transitional replicas was found for all nine series of front vowel sounds and three of the six series of back vowel sounds. Further, in contrast to the previous experiment, simultaneous double-vowel and double-pitch recognition occurred only for the transitional sounds with stepwise attenuation of the harmonics that are not integer multiples of the *D*1 but not for the opposing sounds without and with full attenuation of these harmonics. (ii) Double-pitch recognition without simultaneous double-vowel recognition also occurred. (iii) Again in contrast to the previous experiment, with one exception of a single labelling, double-vowel recognition without simultaneous double-pitch recognition did not occur in the present investigation. Figures 1 and 2 illustrate the main findings of vowel quality shifts in an open–close direction associated with upward shifts of pitch levels for the opposing synthesised replicas (unattenuated harmonics of the natural reference sound versus multiples of *D*1 only) and of the occurrence of double-vowel and double-pitch recognition for the transitional replicas (harmonics as integer multiples of *D*1 unattenuated, all other harmonics stepwise attenuated).

Comparing the individual recognition results of the listeners, for the opposing replicas, listener consensus on vowel recognition proved to be high and recognition of pitch differences proved to be uniform among the listeners. Thus, between-listener differences were marginal. More pronounced between-listener differences occurred for the transitional sounds: The testing of vowel and pitch recognition again revealed distinct recognition profiles for each of the listeners, with similar differences as found in the previous experiment.

As all studies presented in this sixth main chapter and their results showed, the entire line of experimentation on the question of whether vowel recognition relates to pitch (or to a comparable perceptual referencing to a sound pattern repetition over time) produced consistent indications for such a relation. It finally culminated in the evidence of this relation provided by the results of the present study: By manipulating the entire harmonic spectrum of a natural reference vowel sound, two sounds could be produced with equal spectral maxima but with two different recognised vowel qualities associated with two different recognised pitch levels, and the vowel quality shift direction in relation to pitch proved to be consistent for these sounds, that is, rising pitch levels were associated with vowel quality shifts in an open–close direction; in addition, through the above manipulation, it was even possible to produce single sounds for which two vowel qualities and two pitch levels were identified.

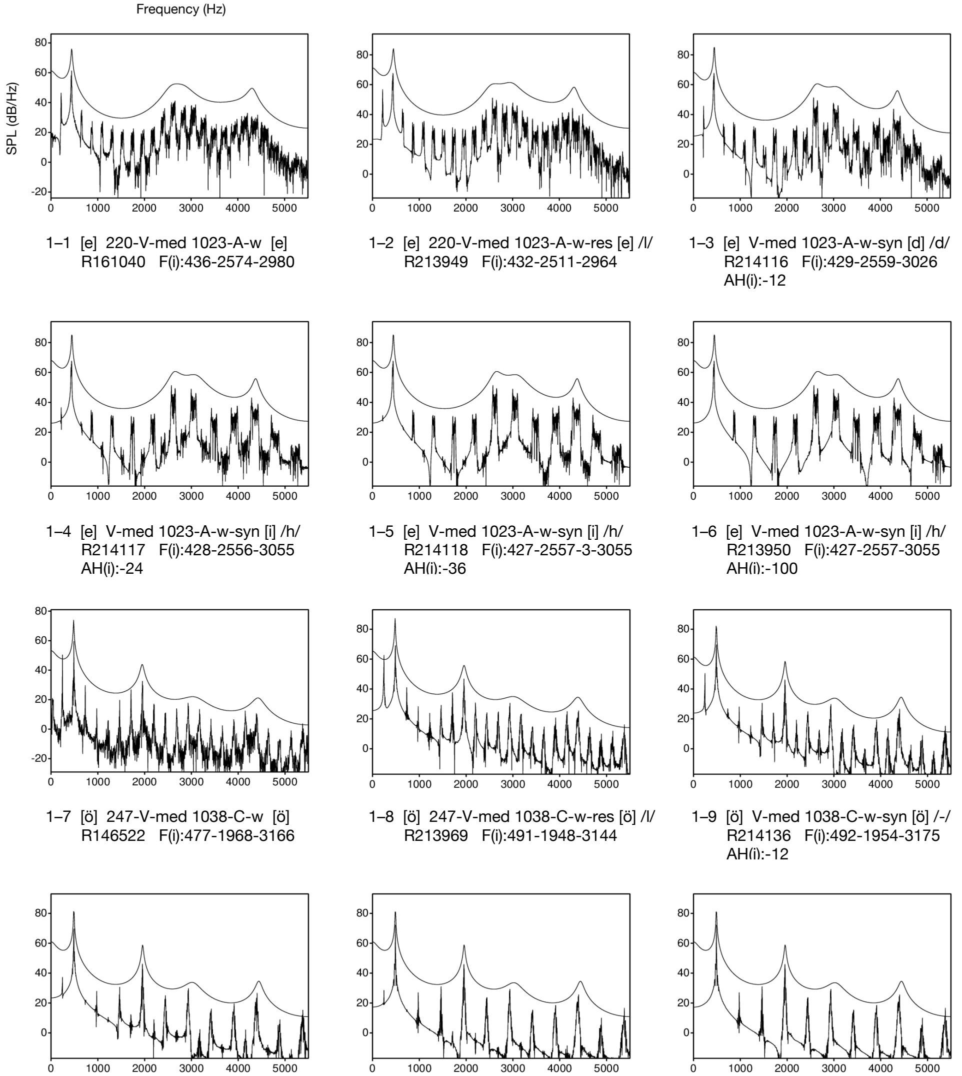
For references, details of experimental design, method and results, an extended discussion including aspects such as listener-specific vowel and pitch recognition, the nonuniform character of vowel and pitch recognition and further remarks with regard to future experiments on the matter, and for the documentation of the sound sample and the results of the investigation (tables including sound links), see the Materials, Chapter M6.10.

For an indication of (re-)synthesis and vowel and pitch level recognition as given in the below figures, including double-vowel and double-pitch recognition, see the Introduction. Note again that, in these figures, the recognition results are given according to Table 3 of Chapter M6.10 (see coloured results in this table). The attenuation of harmonics not being integer multiples of *D*1 frequency is abbreviated as AH(i) and given in dB. A labelling majority for double-vowel and associated double-pitch recognition is given as [d] /d/, and a corresponding labelling minority is given as ([d] /d/). For high-pitched synthesised sounds related to a natural reference sound of /ɛ/ or /ɔ/, vowel recognition results are given as o–c for an open–close direction (for details, see Table 3 of Chapter M6.10).

Figure 1. Harmonic (re-)synthesis based on *D*-patterns of natural voiced reference sounds, including stepwise attenuation of harmonics that are not integer multiples of *D*1: Extensive illustration of the experimental design and the two main findings of vowel quality and pitch level shifts and intermediate double-vowel and double-pitch recognition. Extract of Chapter M6.10, Table 3 (see Series 2, 6, 7, 10 and 13 in this table). For each of the five vowels /e, ø, o, ɛ, ɔ/ investigated, one series of sounds and their spectra are shown in terms of the natural reference sound and all five synthesised replicas (the replica with no attenuation of the harmonics, the three replicas with stepwise attenuation of the levels of the harmonics that are not integer multiples of the *D*1 frequency, and the replica with full attenuation of the harmonics in between the integer multiples of *D*1). According to the vowel and pitch recognition results (labelling majority), for all sound series presented, vowel quality shifts in an open–close direction associated with upward shifts of pitch levels occurred for the opposing synthesised replicas (unchanged harmonics of the natural reference sound versus multiples of *D*1 only), and double-vowel and double-pitch recognition occurred for some of the transitional replicas (harmonics as integer multiples of *D*1 unattenuated, all other harmonics stepwise attenuated).
[C-06-10-F01]

Figure 2. Harmonic (re-)synthesis based on *D*-patterns of natural voiced reference sounds, including stepwise attenuation of harmonics that are not integer multiples of *D*1: Additional reduced illustration of the two main findings. Extract of Chapter M6.10, Table 3 (see Series 1, 3–5, 8, 9, 11, 12, 14 and 15 in this table). For each of the five vowels /e, ø, o, ɛ, ɔ/ and each of the remaining investigated sound series not shown in Figure 1, three synthesised replicas are shown: The replica with no attenuation of the harmonics, a transitional replica with marked double-vowel and double-pitch recognition (for the sound selection, compare with the sound links of the series in the table mentioned), and the replica with full attenuation of the harmonics in between the integer multiples of *D*1. For all series, vowel and pitch shifts and double-vowel and double-pitch recognition correspond to the description given in Figure 1.
[C-06-10-F02]

Figure 1. Harmonic (re-)synthesis based on D-patterns of natural voiced reference sounds, including stepwise attenuation of harmonics that are not integer multiples of D1: Extensive illustration of the experimental design and the two main findings of vowel quality and pitch level shifts and intermediate double-vowel and double-pitch recognition. [C-06-10-F01]



1–1 [e] 220-V-med 1023-A-w [e]
R161040 F(i):436-2574-2980

1–2 [e] 220-V-med 1023-A-w-res [e] /l/
R213949 F(i):432-2511-2964

1–3 [e] V-med 1023-A-w-syn [d] /d/
R214116 F(i):429-2559-3026
AH(i):-12

1–4 [e] V-med 1023-A-w-syn [i] /h/
R214117 F(i):428-2556-3055
AH(i):-24

1–5 [e] V-med 1023-A-w-syn [i] /h/
R214118 F(i):427-2557-3-3055
AH(i):-36

1–6 [e] V-med 1023-A-w-syn [i] /h/
R213950 F(i):427-2557-3055
AH(i):-100

1–7 [ö] 247-V-med 1038-C-w [ö]
R146522 F(i):477-1968-3166

1–8 [ö] 247-V-med 1038-C-w-res [ö] /l/
R213969 F(i):491-1948-3144

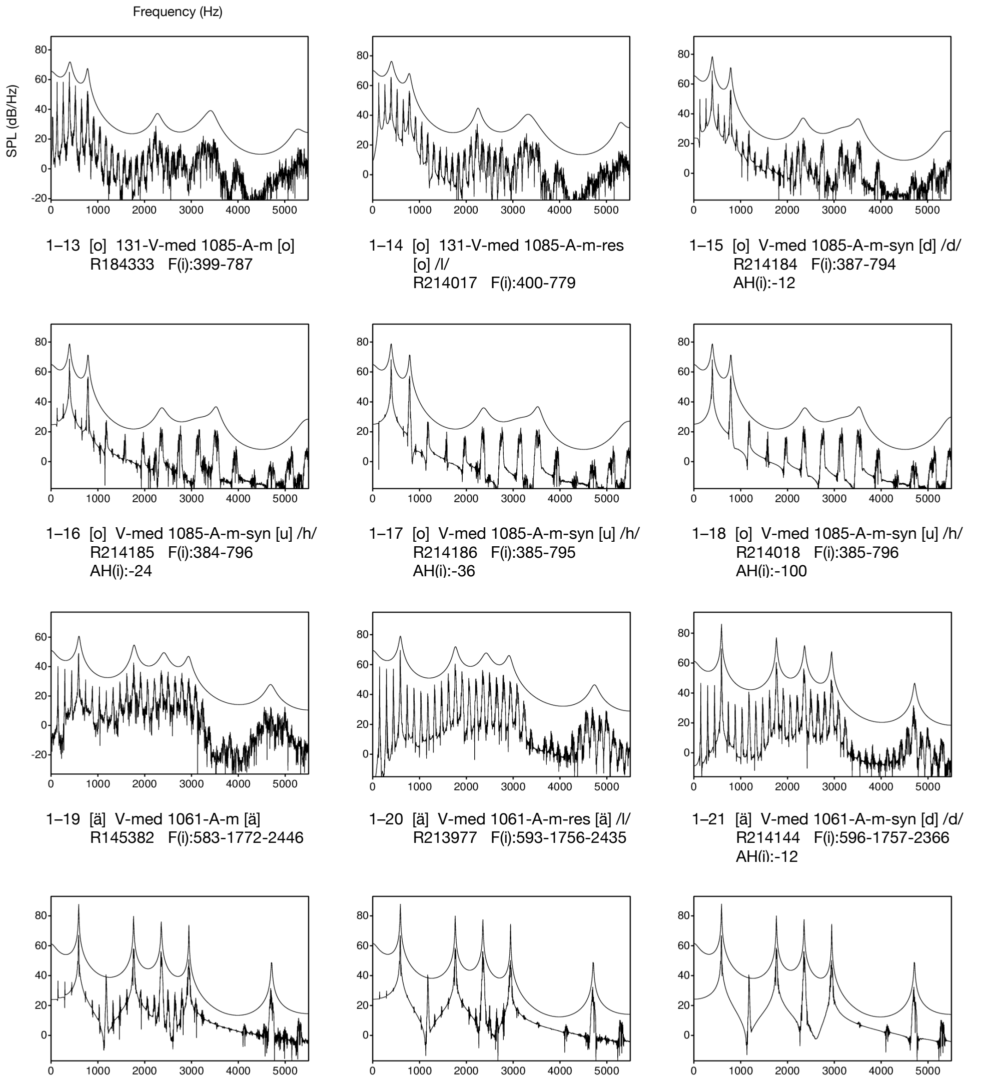
1–9 [ö] V-med 1038-C-w-syn [ö] /-/
R214136 F(i):492-1954-3175
AH(i):-12

1–10 [ö] V-med 1038-C-w-syn [d] /d/
R214137 F(i):492-1956-3181
AH(i):-24

1–11 [ö] V-med 1038-C-w-syn [ü] /h/
R214138 F(i):492-1957-3185
AH(i):-36

1–12 [ö] V-med 1038-C-w-syn [ü] /h/
R213970 F(i):492-1957-3185
AH(i):-100

Figure 1 (continuation). [C-06-10-F01]



1–13 [o] 131-V-med 1085-A-m [o]
R184333 F(i):399-787

1–14 [o] 131-V-med 1085-A-m-res
[o] /l/
R214017 F(i):400-779

1–15 [o] V-med 1085-A-m-syn [d] /d/
R214184 F(i):387-794
AH(i):-12

1–16 [o] V-med 1085-A-m-syn [u] /h/
R214185 F(i):384-796
AH(i):-24

1–17 [o] V-med 1085-A-m-syn [u] /h/
R214186 F(i):385-795
AH(i):-36

1–18 [o] V-med 1085-A-m-syn [u] /h/
R214018 F(i):385-796
AH(i):-100

1–19 [ä] V-med 1061-A-m [ä]
R145382 F(i):583-1772-2446

1–20 [ä] V-med 1061-A-m-res [ä] /l/
R213977 F(i):593-1756-2435

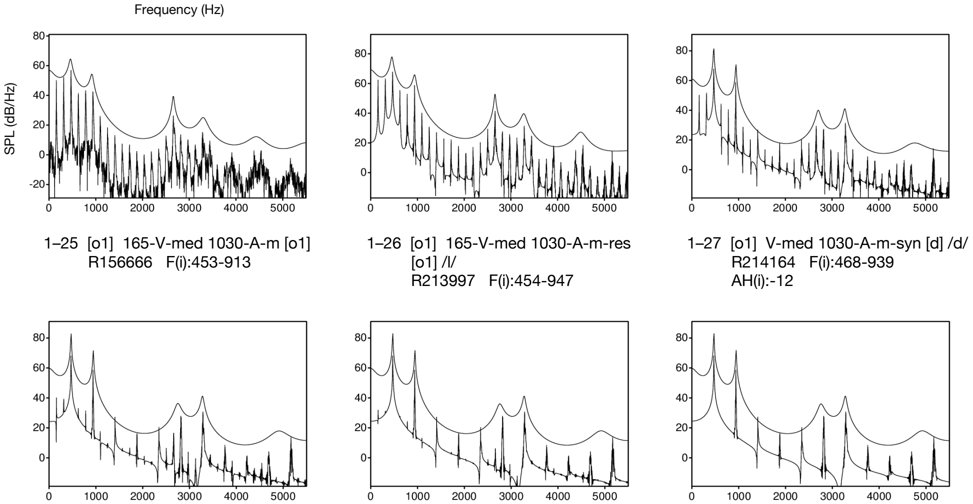
1–21 [ä] V-med 1061-A-m-syn [d] /d/
R214144 F(i):596-1757-2366
AH(i):-12

1–22 [ä] V-med 1061-A-m-syn [d] /d/
R214145 F(i):600-1759-2358
AH(i):-24

1–23 [ä] V-med 1061-A-m-syn [o-c] /h/
R214146 F(i):600-1758-2358
AH(i):-36

1–24 [ä] V-med 1061-A-m-syn [o-c] /h/
R213978 F(i):600-1759-2358
AH(i):-100

Figure 1 (continuation). [C-06-10-F01]



1–25 [o1] 165-V-med 1030-A-m [o1]
R156666 F(i):453-913

1–26 [o1] 165-V-med 1030-A-m-res
[o1] /l/
R213997 F(i):454-947

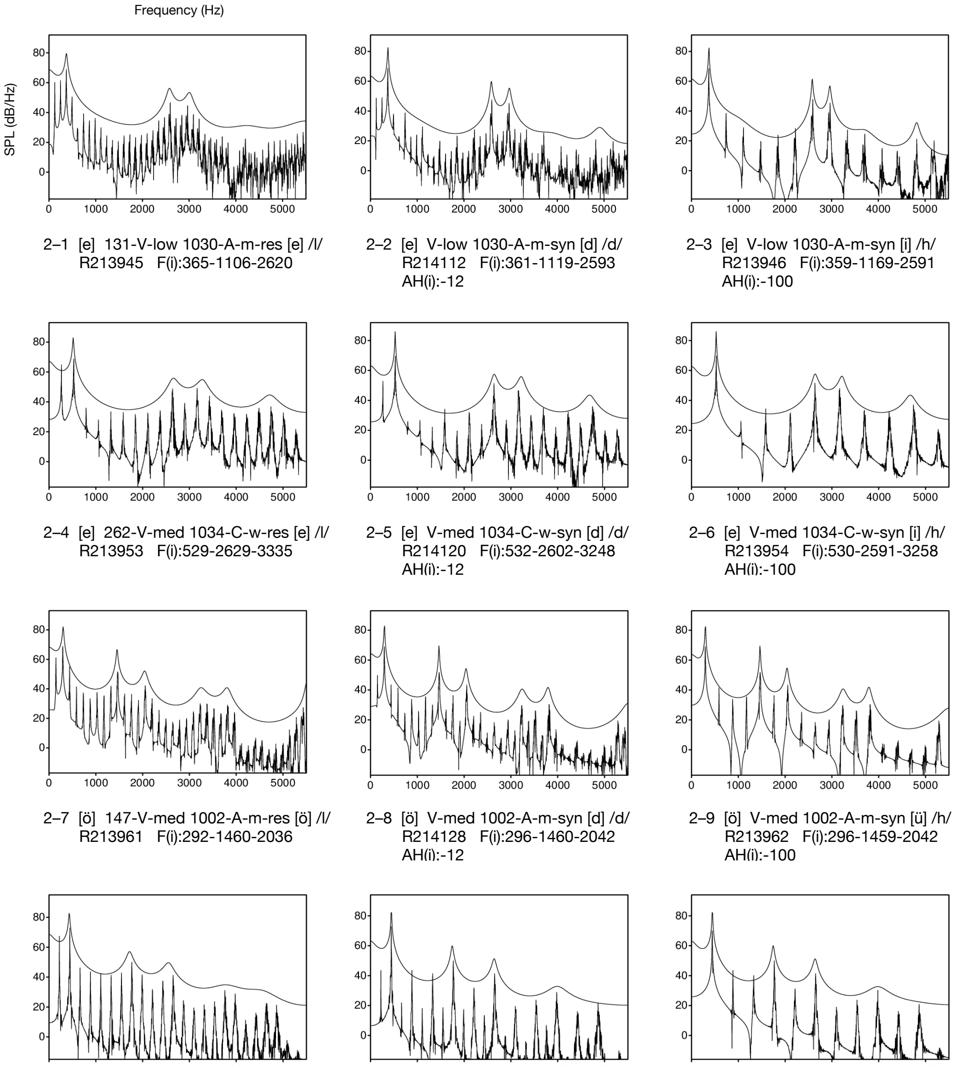
1–27 [o1] V-med 1030-A-m-syn [d] /d/
R214164 F(i):468-939
AH(i):-12

1–28 [o1] V-med 1030-A-m-syn [d] /d/
R214165 F(i):472-922
AH(i):-24

1–29 [o1] V-med 1030-A-m-syn [d] /d/
R214166 F(i):473-924
AH(i):-36

1–30 [o1] V-med 1030-A-m-syn [u] /l/
R213998 F(i):470-912
AH(i):-100

Figure 2. Harmonic (re-)synthesis based on D-patterns of natural voiced reference sounds, including stepwise attenuation of harmonics that are not integer multiples of D1: Additional reduced illustration of the two main findings. [C-06-10-F02]



2–1 [e] 131-V-low 1030-A-m-res [e] /l/
R213945 F(i):365-1106-2620

2–2 [e] V-low 1030-A-m-syn [d] /d/
R214112 F(i):361-1119-2593
AH(i):-12

2–3 [e] V-low 1030-A-m-syn [i] /h/
R213946 F(i):359-1169-2591
AH(i):-100

2–4 [e] 262-V-med 1034-C-w-res [e] /l/
R213953 F(i):529-2629-3335

2–5 [e] V-med 1034-C-w-syn [d] /d/
R214120 F(i):532-2602-3248
AH(i):-12

2–6 [e] V-med 1034-C-w-syn [i] /h/
R213954 F(i):530-2591-3258
AH(i):-100

2–7 [ö] 147-V-med 1002-A-m-res [ö] /l/
R213961 F(i):292-1460-2036

2–8 [ö] V-med 1002-A-m-syn [d] /d/
R214128 F(i):296-1460-2042
AH(i):-12

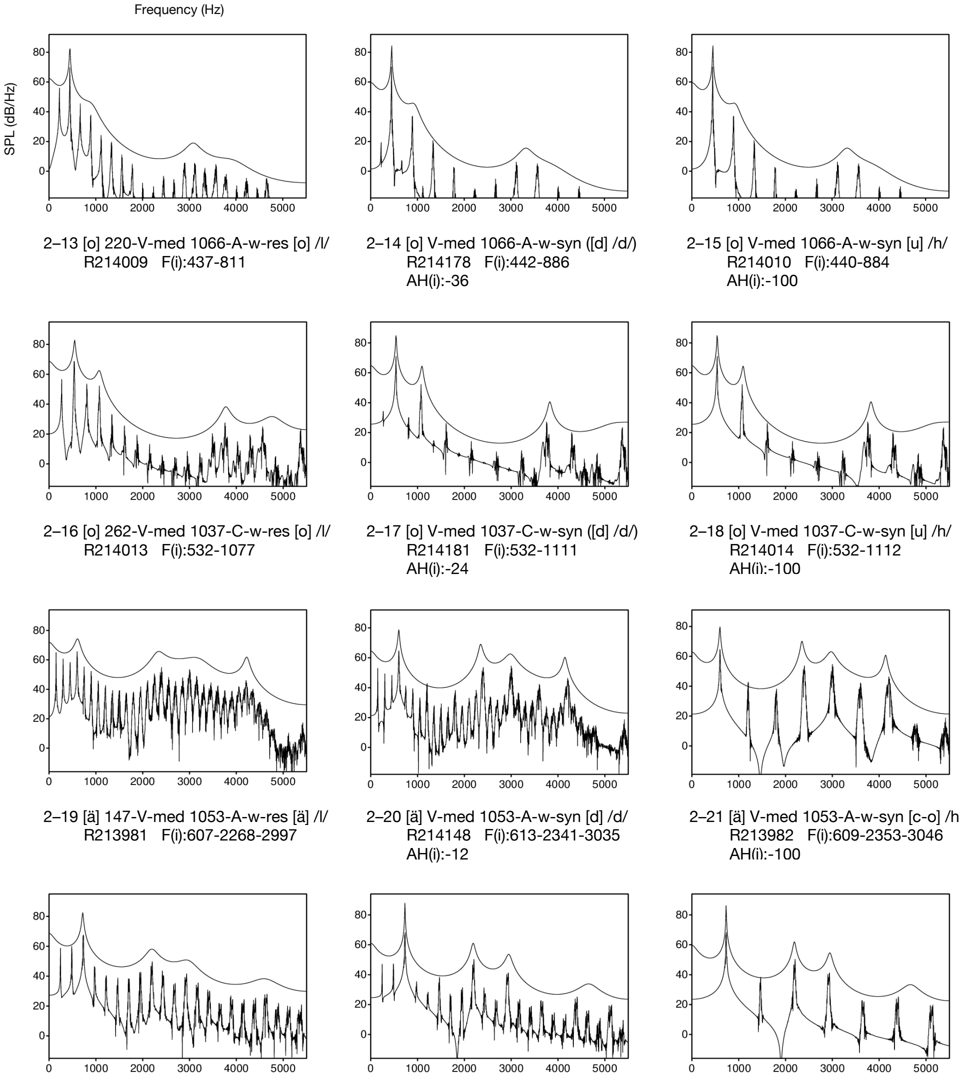
2–9 [ö] V-med 1002-A-m-syn [ü] /h/
R213962 F(i):296-1459-2042
AH(i):-100

2–10 [ö] 220-V-med 1016-A-w-res [ö] /l/
R213965 F(i):446-1723-2628

2–11 [ö] V-med 1016-A-w-syn [d] /d/
R214133 F(i):451-1739-2687
AH(i):-24

2–12 [ö] V-med 1016-A-w-syn [ü] /h/
R213966 F(i):451-1741-2692
AH(i):-100

Figure 2 (continuation). [C-06-10-F02]



2–13 [o] 220-V-med 1066-A-w-res [o] /l/
R214009 F(i):437-811

2–14 [o] V-med 1066-A-w-syn ([d] /d/)
R214178 F(i):442-886
AH(i):-36

2–15 [o] V-med 1066-A-w-syn [u] /h/
R214010 F(i):440-884
AH(i):-100

2–16 [o] 262-V-med 1037-C-w-res [o] /l/
R214013 F(i):532-1077

2–17 [o] V-med 1037-C-w-syn ([d] /d/)
R214181 F(i):532-1111
AH(i):-24

2–18 [o] V-med 1037-C-w-syn [u] /h/
R214014 F(i):532-1112
AH(i):-100

2–19 [ä] 147-V-med 1053-A-w-res [ä] /l/
R213981 F(i):607-2268-2997

2–20 [ä] V-med 1053-A-w-syn [d] /d/
R214148 F(i):613-2341-3035
AH(i):-12

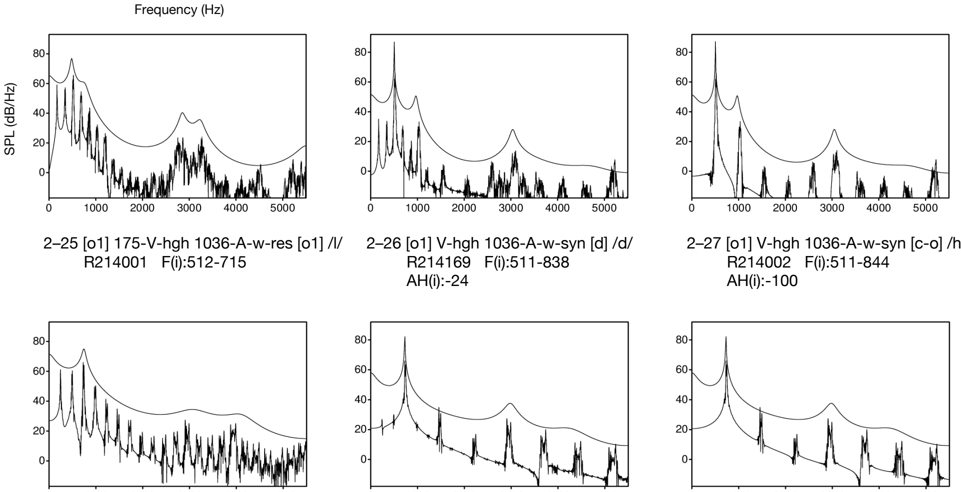
2–21 [ä] V-med 1053-A-w-syn [c-o] /h/
R213982 F(i):609-2353-3046
AH(i):-100

2–22 [ä] 247-V-med 1034-C-w-res [ä] /l/
R213973 F(i):720-2158-3009

2–23 [ä] V-med 1034-C-w-syn [d] /d/
R214140 F(i):729-2159-2972
AH(i):-12

2–24 [ä] V-med 1034-C-w-syn [c-o] /h/
R213974 F(i):729-2161-2972
AH(i):-100

Figure 2 (continuation). [C-06-10-F02]



2–25 [o1] 175-V-hgh 1036-A-w-res [o1] /l/
R214001 F(i):512-715

2–26 [o1] V-hgh 1036-A-w-syn [d] /d/
R214169 F(i):511-838
AH(i):-24

2–27 [o1] V-hgh 1036-A-w-syn [c-o] /h/
R214002 F(i):511-844
AH(i):-100

2–28 [o1] 247-V-med 1054-C-m-res [o1] /l/
R213993 F(i):736-(–)

2–29 [o1] V-med 1054-C-m-syn ([d] /d/)
R214162 F(i):737-1558
AH(i):-36

2–30 [o1] V-med 1054-C-m-syn [c-o] /h/
R213994 F(i):737-1558
AH(i):-100

6.11 Conclusion

Before reaching a final conclusion regarding a general indication of a perceptual and acoustic vowel–pitch relation (or its alternative), the results of the experiments described in this sixth main chapter shall be summarised. These experiments represent an attempt to approach vowel recognition in direct relation to pitch recognition, pitch not or not unambiguously being equal to measured fundamental frequency or spectral characteristics such as *H*1 and/or HCF often equalled with fundamental frequency.

The vowel–pitch relation question was first investigated and discussed with regard to a comparison of natural whispered and voiced sounds and of corresponding synthesised replicas (Chapters 6.1 to 6.3). Given the limitations of a generalisation of the corresponding results (see the discussion section in Chapters M6.1 and M6.3) and formulated as a rough estimation, the investigated sound comparisons and their results indicated that the pitch level of whispered sounds (including whispered-like synthesised replicas) was mostly recognised within an f_o/pitch level range of voiced sounds of c. 200–400 Hz, that is, the pitch level of whispered sounds was mostly recognised as higher than the level of voiced sounds produced at f_o below c. 200 Hz and as lower than the level of voiced sounds produced at f_o of c. 400 Hz or above. Since, according to the definitions given in the literature, whispered sounds do not have a fundamental frequency, this general indication supported the thesis of vowel recognition as being related to pitch independently of whether or not an f_o level can be calculated for the sounds in question. (Note that some additional indications of vowel and pitch recognition differences were also found, which relate to age and gender differences of the speakers and the comparison of natural and/or synthesised sounds.) Because of these indications, we have suggested that whispered vowel sounds may be perceived as having a pitch comparable to voiced sounds, and thus the link between spectral characteristics and related vowel recognition of whispered and voiced sounds may be pitch.

The vowel–pitch relation was then investigated and discussed with regard to the "missing fundamental" phenomenon (Chapter 6.4). It was argued that suppression of *H*1 or *H*1–*H*2 of natural vowel sounds did not affect pitch recognition as such and that occurring vowel quality shifts for sounds with suppressed *H*1 or *H*1–*H*2 were not due to pitch level shifts but to HP sound filtering. It was the aim of the investigation of the "missing fundamental" phenomenon to document that vowel recognition does not systematically relate to *H*1. Since *H*1 is often equated with fundamental frequency, the demonstration that the pitch level of vowel sounds does not relate to *H*1 is of major importance.

In a third step, the vowel–pitch relation was investigated and discussed with regard to so-called sinewave vowel sounds (three-sinusoid sound synthesis, the sinusoids relating to estimated statistical *F*-patterns or estimated *F*-patterns of single natural vowel sounds), their spectra not manifesting a harmonicity of the sinusoids (Chapters 6.5 and 6.6). Vowel recognition of sounds of this type was again indicated to relate to pitch: In most cases, vowel qualities or vowel openness of synthesised sounds based on estimated *F*-patterns of close vowels were recognised according to vowel intention, with the pitch of these sounds being recognised at lower levels than the levels of the close-mid and open-mid vowel sounds investigated. In contrast, vowel qualities of sounds related to *F*-patterns of close-mid and open-mid vowels were confused in terms of quality shifts in an open–close direction when compared with vowel intention, and the pitch of these sounds was recognised at middle or higher levels. Thus, the findings for close, close-mid and open-mid vowels supported the vowel–pitch relation thesis for sounds without an HCF in their spectra that could be compared with an HCF of natural vowel sounds. At the same time, the findings demonstrated that spectral maxima do not relate to vowel recognition per se, independently of pitch. Exceptions were the findings for the sounds of /a/, which were mostly recognised in the /a-ɔ/ range despite comparatively high first sinusoid frequencies and higher recognised pitch levels, once again indicating a nonuniform character of spectral representation of vowel quality.

At this stage of investigation, the vowel–pitch relation thesis (or its alternative) was supported for sounds with no fundamental frequency (whispered vowel sounds) and for sounds with no HCF (sinewave vowel sounds), and it was argued that pitch level could not, in general, be equated with *H*1 frequency.

In a fourth step, the vowel–pitch relation was investigated and discussed with regard to harmonic sinewave synthesis based on “incomplete” harmonic series: Synthesised sounds were investigated which related to three sinusoids or a single lower sinusoid and a frequency band of higher sinusoids equal in amplitude (Chapter 6.7), all sinusoid configurations possessing harmonicity. In short, a shift in recognised vowel quality could be triggered either by a change in S_1 without a change in HCF or, inversely, by a change in HCF without a change in S_1. Further, and most importantly, if a change in HCF triggered a vowel quality shift, it was associated with a pitch level shift. (However, this association was somewhat weak for the first of the two experiments conducted.) The investigation aimed to document two causes of vowel quality shifts: a change of a relative spectral maximum or a change of the sound periodicity expressed in HCF of the harmonics in the spectrum.

During the pilot synthesis studies based on sinusoids and during the conducting of the vowel synthesis experiments discussed in Chapters 6.5 to 6.7, the listeners involved in the recognition tasks repeatedly reported two experiences: They perceived many of the sounds as highly artificial, and because of the very specific attention given to sound characteristics during the tests, they could recognise two vowel qualities and/or two (or even more) pitch levels for numerous cases of sounds. Therefore, in a fifth step, the investigation of sinewave-like vowel sounds was resumed, developed and extended (Chapter 6.8): Harmonic synthesis was conducted based on three extracted dominant harmonics and their dynamic contour of natural reference sounds, the frequencies of the selected harmonics being near the estimated formant frequencies of the reference sounds. In addition, for each vowel quality investigated and for each of the three speaker groups of men, women and children, three sounds were investigated to include spectral variation in the examination. Finally, vowel and pitch recognition were not only tested for dominant or prominent vowel qualities and/or pitch levels but also for simultaneously occurring secondary qualities and/or secondary levels. Concerning the recognition of dominant or prominent vowel qualities and pitch levels, the results obtained were in line with the above results for sinewave vowel sounds based on sinusoids, the recognised pitch level in most cases being associated with the frequency of the first harmonic of synthesis. Concerning the recognition of secondary vowel qualities and pitch levels, numerous sounds were recognised by single listeners as having two vowel qualities and two pitch levels or two pitch levels only (double-vowel-only recognition was very rare). This extended investigation of sinewave-like vowel sounds aimed to document the vowel–pitch relation for synthesised sounds with more natural sound quality and with HCF of the three harmonics used in synthesis, to crosscheck the vowel–pitch relation for different spectral maxima configurations of sounds of a given vowel and to include (and introduce) the examination of double-vowel and/or double-pitch recognition into the experimental design.

In a sixth and final step, an attempt was made to oppose two sounds with equal spectral maxima but different recognised vowel qualities and pitch levels because of different HCF and, at the same time, also to investigate the transitional sounds in between them (Chapters 6.9

and 6.10): If, based on the harmonic spectrum of a natural reference sound and its recognised vowel quality and pitch level, a spectral manipulation is made so as to create a second sound with unchanged spectral maxima but with different HCF, recognised vowel quality and pitch level, it can be expected that, conducting the spectral manipulation step by step, sounds with double-vowel and/or double-pitch may occur. As the last two experiments showed, this could indeed be demonstrated for sound synthesis with stepwise attenuation of either low harmonics that are not part of a *D*-pattern of a natural reference sound or of harmonics that are not integer multiples of its first dominant harmonic (with HCF frequency of all dominant harmonics being above *H*1 frequency). Thus, in addition to the comparison of two opposing sounds with different recognised pitch levels, the vowel–pitch relation could also be demonstrated for single sounds with double-vowel and double-pitch recognition, strongly underpinning the vowel–pitch relation thesis: Single sounds for which two vowel qualities and two pitch levels can be recognised and for which shift directions are consistent and predictable – rising pitch level associated with vowel quality shift in an open–close direction – would not occur if the perceptual process did not relate vowel recognition to pitch (or its alternative; see below).

In these terms, we conclude that vowel recognition relates to pitch (or its alternative), independently of whether or not f_o, *H*1 or HCF can be assessed. Thus, the observation of a relation of the vowel spectrum of natural sounds to f_o of sound production turns out to be an indication of the vowel–pitch relation (or its alternative). Evidently, this calls for a change in the paradigm of our understanding of the human voice and the vowel sound, and with this, of our understanding of speech.

The vowel–pitch relation (or its alternative) and, with it, occurring cases of sounds with double-vowel and/or double-pitch recognition represent a core phenomenon of vowel acoustics and recognition. The demonstration thereof is one of the most important achievements of this treatise. Hence, sounds with double-vowel and double-pitch recognition are of primary importance because evidence for the relation in question could be given for single sounds with single acoustic sound characteristics. Consequently, the relation was demonstrated to result from a primarily perceptual process of vowel recognition. At the same time, in a definitive manner, the phenomenon contradicts the thesis of spectral maxima or filter curves being vowel quality-specific *per se.*

The demonstration of the vowel–pitch relation proved to be dependent on the type of synthesis applied and the resulting sound quality and on the investigation of vowel qualities, pitch levels and ranges, individual spectral energy configurations for sound production of a vowel, listening test conditions and recognition strategies of individual listeners. Further, the fact that there were numerous cases of double-pitch recognition only and rare cases of double-vowel recognition only needs to be considered when interpreting the results. (For a detailed discussion, see Chapters M6.9 and M6.10).

The occurrence of rare cases of vowel quality shifts without pitch shifts are difficult to interpret: On the one hand, as argued in Chapter M6.9, these findings may be a result that is to be expected considering the specific timbre of some of the sounds investigated and also considering the kind of recognition tests performed. Concerning sound timbre, no vowel quality shifts without pitch level shifts occurred in the experiments discussed in Chapter 6.10 (disregarding one single labelling), with the synthesis in this experiment resulting in more natural-like vowel sounds when compared to the synthesis of the other experiments of this main chapter. Concerning recognition tests, except for one experiment (see Chapter 6.8), cases of vowel quality shifts without pitch level shifts occurred (i) only for separately testing double-vowel and double-pitch recognition, (ii) only for some of the listeners, and (iii) associated with listener-specific labelling inconsistencies. In future studies, recognition tests may be developed further, including simultaneous double-vowel and double-pitch recognition and relating the two recognised vowel qualities and two pitch levels to one another to clarify the matter. On the other hand, the findings may indicate that eliciting vowel and pitch recognition through verbal instruction by the interviewer and verbal response by the listener does not allow for uncovering all basic aspects of the actual analytical recognition process (see below). Therefore, we cautiously conclude that vowel recognition relates to pitch or a comparable perceptual referencing to a sound pattern repetition over time. (Note in this context that we could not replicate the claim of Robinson and Patterson, 1995, that a single period of a vowel sound is sufficient for vowel recognition: According to the author's estimate, single periods extracted from the middle of the opposing sounds of the series investigated in Chapter 6.10 do not allow for the recognition of the vowel qualities of the reference sounds and the related open–close quality shifts; however the repetition of these single periods over a time of 1 sec. do allow for vowel quality and quality shift recognition.)

To elaborate further on the perceptual character of the recognition experiments conducted, vowel and pitch recognition are related to listener-specific sound perception and recognition strategies. Moreover, recognition experiments for sounds produced with extensively varying production parameters, as discussed here – including sound synthesis and sound manipulations – require special skills and familiarisation on the part of the listeners. But even for the trained singers, actresses and actors involved in the reported listening tests, substantial between-listener differences with regard to recognition results were sometimes observed. This finding has to be expected for the kind of investigation discussed here: As for vowel quality recognition, listeners differ above all in their mapping of vowel boundaries, and as for pitch recognition, listeners employ individual analytic strategies that have a strong impact on sounds for which sound timbre, vowel quality and pitch concur and sometimes compete. In addition, possible musical interval mismatches and the context of sound presentation in the listening tests also influence the recognition results.

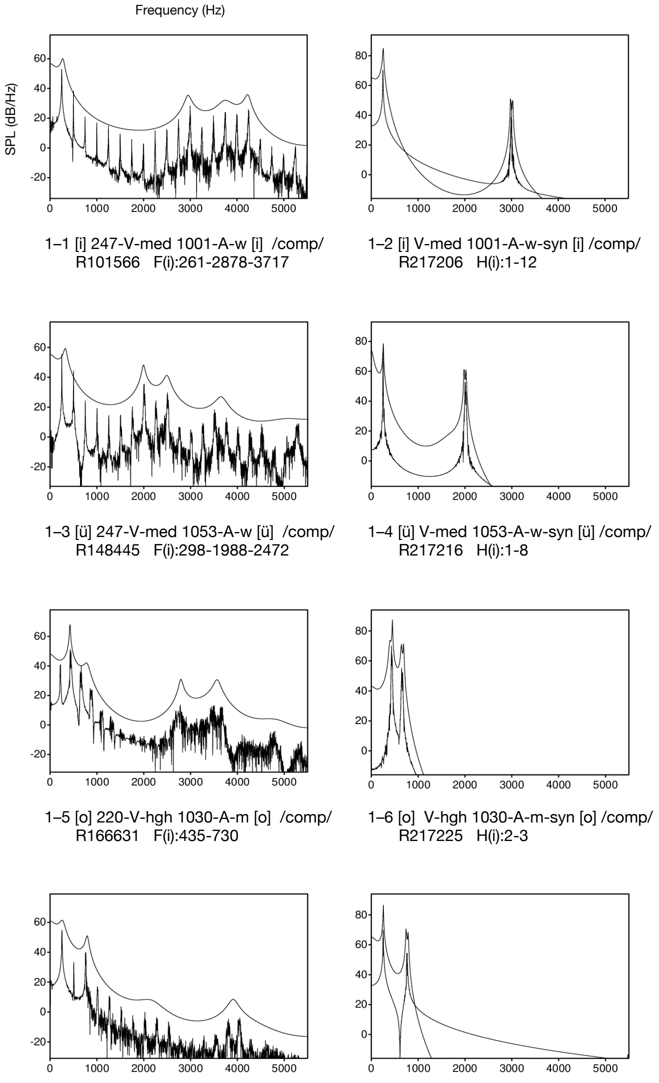
Indeed, the use of the term "recognition" to describe a verbal statement of a listener answering a test task is itself to be questioned. To give an example, some listeners may be able to verbally assign parallel vowel quality and pitch level shifts for many sounds presented in the reported experiments, while others may only be able to verbally assign vowel quality shifts without being able to assign a pitch level difference between the sounds verbally. However, this may not be due to a lack of a general vowel–pitch relation for all listeners, but it may be due to the difficulty of "conscious" analytic pitch level differentiation and corresponding verbal denotation. Therefore, it is questionable whether the actual structure of the perceptual process of vowel quality recognition is always detectable in recognition tasks involving verbal pitch-level assignments. In this sense, the methodological question of investigating interrelated vowel and pitch perception and recognition is in itself a matter of future research. (Note in this context that so-called poor-pitch perception is not reported as being directly associated with vowel confusion.)

However, in our understanding, all these aspects do not counter a general vowel–pitch relation thesis (or its alternative).

Besides the results for associated vowel and pitch recognition, the fact that highly recognisable vowel sounds with a natural-like quality can be synthesised based on only two, three or four extracted harmonics and their dynamic course (see the experiments discussed in Chapters 6.8 and 6.9) is remarkable. Notably, synthesised sounds of this type provide evidence that the spectral shape in terms of an estimated spectral envelope, including the fine structure of the vowel spectrum, is by no means an acoustic representation of vowel quality that could substitute the respective *F*-pattern: There is no spectral fine structure in synthesised sounds based on two, three or four extracted harmonics of a natural reference sound, and no common spectral shape concept accounts for the synthesised sounds. In addition to the sounds documented in Chapters 6.8 and 6.9 (and the corresponding chapters in the Materials), Figure 1 illustrates this phenomenon by presenting sounds that were synthesised based on two or three harmonics of natural reference sounds: For these sounds, not only the intended vowel quality but also the intended pitch level is maintained in synthesis (author's estimate; occurring double-vowel and/or double-pitch recognition disregarded). In this context, special attention should be given to the surprisingly good sound timbre produced by the harmonic synthesiser HarmSyn, even if synthesis is based on only two or three harmonics of a natural reference sound.

Figure 1. Recognisable synthesised sounds of the vowels /i, y, o, u/ and /e, ø, ɛ, a/, synthesis based on two or three harmonics of natural reference sounds, with both intended vowel quality and intended pitch level maintained (author's estimate). Sounds 1–8 = four sound pairs of the vowels /i, y, o, u/. For each pair, a natural reference sound and a related sound that was synthesised based on two harmonics of the reference sound are shown. Sounds 9–16 = four comparable sound pairs of the vowels /e, ø, ɛ, a/, sound synthesis based on three harmonics.
[C-06-11-F01] ↗

Figure 1. Recognisable synthesised sounds of the vowels /i, y, o, u/ and /e, ø, ɛ, a/, synthesis based on two or three harmonics of natural reference sounds, with both intended vowel quality and intended pitch level maintained (author's estimate). [C-06-11-F01]



1–1 [i] 247-V-med 1001-A-w [i] /comp/
R101566 F(i):261-2878-3717

1–2 [i] V-med 1001-A-w-syn [i] /comp/
R217206 H(i):1-12

1–3 [ü] 247-V-med 1053-A-w [ü] /comp/
R148445 F(i):298-1988-2472

1–4 [ü] V-med 1053-A-w-syn [ü] /comp/
R217216 H(i):1-8

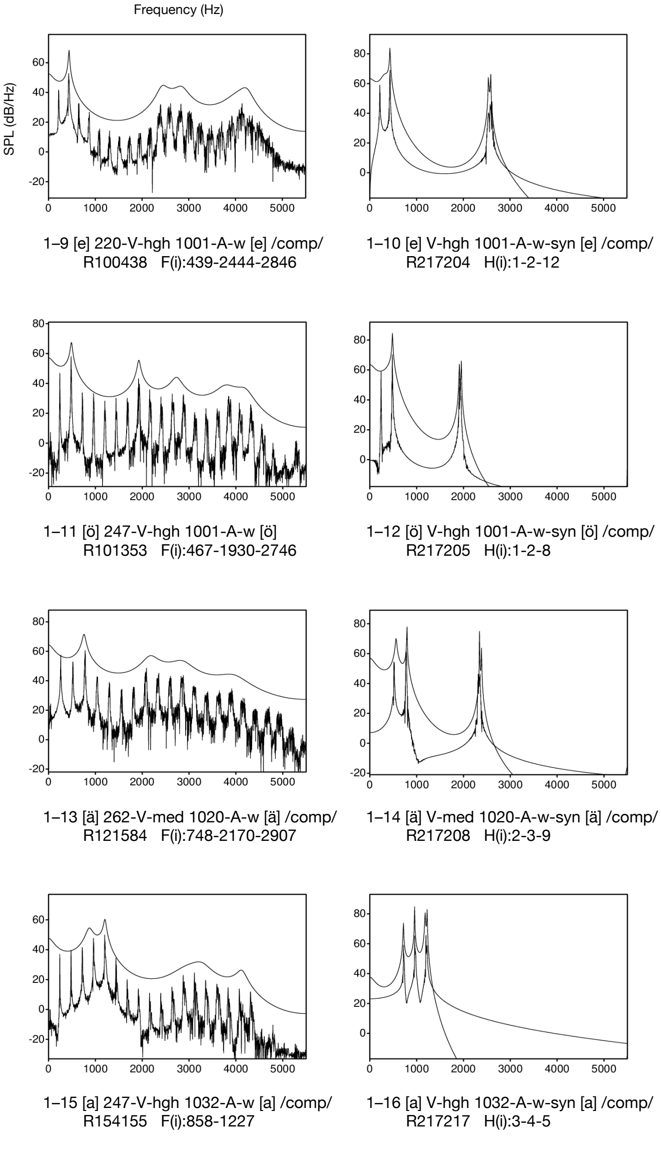
1–5 [o] 220-V-hgh 1030-A-m [o] /comp/
R166631 F(i):435-730

1–6 [o] V-hgh 1030-A-m-syn [o] /comp/
R217225 H(i):2-3

1–7 [u] 247-V-med 1027-A-w [u] /comp/
R118444 F(i):260-791

1–8 [u] V-med 1027-A-w-syn [u] /comp/
R217207 H(i):1-3

Figure 1 (continuation). [C-06-11-F01]



1–9 [e] 220-V-hgh 1001-A-w [e] /comp/
R100438 F(i):439-2444-2846

1–10 [e] V-hgh 1001-A-w-syn [e] /comp/
R217204 H(i):1-2-12

1–11 [ö] 247-V-hgh 1001-A-w [ö]
R101353 F(i):467-1930-2746

1–12 [ö] V-hgh 1001-A-w-syn [ö] /comp/
R217205 H(i):1-2-8

1–13 [ä] 262-V-med 1020-A-w [ä] /comp/
R121584 F(i):748-2170-2907

1–14 [ä] V-med 1020-A-w-syn [ä] /comp/
R217208 H(i):2-3-9

1–15 [a] 247-V-hgh 1032-A-w [a] /comp/
R154155 F(i):858-1227

1–16 [a] V-hgh 1032-A-w-syn [a] /comp/
R217217 H(i):3-4-5

7 Spectral Variation of Vowel Sounds and its Nonuniform Character – Broadening the Documentation of the Variation Extent

7.1 Different Vowel-Related Spectral Peak Numbers

As was discussed when describing the relation of the lower vowel spectrum to f_o and the resulting formant pattern and spectral shape ambiguity (see Chapters 2 and 3), and as was also indicated in the documentation of vocal effort-related spectral variation (see Chapter 5.4), the vowel spectrum exhibits a nonuniform character concerning these three aspects. Earlier, we have also discussed further aspects of nonuniform spectral variation observed for natural sounds of a given vowel in vowel-related frequency ranges that obstruct the formulation of a general concept of relating recognised vowel quality to a specific spectral peak pattern or an average spectral shape. All these aspects are taken up, completed, discussed and documented anew in the following chapters based on the Zurich Corpus, describing the nonuniform relation of recognised vowel quality and the vowel spectrum concerning:

- The inconstant number of vowel-related spectral peaks
- The occurrence of inversions of vowel-related relative spectral energy maxima and minima
- The occurrence of flat or sloping vowel-related spectral energy distribution
- f_o variation < 250 Hz not affecting *F*-patterns and spectral envelopes
- The relation of the lower vowel spectrum to f_o being different for different vowel qualities
- The fine structure of spectral energy distribution having an impact on the relation of the lower vowel spectrum to f_o, without and with vocal effort variation

This first chapter addresses the question of an inconstant number of spectral peaks for sounds of a given vowel in their vowel-related frequency ranges.

It is well known that sounds of back vowels can manifest only one spectral peak < 1–1.5 kHz instead of the expected two peaks. This phenomenon is generally understood as being a consequence of two formants close in frequency (formant merger). Comparably, attempts were made to relate the *F*-patterns of synthesised sounds of front vowels that are based on either two or three formants, with *F*2 of the two-formant sounds (often termed *F*2-prime) being in between *F*2–*F*3 of the three-formant sounds. Both phenomena are discussed within the concept of a spectral "centre of gravity" effect in terms of an auditory spectral averaging process. Further, according to the literature, in some sound spectra of some speakers, an additional spectral envelope peak may occur between the expected first and second or second and third formant, a manifestation generally understood as spurious formants unrelated to vowel quality. Finally, as discussed in Chapter 5.1, the spectra of vowel sounds produced with breathy phonation may manifest an increased amplitude of the first harmonic, which sometimes shows the highest energy level in the spectrum.

However, these types of spectral peak patterns that deviate from the general assumption of vowel-specific peak numbers were barely investigated systematically, including a variation of basic production parameters, and our own earlier investigations did not support a general explanation of different formant numbers of vowel sounds as being a phenomenon of formant merging or *F*2-prime or spurious formants. Therefore, and to again document the possible variability of the vowel spectrum based on the newly compiled Zurich Corpus and to embed it into the line of argument of this treatise, a corresponding study was conducted addressing two questions: What is the variation of spectral peak numbers that can be observed for natural vowel sounds, including different phonation types and different levels of f_o for the voiced sounds? Do sound spectra manifesting "unexpected" spectral peak numbers generally comply with the concepts of formant merging, *F*2-prime or spurious formants?

Based on the inspection of the corpus, for each of the eight long Standard German vowels, two series of sounds produced by different speakers of different speaker groups (children, women and men) were compiled for which the spectra manifested a varying number of vowel-related spectral peaks: One sound series included voiced, breathy and creaky sounds (sounds manifesting a harmonic or quasi-harmonic spectrum), the other series included whispered sounds (sounds lacking a harmonic spectrum). Sound production included variation of f_o, vocal effort, vowel context (V and CVCV context) and production style. With few exceptions, the selected sounds were fully recognised in the standard listening test conducted when creating the corpus (100% vowel recognition rate matching vowel intention).

As a result and adding to insights of earlier experiments presented in the previous main chapters, the two sound samples compiled document the below main aspects regarding the observable variation of spectral peak numbers in vowel-related frequency ranges and their relation to the concepts of formant merging, *F*2-prime or spurious formants.

Sounds of /u, o, a/ produced with voiced, breathy or creaky phonation can manifest either one or two lower spectral peaks < 1.5 kHz or, for /o, a/, even three peaks. In addition, sounds with dominant *H*1 and weak or "undetectable" first (expected) peak, or with a rippled spectrum < 1–1.5 kHz, or with dominant *H*1 or *H*1 and *H*2 and a subsequent single peak < 1.5 kHz, also occur. Sounds of /o, a/ produced with whispered phonation can manifest one to three spectral peaks or a frequency band of prominent spectral energy.

Sounds of /ɛ, ø, e, y, i/ produced with voiced, breathy or creaky phonation can manifest many different types of peak structures such as (i) weak or barely definable first spectral peak structure < 1 kHz in terms of sloping spectral energy, (ii) dominant *H*1 or *H*1 and *H*2 for voiced or breathy sounds or a low spectral peak for creaky sounds preceding the first expected peak, (iii) prominent rippled low spectral energy or two low spectral peaks for creaky sounds, (iv) no separating peak structure for the frequency range of the first two expected spectral peaks, (v) "absent" expected second or third spectral peak, (vi) weak higher peak structure and low energy in the corresponding frequency range, and (vii) sounds at high f_o levels with unassessable spectral peak structure. Sounds of these front vowels produced with whispered phonation can manifest "split" lower spectral peaks < 1.5 kHz in terms of two peaks in the frequency range of an expected single peak (in some cases of sounds of /ɛ/, even three peaks were indicated), and they can also manifest only one peak in the vowel-related frequency range > c. 1.3 kHz.

Figures 1 and 2 illustrate these main indications of formant number alteration for sounds of back vowels and /a/, and Figures 3–5 illustrate the phenomenon for sounds of front vowels.

Concerning the formant merging concept, the most apparent contradiction to formant merging as an explanation for different spectral peak numbers was observed for sounds of back vowels: Comparing sounds of /u/ or /o/ produced at similar f_o levels which manifested either only one or two (expected) spectral peak(s) showed that the first peak frequencies were similar but, for the sounds with only one peak, an (expected) second peak was "missing". This finding indicated that, for sounds of these two vowels, the second spectral peak might affect sound timbre but not vowel quality (for illustration, see the first two sound pairs in Figure 1). Concerning sounds of front vowels, sounds with three expected vowel-related spectral peaks and sounds with only two peaks occurred for both cases of an “absent” second or third peak. This observation contradicted, in its turn, the concepts of formant merging or an *F*2-prime in between (expected) *F*2 and *F*3.

As for the notion that assumed spurious formants could explain a higher number of peaks than expected based on phonetic knowledge, sounds with prominent *H*1 preceding an expected low spectral peak, creaky sounds with three lower peaks, and whispered sounds of front vowels with two lower peaks < 1 kHz can hardly be attributed to rare effects of a speaker’s individual production characteristics. Rather, they have to be accounted for as part of the occurring general spectral variation of vowel sounds.

Concluding, any phenomenological investigation of natural vowel sounds will provide evidence that, firstly, a vowel quality is not acoustically defined by a constant number of spectral peaks *in general* and, secondly, the inconstant peak numbers documented here cannot *generally* be explained by formant merging, *F*2-prime or spurious formants. Thus, a vowel sound does not relate to a specific number of vowel quality-related spectral peaks, again indicating that vowel perception cannot be approached within a spectral peak-picking concept.

Concerning *F*-pattern estimation, once again, there was only a weak or no methodological substantiation for many of the documented sounds with spectral peak patterns deviating from the general expectation.

For references, extended background information, details of method and results, extended discussion and documentation of results (tables including sound links), see the Materials, Chapter M7.1.

Note that all figures below represent extracts of the entire sound sample investigated, as documented in Chapter M7.1.

Figure 1. Sounds of /u, o, a/ produced by women and men manifesting different spectral peak numbers < 1.5 kHz. Sounds 1–6 = three sound pairs of /u/, /o/ and /a/ illustrating present or absent expected second spectral peak < 1.5 kHz. Sounds 7–9 = three sounds of /o/ and /a/ illustrating three spectral peaks < 1.5 kHz or a “rippled” lower spectrum. Sounds 10–12 = three sounds of /a/ illustrating cases of a single peak, a frequency band with prominent spectral energy, and three peaks < 1.5 kHz for whispered sounds.
[C-07-01-F01]

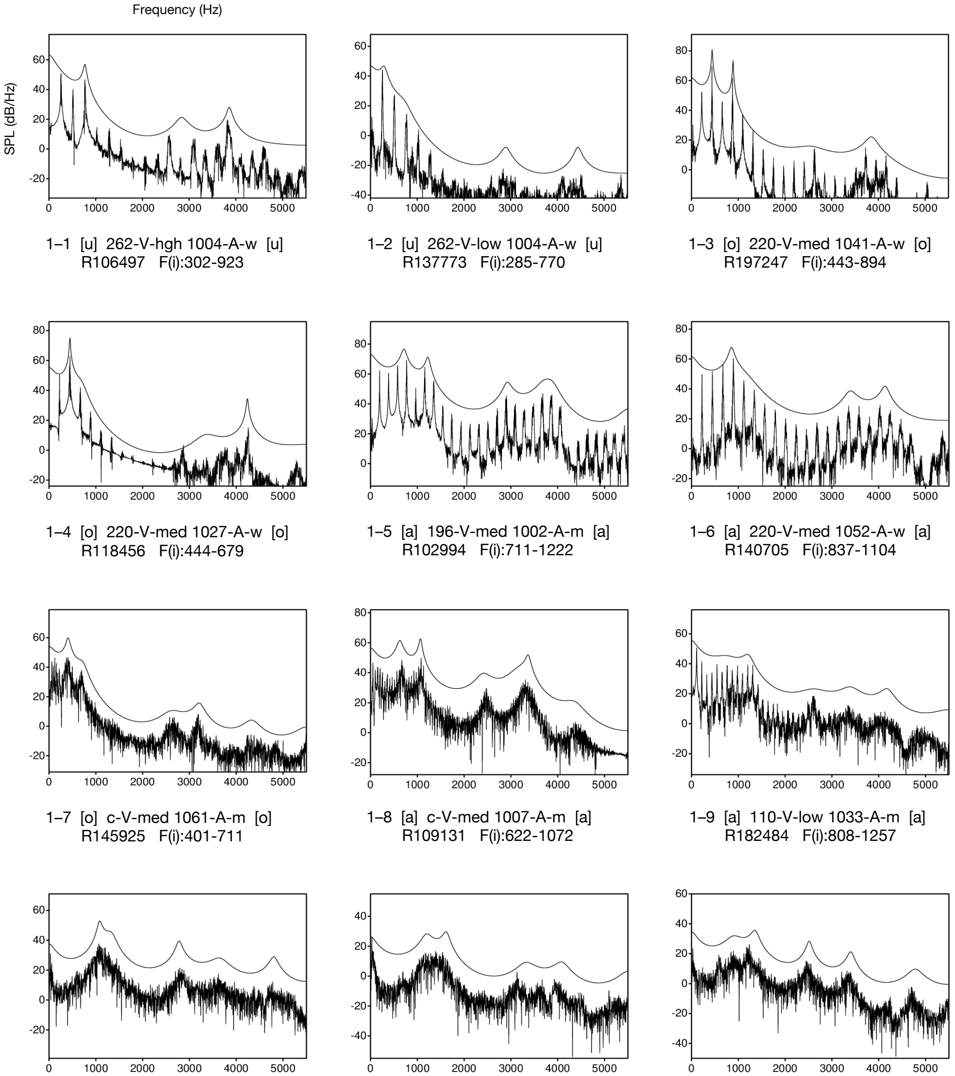
Figure 2. Spectral peak variation for a series of sounds of /a/ produced by a child. 11 sounds (including variation of f_o and phonation type) that manifest different patterns of relative spectral energy minima and maxima are shown.
[C-07-01-F02]

Figure 3. Sounds of /ɛ, e, ø/ produced by women and men manifesting different types of spectral peak structure. Sounds 1–3 = three sounds of /ɛ/ illustrating barely definable first spectral peak structure < 1 kHz in terms of sloping spectral energy. Sounds 4–6 = three sounds of /ɛ/ illustrating dominant *H*1 or *H*1 and *H*2 or a low spectral peak preceding the first expected peak. Sounds 7–9 = three sounds of /e/ illustrating two prominent lower spectral peaks < 1.5 kHz or a “rippled” lower spectral energy. Sounds 10–12 = three sounds of /ø/ illustrating the lack of separated spectral peaks for the frequency range of the first two expected peaks.
[C-07-01-F03]

Figure 4. Sounds of /e, y, ø/ produced by children, women and men manifesting different types of spectral peak structure. Sounds 1–4 = two sound pairs of /e/ illustrating present or absent expected second spectral peak. Sounds 5–8 = two sound pairs of /y/ illustrating present or absent expected third spectral peak. (Note that each of these four pairs was produced by a single speaker.) Sounds 9 and 10 = two sounds of /ø/ illustrating weak higher peak structure and low spectral energy in the corresponding frequency range.
[C-07-01-F04]

Figure 5. Sounds of /ɛ, e, y, i, ø/ produced by women and men manifesting different types of spectral peak structure. Sounds 1–6 = six sounds of /ɛ, e, y, i/ produced at higher f_o levels, illustrating fully or partially unassessable spectral peak structure. Sounds 7–9 = three whispered sounds of /ɛ, ø, e/ illustrating two or three spectral peaks in the lower frequency range of an expected single peak. Sounds 10 and 11 = a whispered sound pair of /e/ illustrating one or two peak(s) > 1.3 kHz.
[C-07-01-F05]

Figure 1. Sounds of /u, o, a/ produced by women and men manifesting different spectral peak numbers < 1.5 kHz. [C-07-01-F01]



1–1 [u] 262-V-hgh 1004-A-w [u]
R106497 F(i):302-923

1–2 [u] 262-V-low 1004-A-w [u]
R137773 F(i):285-770

1–3 [o] 220-V-med 1041-A-w [o]
R197247 F(i):443-894

1–4 [o] 220-V-med 1027-A-w [o]
R118456 F(i):444-679

1–5 [a] 196-V-med 1002-A-m [a]
R102994 F(i):711-1222

1–6 [a] 220-V-med 1052-A-w [a]
R140705 F(i):837-1104

1–7 [o] c-V-med 1061-A-m [o]
R145925 F(i):401-711

1–8 [a] c-V-med 1007-A-m [a]
R109131 F(i):622-1072

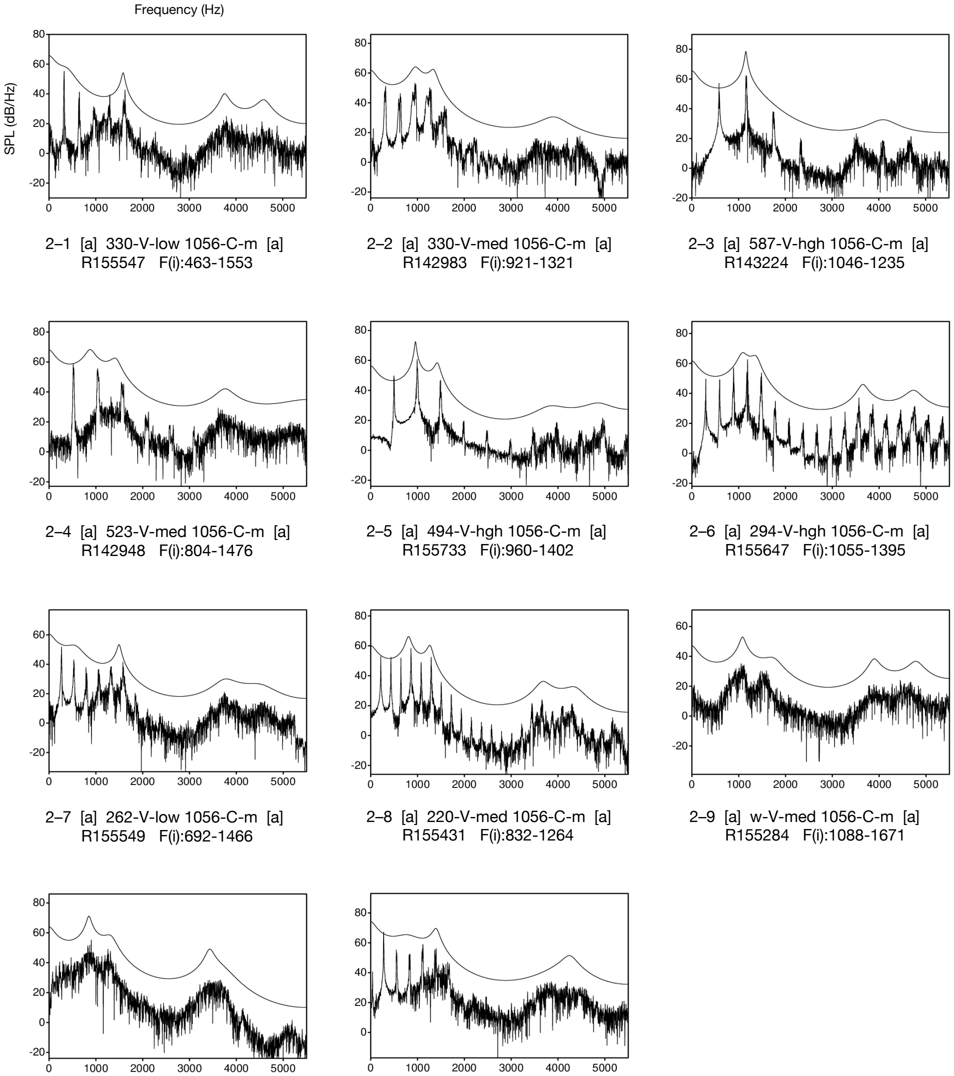
1–9 [a] 110-V-low 1033-A-m [a]
R182484 F(i):808-1257

1–10 [a] w-V-med 1063-A-m [a]
R149592 F(i):1044-1324

1–11 [a] w-sVsV-med 1088-A-w [a]
R192679 F(i):1207-1541

1–12 [a] w-sVsV-med 1079-A-m [a]
R202575 F(i):900-1294

Figure 2. Spectral peak variation for a series of sounds of /a/ produced by a child. [C-07-01-F02]



2–1 [a] 330-V-low 1056-C-m [a]
R155547 F(i):463-1553

2–2 [a] 330-V-med 1056-C-m [a]
R142983 F(i):921-1321

2–3 [a] 587-V-hgh 1056-C-m [a]
R143224 F(i):1046-1235

2–4 [a] 523-V-med 1056-C-m [a]
R142948 F(i):804-1476

2–5 [a] 494-V-hgh 1056-C-m [a]
R155733 F(i):960-1402

2–6 [a] 294-V-hgh 1056-C-m [a]
R155647 F(i):1055-1395

2–7 [a] 262-V-low 1056-C-m [a]
R155549 F(i):692-1466

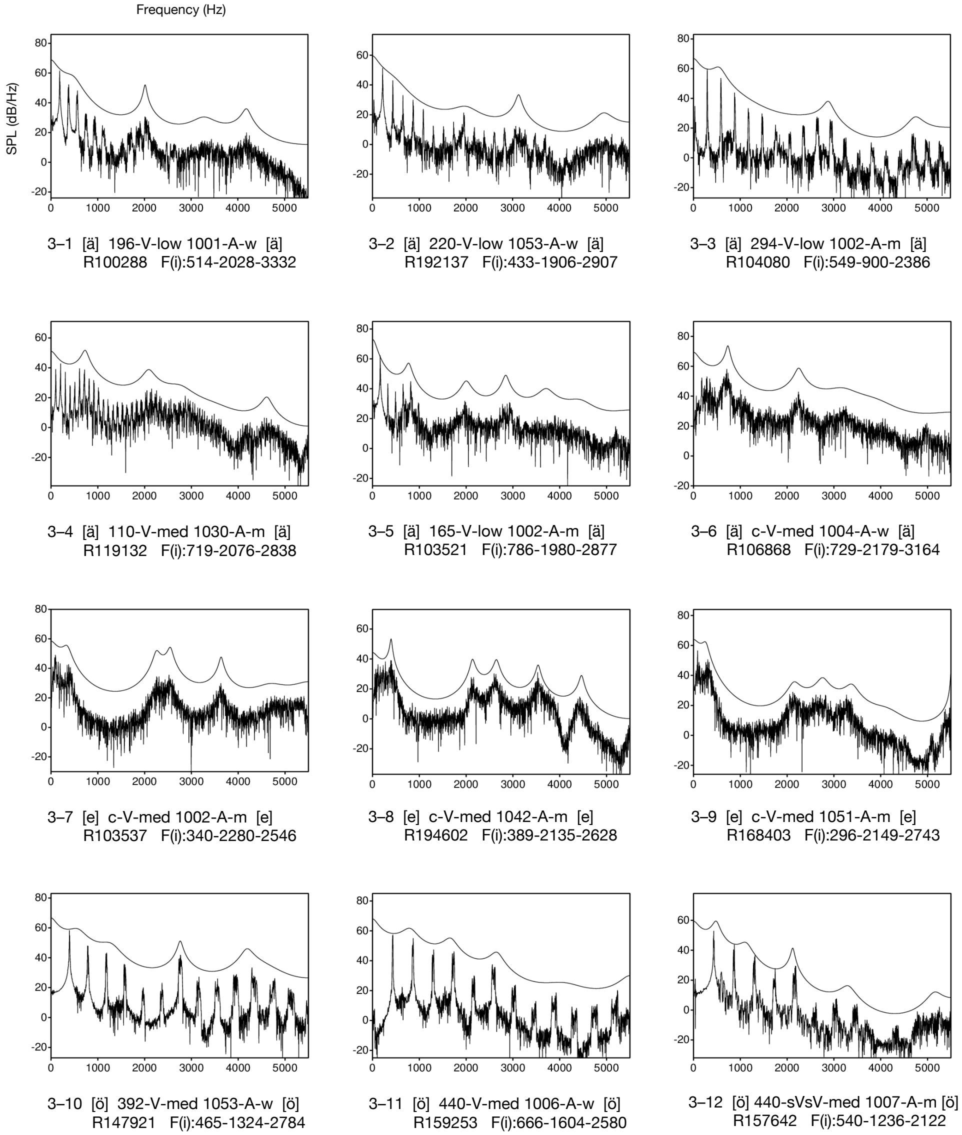
2–8 [a] 220-V-med 1056-C-m [a]
R155431 F(i):832-1264

2–9 [a] w-V-med 1056-C-m [a]
R155284 F(i):1088-1671

2–10 [a] c-V-med 1056-C-m [a]
R155307 F(i):832-1278

2–11 [a] 262-V-low 1056-C-m [a]
R155247 F(i):697-1390

Figure 3. Sounds of /ɛ, e, ø/ produced by women and men manifesting different types of spectral peak structure. [C-07-01-F03]



3–1 [ä] 196-V-low 1001-A-w [ä]
R100288 F(i):514-2028-3332

3–2 [ä] 220-V-low 1053-A-w [ä]
R192137 F(i):433-1906-2907

3–3 [ä] 294-V-low 1002-A-m [ä]
R104080 F(i):549-900-2386

3–4 [ä] 110-V-med 1030-A-m [ä]
R119132 F(i):719-2076-2838

3–5 [ä] 165-V-low 1002-A-m [ä]
R103521 F(i):786-1980-2877

3–6 [ä] c-V-med 1004-A-w [ä]
R106868 F(i):729-2179-3164

3–7 [e] c-V-med 1002-A-m [e]
R103537 F(i):340-2280-2546

3–8 [e] c-V-med 1042-A-m [e]
R194602 F(i):389-2135-2628

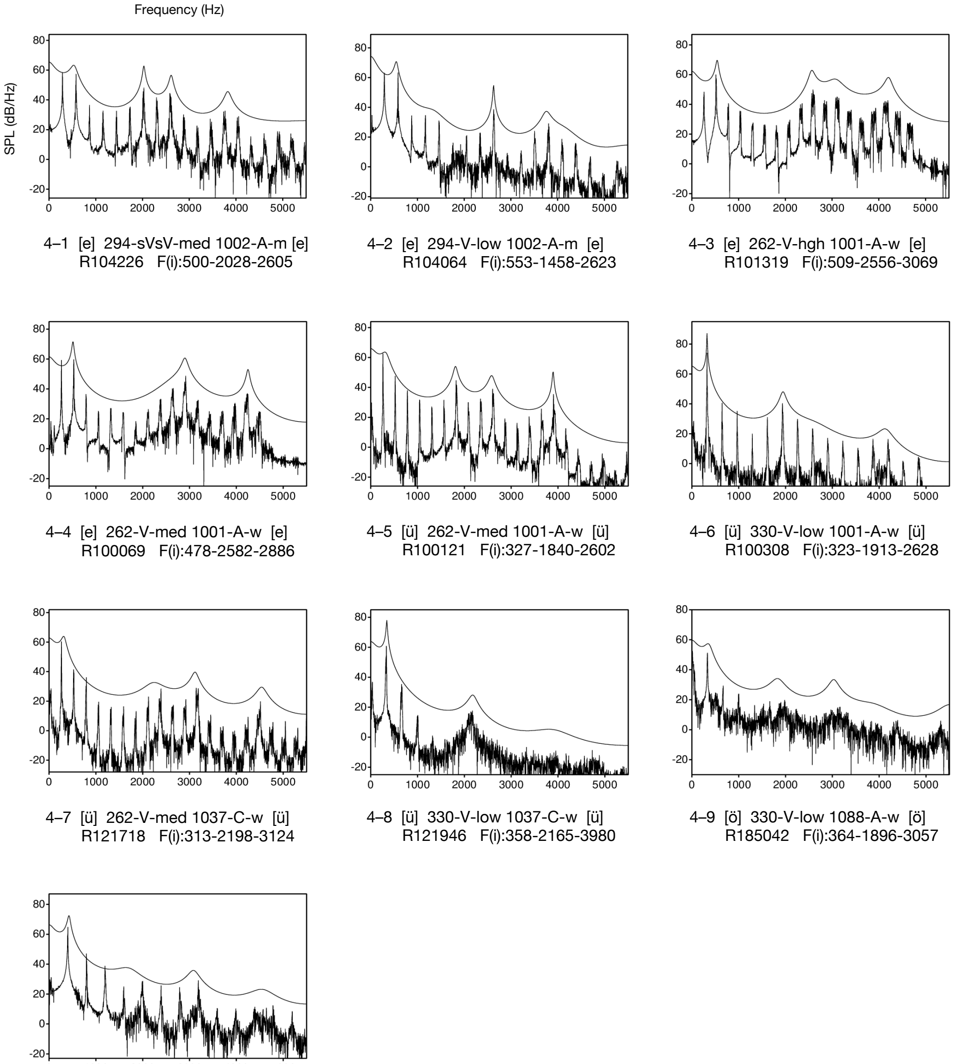
3–9 [e] c-V-med 1051-A-m [e]
R168403 F(i):296-2149-2743

3–10 [ö] 392-V-med 1053-A-w [ö]
R147921 F(i):465-1324-2784

3–11 [ö] 440-V-med 1006-A-w [ö]
R159253 F(i):666-1604-2580

3–12 [ö] 440-sVsV-med 1007-A-m [ö]
R157642 F(i):540-1236-2122

Figure 4. Sounds of /e, y, ø/ produced by children, women and men manifesting different types of spectral peak structure. [C-07-01-F04]



4–1 [e] 294-sVsV-med 1002-A-m [e]
R104226 F(i):500-2028-2605

4–2 [e] 294-V-low 1002-A-m [e]
R104064 F(i):553-1458-2623

4–3 [e] 262-V-hgh 1001-A-w [e]
R101319 F(i):509-2556-3069

4–4 [e] 262-V-med 1001-A-w [e]
R100069 F(i):478-2582-2886

4–5 [ü] 262-V-med 1001-A-w [ü]
R100121 F(i):327-1840-2602

4–6 [ü] 330-V-low 1001-A-w [ü]
R100308 F(i):323-1913-2628

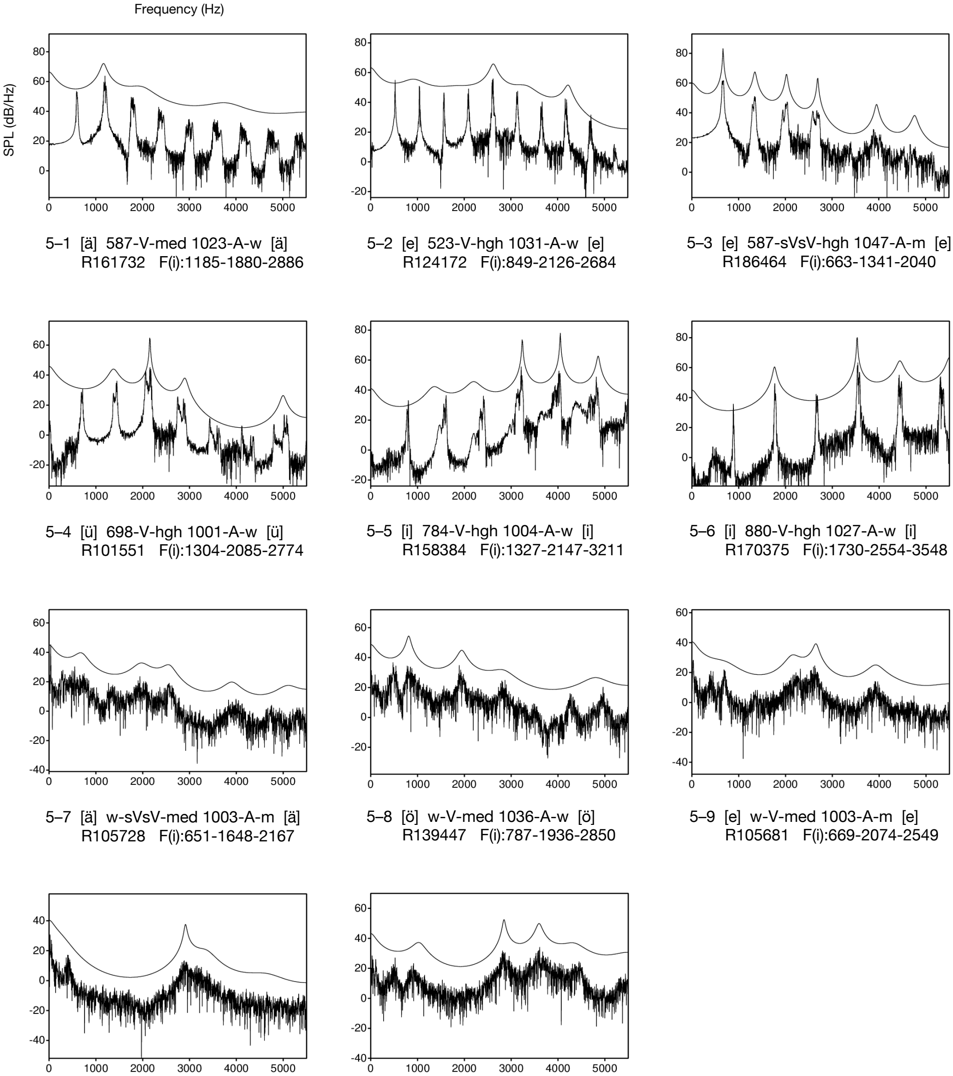
4–7 [ü] 262-V-med 1037-C-w [ü]
R121718 F(i):313-2198-3124

4–8 [ü] 330-V-low 1037-C-w [ü]
R121946 F(i):358-2165-3980

4–9 [ö] 330-V-low 1088-A-w [ö]
R185042 F(i):364-1896-3057

4–10 [ö] 392-V-low 1054-C-m [ö]
R177419 F(i):433-1570-3096

Figure 5. Sounds of /ɛ, e, y, i, ø/ produced by women and men manifesting different types of spectral peak structure. [C-07-01-F05]



5–1 [ä] 587-V-med 1023-A-w [ä]
R161732 F(i):1185-1880-2886

5–2 [e] 523-V-hgh 1031-A-w [e]
R124172 F(i):849-2126-2684

5–3 [e] 587-sVsV-hgh 1047-A-m [e]
R186464 F(i):663-1341-2040

5–4 [ü] 698-V-hgh 1001-A-w [ü]
R101551 F(i):1304-2085-2774

5–5 [i] 784-V-hgh 1004-A-w [i]
R158384 F(i):1327-2147-3211

5–6 [i] 880-V-hgh 1027-A-w [i]
R170375 F(i):1730-2554-3548

5–7 [ä] w-sVsV-med 1003-A-m [ä]
R105728 F(i):651-1648-2167

5–8 [ö] w-V-med 1036-A-w [ö]
R139447 F(i):787-1936-2850

5–9 [e] w-V-med 1003-A-m [e]
R105681 F(i):669-2074-2549

5–10 [e] w-sVsV-med 1048-A-w [e]
R130609 F(i):559-2468-3065

5–11 [e] w-V-med 1053-A-w [e]
R148877 F(i):953-2804-3562

7.2 Inversions of Vowel-Related Relative Spectral Energy Maxima and Minima

The second observation discussed in this main chapter concerns the occurrence of inversions of relative spectral energy maxima and minima for sounds of a given vowel in vowel-related frequency ranges. As explained in the previous chapter, comparing sounds of the back vowels /u, o/ of the Zurich Corpus, the sounds produced at similar f_o levels with either two (expected) spectral peaks or only one single peak < 1.5 kHz showed that the first peak frequencies for those sounds were similar in most cases but, for the sounds with only one peak, the second (expected) peak was "missing". However, as discussed in the Preliminaries (see p. 62 and pp. 183–186), if sounds of these vowels produced at lower or middle levels of f_o and manifesting two spectral peaks were compared with other sounds produced at middle or higher f_o with only one spectral peak, inverse relative spectral maxima and minima in the form of inverse spectral envelope curves ≤ 1.5 kHz occurred, without any change in vowel recognition: Thus, whereas a relative energy minimum in between two peaks in the spectrum can be manifest for one sound of a vowel, a single spectral energy maximum can be manifest for another sound of that vowel. The same holds true for comparisons between the respective calculated filter curves and estimated formant patterns (if the estimation is methodologically substantiated). Similar observations were made for sounds of /a/, but they did not systematically relate to f_o variation. For sounds of front vowels, such inversions were, in their turn, observed for the vowel-specific frequency range > 1 kHz, but they were often related to marked vocal effort variations.

To again document the possible variability of the vowel spectrum on the new basis of the Zurich Corpus and to embed it in the line of argument of this treatise, a corresponding study was conducted: Based on the inspection of the corpus, for each of the three long Standard German vowels /u, o, a/ and each of the three speaker groups of men, women and children, sound pairs were compiled for which the first sound manifested two distinct relative spectral energy maxima < 1.5 kHz and the second sound manifested a single maximum in between the two maxima of the first sound. The study was limited to sounds of back vowels and /a/ because, for sounds of these vowels, occurring inversions concern the entire vowel-related spectral frequency range and, according to our previous experiences in the context of the investigation discussed in the Preliminaries, the inversions can often be observed not only for sounds with marked vocal effort variation but also for sounds produced without intended effort variation. The sound sample created was limited to sounds produced in nonstyle mode at various f_o levels with medium vocal effort in V context, and for which the listening test conducted when creating the corpus provided a 100% recognition rate matching vowel intention.

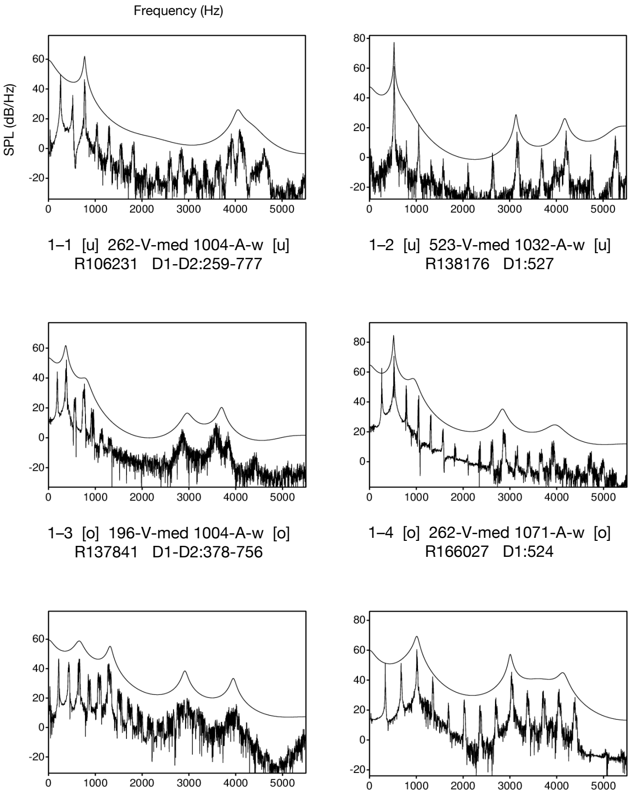
As a result, for each of the three vowels, numerous sound pairs with inverted spectral energy minima and maxima were found in the corpus, as was to be expected against the background of our earlier studies. Further, when analysing the frequency distance between the two lower spectral peaks of the first sound of a pair, for some of these sounds, the frequency distance in Bark was near the upper limit of 3.5 Bark (upper limit for spectral integration within the “centre of gravity” approach) or even exceeding this limit. Thus, in its turn, the occurrence of inverted spectral energy minima and maxima for sounds of a vowel cannot, *in general,* be explained by spectral integration. Finally, when comparing the f_o levels of the two sounds of a pair, the level of the first sound was generally found to be markedly below the level of the second sound.

For exemplary documentation of the inversion phenomenon in this treatise, three sound pairs per vowel produced by men, women and children were selected. For each single sound pair, the first sound manifested two distinct relative spectral energy maxima < 1.5 kHz, including two single harmonics forming the tips of the peaks, and the second sound manifested a single distinct relative spectral energy maximum, including a single harmonic forming the peak tip, this single peak lying in between the two peaks of the first sound. As a result, a compilation of three sound pairs per vowel (9 sound pairs and 18 single sounds in total) was created. As a sample of this documentation, Figure 1 illustrates the inversion phenomenon, presenting a sound pair for each of the three vowels investigated.

For references, details of method and results, extended discussion and documentation of results (table including sound links), see the Materials, Chapter M7.2.

Figure 1. Sound pairs of /u, o, a/ produced by women illustrating inversions of vowel-related relative spectral energy maxima and minima. Extracts of Chapter M7.2, Table 1 (see Series 2, 5 and 8 in this table). Spectral peak frequencies in terms of calculated frequencies of the dominant harmonics $D_{(i)}$ < 1.5 kHz are indicated. Sounds 1 and 2 = two sounds of /u/; note that, for the first sound of the pair, the difference between D_1 and D_2 is 4.48 Bark. Sounds 3 and 4 = two sounds of /o/; for the first sound of the pair, the difference between D_1 and D_2 is 3.13 Bark. Sounds 5 and 6 = two sounds of /a/; for the first sound of the pair, the difference between D_1 and D_2 is 4.02 Bark.
[C-07-02-F01]

Figure 1. Sound pairs of /u, o, a/ produced by women illustrating inversions of vowel-related relative spectral energy maxima and minima. [C-07-02-F01]



1–1 [u] 262-V-med 1004-A-w [u]
R106231 D1-D2:259-777

1–2 [u] 523-V-med 1032-A-w [u]
R138176 D1:527

1–3 [o] 196-V-med 1004-A-w [o]
R137841 D1-D2:378-756

1–4 [o] 262-V-med 1071-A-w [o]
R166027 D1:524

1–5 [a] 220-V-med 1088-A-w [a]
R184426 D1-D2:651-1302

1–6 [a] 330-V-med 1046-A-w [a]
R159930 D1:1014

7.3 Flat or Sloping Vowel-Related Spectral Energy Distribution in Natural Vowel Sounds

The third observation discussed in this main chapter concerns sounds with flat or sloping spectral energy distribution in the entire vowel-related frequency range or in the upper part of that range. Early synthesis experiments reported in the literature already indicated that a stepwise increase in the number of harmonics from *H*1 upwards with equal harmonic levels resulted in recognisable vowel sounds, vowel quality shifting incrementally from /u/ to /o/ to /ɔ/ to /ɑ/ and finally to /a/. The same held true for a stepwise increase in the number of harmonics with decreasing harmonic levels (with a spectral slope). Furthermore, a comparable demonstration was given for recognisable sounds of front vowels synthesised based on one or two low harmonic(s) and a band of harmonics in the higher frequencies, with vowel-related series of the higher harmonics equal or decreasing in their levels. (The next chapter will discuss our re-examination of some of these phenomena in vowel synthesis.) Finally, recognisable sounds of front vowels are reported, which were synthesised with only *F*1 as a well-specified spectral peak, combined with a broad higher frequency region of energy with no marked peaks.

Concerning natural sounds, in the Preliminaries (pp. 57–58 and 147–157), we have discussed and documented cases of sounds of back vowels and /a/ whose spectra exhibited a series of harmonics with quasi-identical or with continuously decreasing amplitudes in the lower frequency range < c. 1.5 kHz (flat or sloping spectral energy distribution in vowel-related frequency ranges). We have also discussed and documented cases of sounds of front vowels that manifested series of harmonics with quasi-identical amplitudes in the higher frequency range > c. 1.3 kHz (flat spectral energy distribution in the vowel-related portions of higher frequency ranges).

To replicate and expand on this aspect of the vowel spectrum and to embed it into the line of argument of this treatise, based on the new sound sample of the Zurich Corpus, a study was conducted to provide exemplary compilations of natural sounds with flat or sloping spectral energy distribution either throughout the entire vowel-related frequency range or in the higher part of it: Inspecting the corpus, for each of the eight long Standard German vowels, exemplary samples of sounds produced by speakers of all three speaker groups (children, women and men) with voiced or breathy phonation and including variation of f_o, vocal effort, vowel context (V and sVsV context) and production style were compiled for which the spectra exhibited flat or sloping spectral energy distribution in vowel-related frequency ranges. The sample investigated was limited to fully recognised sounds (100% vowel recognition rate according to the standard listening test conducted when creating the corpus, matching vowel intention).

As a result and confirming earlier findings, for each of the eight vowels, numerous sounds with both flat and sloping spectra or spectral parts were found, of which only a sample was selected for documentation in this treatise. For the sounds of /u, o, a/, above all, two general types of either flat or sloping energy distribution < c. 1.5 kHz were observed: As a tendency, for the lower range of f_o < c. 250 Hz, the selected sounds related to low vocal effort in voiced phonation or to breathy phonation. No vocal effort-specific relation of the selected sounds was manifest for the frequencies above this range of f_o. For the sounds of /ɛ/, /ø/ and /e/, dependent on the f_o level of the sounds, above all, three types of spectral manifestations were observed for the vowel-related spectral frequency range: A spectral peak or prominent frequency band in the lower range and a flat energy distribution in the higher range; only sloping energy distribution; only flat energy distribution. For the sounds of /y/ and /i/, also dependent on the f_o level of the sounds, above all, two types of spectral manifestations were observed for the vowel-related spectral frequency range: A spectral peak in the lower frequency range associated with flat energy distribution in the higher range, or only flat energy distribution. As extracts of the entire investigated sound sample, Figures 1–3 exemplify this phenomenon.

In conclusion, flat or sloping energy distribution in a vowel spectrum proved not to be a rare phenomenon of vowel sounds, and it was not limited to a specific type of vowel production. This finding indicates, in its turn, that vowel quality recognition cannot *generally* be attributed to discrimination of spectral peaks.

In the context of vowel sounds with flat or sloping spectra or spectral parts, once again, methodological substantiation for formant pattern estimation often proved to be weak or lacking, as was the case for vowel sounds with different vowel-related spectral peak numbers. Further, for some comparisons of sounds of different vowels with flat or sloping spectral energy distribution, the vowel-related spectral difference was barely understood based on existing phonetic knowledge. Figure 4 illustrates three cases of such comparisons.

For references, extended background information, details of method and results, extended discussion and documentation of results (table including sound links), see the Materials, Chapter M7.3.

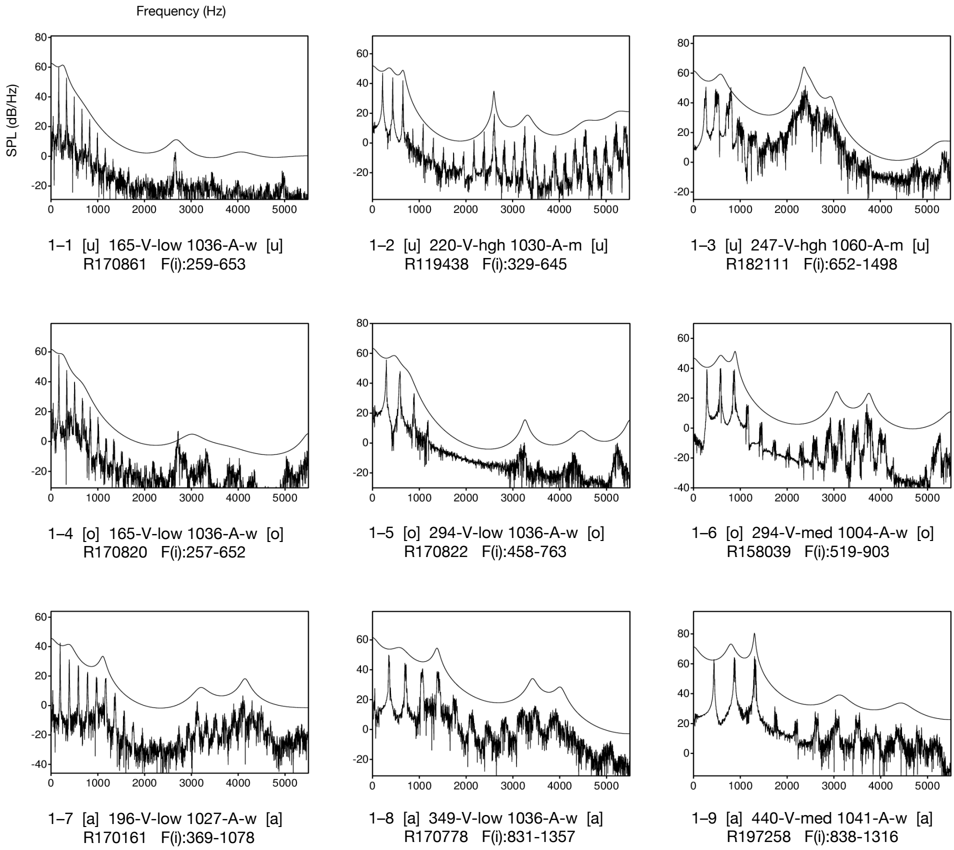
Figure 1. Sounds of /u, o, a/ produced by women and men manifesting flat or sloping spectral energy distribution in the entire vowel-related frequency range. Extracts of Chapter M7.3, Table 1 (see Series 1–3 in this table).
[C-07-03-F01] ↗

Figure 2. Sounds of /ɛ, ø/ produced by women and men manifesting flat or sloping spectral energy distribution in the entire vowel-related frequency range or just a part of it. Extracts of Chapter M7.3, Table 1 (see Series 4 and 5 in this table).
[C-07-03-F02] ↗

Figure 3. Sounds of /e, y, i/ produced by children, women and men manifesting flat or sloping spectral energy distribution in the entire vowel-related frequency range or just a part of it. Extracts of Chapter M7.3, Table 1 (see Series 6–8 in this table).
[C-07-03-F03] ↗

Figure 4. Comparison of sounds of /u/ and /o/, and of /o/ and /a/, for which vowel-related spectral differences are difficult to understand based on existing phonetic knowledge. Extracts of the sound sample shown in Figure 1. Sounds 1 and 2 = comparison of a sound of /u/ with a sound of /o/ for which the harmonic spectrum < 1 kHz shows no clearly identifiable difference according to existing phonetic knowledge despite similar f_o levels. Sounds 3 and 4 = second comparison of a sound of /u/ with a sound of /o/ for which neither *F*-pattern estimation nor a direct comparison of the spectrograms allows for an identifiable difference according to existing phonetic knowledge. Sounds 5 and 6 = comparison of a sound of /o/ with a sound of /a/ for which *F1* estimation applying a speaker group-related default parameter for the maximum number of formants of LPC analysis results in a lower *F1* for the sound of /a/ than for the sound of /o/, in contrast to values given in formant statistics.
[C-07-03-F04] ↗

Figure 1. Sounds of /u, o, a/ produced by women and men manifesting flat or sloping spectral energy distribution in the entire vowel-related frequency range. [C-07-03-F01]



1–1 [u] 165-V-low 1036-A-w [u]
R170861 F(i):259-653

1–2 [u] 220-V-hgh 1030-A-m [u]
R119438 F(i):329-645

1–3 [u] 247-V-hgh 1060-A-m [u]
R182111 F(i):652-1498

1–4 [o] 165-V-low 1036-A-w [o]
R170820 F(i):257-652

1–5 [o] 294-V-low 1036-A-w [o]
R170822 F(i):458-763

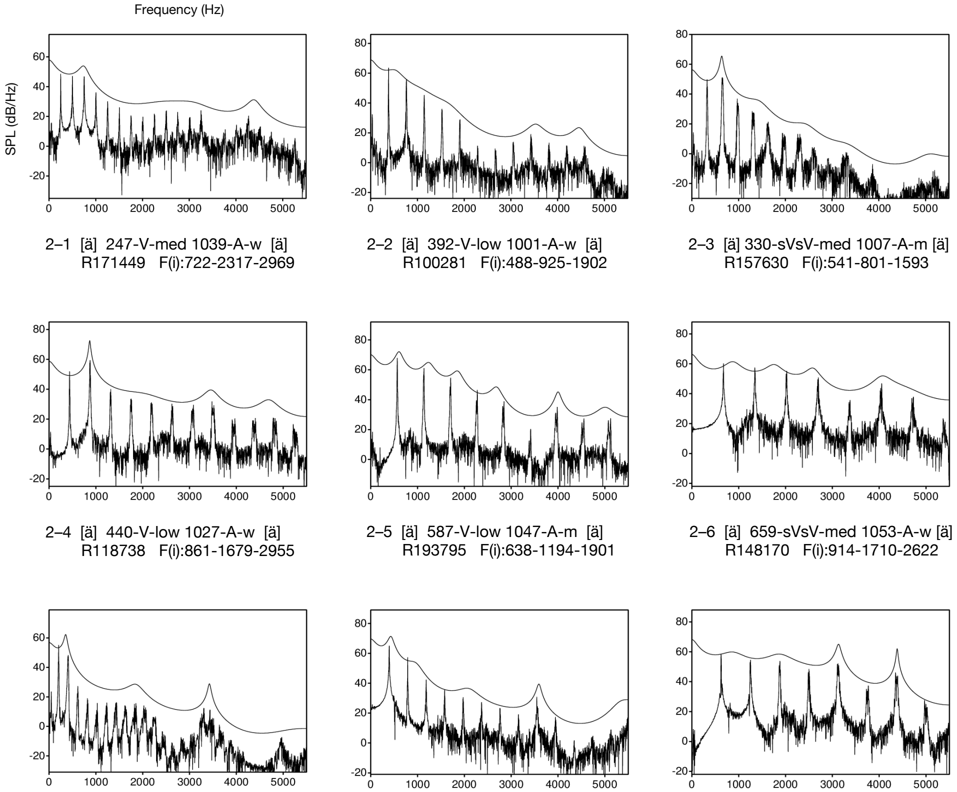
1–6 [o] 294-V-med 1004-A-w [o]
R158039 F(i):519-903

1–7 [a] 196-V-low 1027-A-w [a]
R170161 F(i):369-1078

1–8 [a] 349-V-low 1036-A-w [a]
R170778 F(i):831-1357

1–9 [a] 440-V-med 1041-A-w [a]
R197258 F(i):838-1316

Figure 2. Sounds of /ɛ, ø/ produced by women and men manifesting flat or sloping spectral energy distribution in the entire vowel-related frequency range or just a part of it. [C-07-03-F02]



2–1 [ä] 247-V-med 1039-A-w [ä]
R171449   F(i):722-2317-2969

2–2 [ä] 392-V-low 1001-A-w [ä]
R100281   F(i):488-925-1902

2–3 [ä] 330-sVsV-med 1007-A-m [ä]
R157630   F(i):541-801-1593

2–4 [ä] 440-V-low 1027-A-w [ä]
R118738   F(i):861-1679-2955

2–5 [ä] 587-V-low 1047-A-m [ä]
R193795   F(i):638-1194-1901

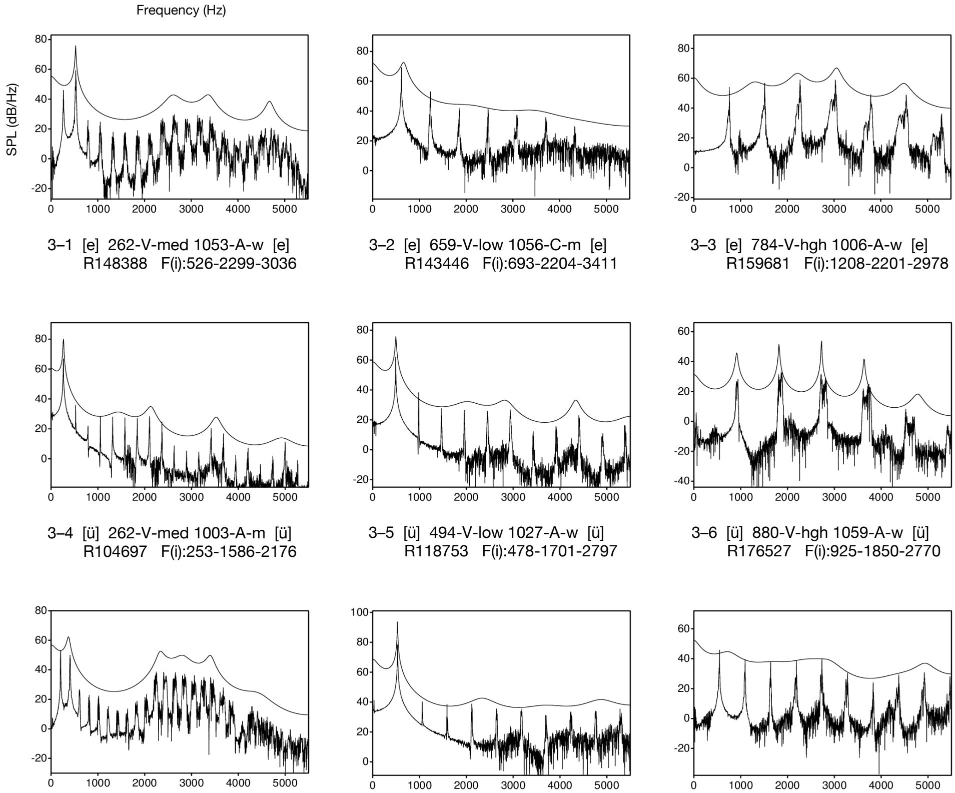
2–6 [ä] 659-sVsV-med 1053-A-w [ä]
R148170   F(i):914-1710-2622

2–7 [ö] 196-V-low 1063-A-m [ö]
R149312   F(i):354-1381-2012

2–8 [ö] 392-V-med 1063-A-m [ö]
R173465   F(i):467-1108-2029

2–9 [ö] 659-V-hgh 1086-A-w [ö]
R188623   F(i):925-2082-3112

Figure 3. Sounds of /e, y, i/ produced by children, women and men manifesting flat or sloping spectral energy distribution in the entire vowel-related frequency range or just a part of it. [C-07-03-F03]



3–1 [e] 262-V-med 1053-A-w [e]
R148388 F(i):526-2299-3036

3–2 [e] 659-V-low 1056-C-m [e]
R143446 F(i):693-2204-3411

3–3 [e] 784-V-hgh 1006-A-w [e]
R159681 F(i):1208-2201-2978

3–4 [ü] 262-V-med 1003-A-m [ü]
R104697 F(i):253-1586-2176

3–5 [ü] 494-V-low 1027-A-w [ü]
R118753 F(i):478-1701-2797

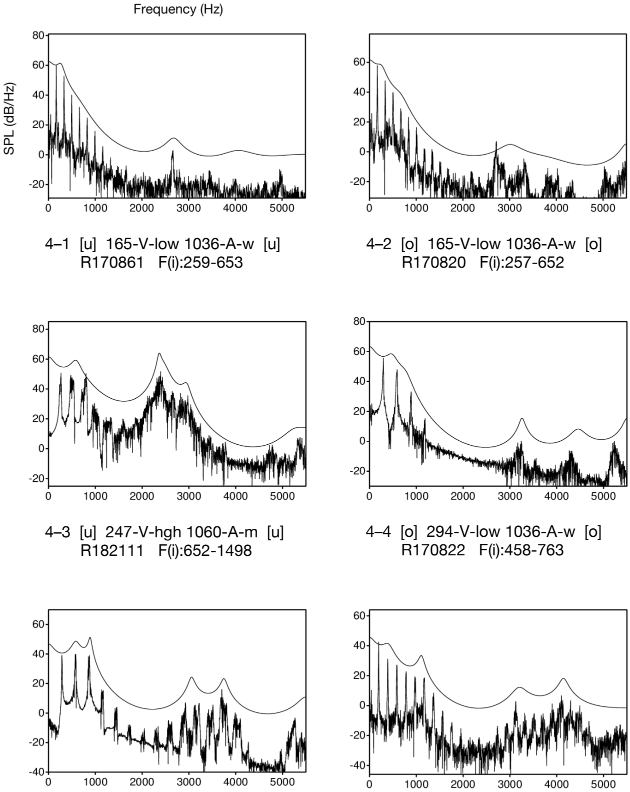
3–6 [ü] 880-V-hgh 1059-A-w [ü]
R176527 F(i):925-1850-2770

3–7 [i] 196-V-hgh 1044-A-m [i]
R125743 F(i):382-2281-2661

3–8 [i] 523-V-med 1104-C-m [i]
R195823 F(i):528-2491-3700

3–9 [i] 523-V-med 1060-A-m [i]
R196122 F(i):600-1657-2503

Figure 4. Comparison of sounds of /u/ and /o/, and of /o/ and /a/, for which vowel-related spectral differences are difficult to understand based on existing phonetic knowledge. [C-07-03-F04]



4–1 [u] 165-V-low 1036-A-w [u]
R170861 F(i):259-653

4–2 [o] 165-V-low 1036-A-w [o]
R170820 F(i):257-652

4–3 [u] 247-V-hgh 1060-A-m [u]
R182111 F(i):652-1498

4–4 [o] 294-V-low 1036-A-w [o]
R170822 F(i):458-763

4–5 [o] 294-V-med 1004-A-w [o]
R158039 F(i):519-903

4–6 [a] 196-V-low 1027-A-w [a]
R170161 F(i):369-1078

7.4 Flat Vowel-Related Spectral Energy Distribution in Synthesised Vowel Sounds

In the context of discussing vowel-related flat or sloping spectral energy distribution, the vowel recognition of correspondingly synthesised sounds is worth considering: As mentioned in the introduction of the previous chapter, early synthesis experiments already demonstrated sounds of back vowels and of /a/ with flat spectra in terms of series of consecutive harmonics equal in amplitude or with sloping harmonic levels. To expand on the aspect of flat vowel-related spectral energy distribution and include vowel synthesis, and to replicate and re-examine the above indications of earlier studies reported in the literature, in a corresponding study, three different types of series of consecutive harmonics equal in amplitude were investigated: (i) Harmonic series as a result of a stepwise increase in the number of consecutive harmonics from *H*1 to *H*1–*H*20, (ii) harmonic series as a result of a stepwise decrease of the number of consecutive harmonics from *H*1–*H*20 to *H*20, and (iii) harmonic series as a result of a stepwise increase in the number of consecutive harmonics from a middle harmonic to a series of 11 harmonics at a maximum (e.g., from *H*2 to *H*12, or from *H*3 to *H*13 and so forth, with the last series being from *H*15 to *H*20). Harmonics were always multiples of 200 Hz. On this basis, monotonous sounds of 1.2 sec. (with a 0.1 sec. fade in/out) were produced using the SinSyn tool. As a result, a sample of 190 synthesised sounds was created. Finally, for all synthesised sounds, vowel recognition was tested according to the standard procedure of the Zurich Corpus and involving the five standard listeners.

According to the vowel recognition results, configurations of series of equal-amplitude harmonics related to recognisable sounds were found for all long Standard German vowels except /u/. (Recognisable sounds of /u/ related to synthesis based on single harmonics.) Thus, the experiment confirmed earlier indications that it is possible to synthesise recognisable vowel sounds based on entirely flat harmonic spectra in terms of series of consecutive equal-amplitude harmonics in various frequency bands. However, for the f_o level investigated (resulting from the HCF of the harmonic series), the number of harmonic configurations and synthesised sounds per vowel, as well as the related recognition rates, varied strongly, with the most contrasting findings found for the comparison of sounds of /ɛ, a/ with those of /ø/: Numerous sounds with their different harmonic configurations were recognised as /ɛ/ or /a/ by all listeners, but only one sound related to a single harmonic configuration was recognised as /ø/ with a weak labelling majority (3/5 listeners).

For documentation in this treatise, only synthesised sounds with a recognition rate of 100% were selected from the entire sound sample investigated, including sounds of /u/ related to a single harmonic (see Chapter M7.4). As an extract of this reduced sample of fully recognised sounds, for each of the vowels /u, o, ɔ, a, ɛ, e, y, i/, one sound example is shown in Figure 1, illustrating the phenomenon in question.

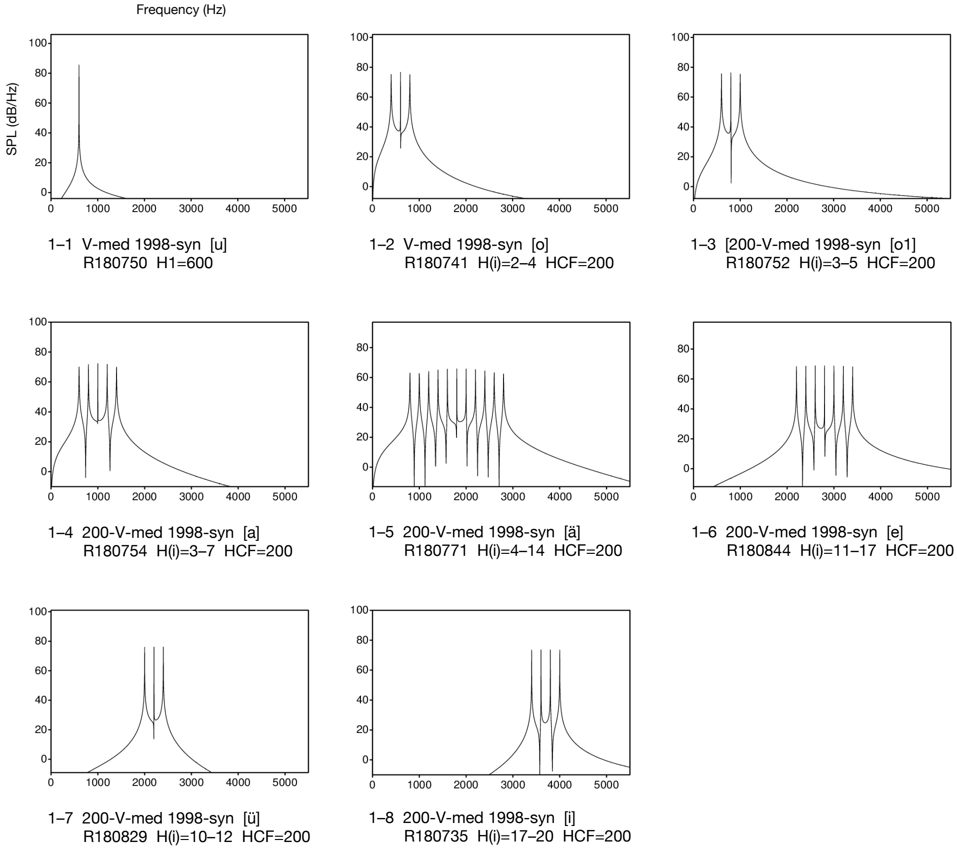
The strong variation in the number of clearly recognised sounds for different vowel qualities is difficult to interpret since only harmonics as multiples of 200 Hz were investigated, and the question of the role of *H*1 reference frequency (and corresponding HCF) in this type of experimentation was left open. This question should be addressed in future studies. (For some indications on the matter, see Chapter M7.4.)

In sum, the results of the present study were in line with the results of earlier synthesis experiments reported in the literature and expanded on the aspect that recognisable sounds could be produced in vowel synthesis based on flat harmonic spectra for all long Standard German vowels and also for /ɔ/. Thereby, the most impressive cases of sounds of this kind concerned the sounds recognised as /a/ and /ɛ/ since the related frequency bands were very large and opposed to any concept of spectral peak structure. Furthermore, the results of the present study paralleled the findings of vowel recognition for synthesised sounds of front vowels discussed in Chapter 6.7, synthesis based on series of equal-amplitude harmonics > 1 kHz combined with a single lower harmonic < 1 kHz.

For references, details of method and results, extended discussion and documentation of results (tables including sound links), see the Materials, Chapter M7.4.

Figure 1. Recognised synthesised sounds of the vowels /u, o, ɔ, a, ɛ, e, y, i/, with synthesis based on an entirely flat energy distribution in vowel-related frequency bands or on a single sinusoid. Extract of Chapter M7.4, Table 1 (see Series 2 and 3 in this table). For each of the vowels /o, ɔ, a, ɛ, e, y, i/, a sound is shown which was synthesised based on a band of equal-amplitude sinusoids in a harmonic relation, with the sounds being unanimously recognised in the recognition test. A fully recognised sound of /u/ is added, synthesised based on a single sinusoid.
[C-07-04-F01]

Figure 1. Recognised synthesised sounds of the vowels /u, o, ɔ, a, ɛ, e, y, i/, with synthesis based on an entirely flat energy distribution in vowel-related frequency bands or on a single sinusoid. [C-07-04-F01]



1–1 V-med 1998-syn [u]
R180750 H1=600

1–2 V-med 1998-syn [o]
R180741 H(i)=2–4 HCF=200

1–3 [200-V-med 1998-syn [o1]
R180752 H(i)=3–5 HCF=200

1–4 200-V-med 1998-syn [a]
R180754 H(i)=3–7 HCF=200

1–5 200-V-med 1998-syn [ä]
R180771 H(i)=4–14 HCF=200

1–6 200-V-med 1998-syn [e]
R180844 H(i)=11–17 HCF=200

1–7 200-V-med 1998-syn [ü]
R180829 H(i)=10–12 HCF=200

1–8 200-V-med 1998-syn [i]
R180735 H(i)=17–20 HCF=200

7.5 Sounds of Close and Close-Mid Vowels for Which Marked f_o Variation < 250 Hz Does Not Affect Estimated Formant Patterns and Spectral Envelopes

In the context of discussing the nonuniform character of spectral variation when observing sounds of a given vowel, a further observation concerns the nonuniform relation of the lower vowel spectrum to f_o for different frequency ranges of f_o variation. As indicated in the Preliminaries (p. 159) and discussed in Chapter 2.1, vocalises of close vowels showed a marked relation of the lower vowel spectrum to f_o but only from f_o levels above c. 200–300 Hz (depending on vowel quality). Correspondingly, as shown in Chapter 3.3, the same held true for f_o variation in vowel synthesis resulting in formant pattern and spectral shape ambiguity (see also f_o levels and ranges of the comparison of natural vowel sounds with ambiguous *F*-patterns and spectral envelopes in Chapter 3.5). Taking, in addition, our general experiences of the extensive inspection of the Zurich Corpus into account, we assumed that, for sounds of close and close-mid vowels, formant pattern and spectral shape ambiguity may generally occur for an approximately one-octave (or more) increase in f_o, if (and only if) the higher f_o levels of comparison are ≥ c. 300 Hz. (However, for some exceptions of synthesised sounds, see Chapters 3.1, 3.2. and 7.8; see also Chapter 7.7 for cases of sounds of open-mid and open vowels that indicate possible vowel quality shifts due to f_o variation, including an upper f_o frequency of 300 Hz.) In contrast, no differences in the spectral energy distribution and the related maxima may be manifest for one-octave differences of f_o if all f_o levels of compared sounds are below c. 250 Hz. To some extent, we expected that the same holds true for sounds of close-mid vowels, above all, if all f_o levels of the compared sounds are below c. 200 Hz.

Based on indications of an earlier study on this matter, a corresponding new documentary study was conducted in the context of the present treatise: Inspecting voiced sounds of the Zurich Corpus produced by men in nonstyle mode with medium vocal effort in V context at f_o ≤ 250 Hz, for each of the long close and close-mid Standard German vowels, sound pairs of individual speakers and related spectra and estimated *F*-patterns were investigated, and a sample of numerous pairs per vowel was compiled for which (i) the f_o levels differed by approximately one octave or more, but (ii) the spectral envelope and the estimated *F*-patterns < 1 kHz did not indicate marked differences. The sound sample investigated was limited to fully recognised sounds (100% vowel recognition rate according to the standard listening test conducted when creating the corpus, matching vowel intention). As a further selection criterion, for each of the two sounds of a pair, vowel quality was investigated in resynthesis by the author using the KlattSyn tool, resynthesis based on the estimated *F*-pattern of a sound but applying both f_o levels of a pair. Sounds were selected for which no marked vowel quality shift occurred in resynthesis for both f_o levels applied.

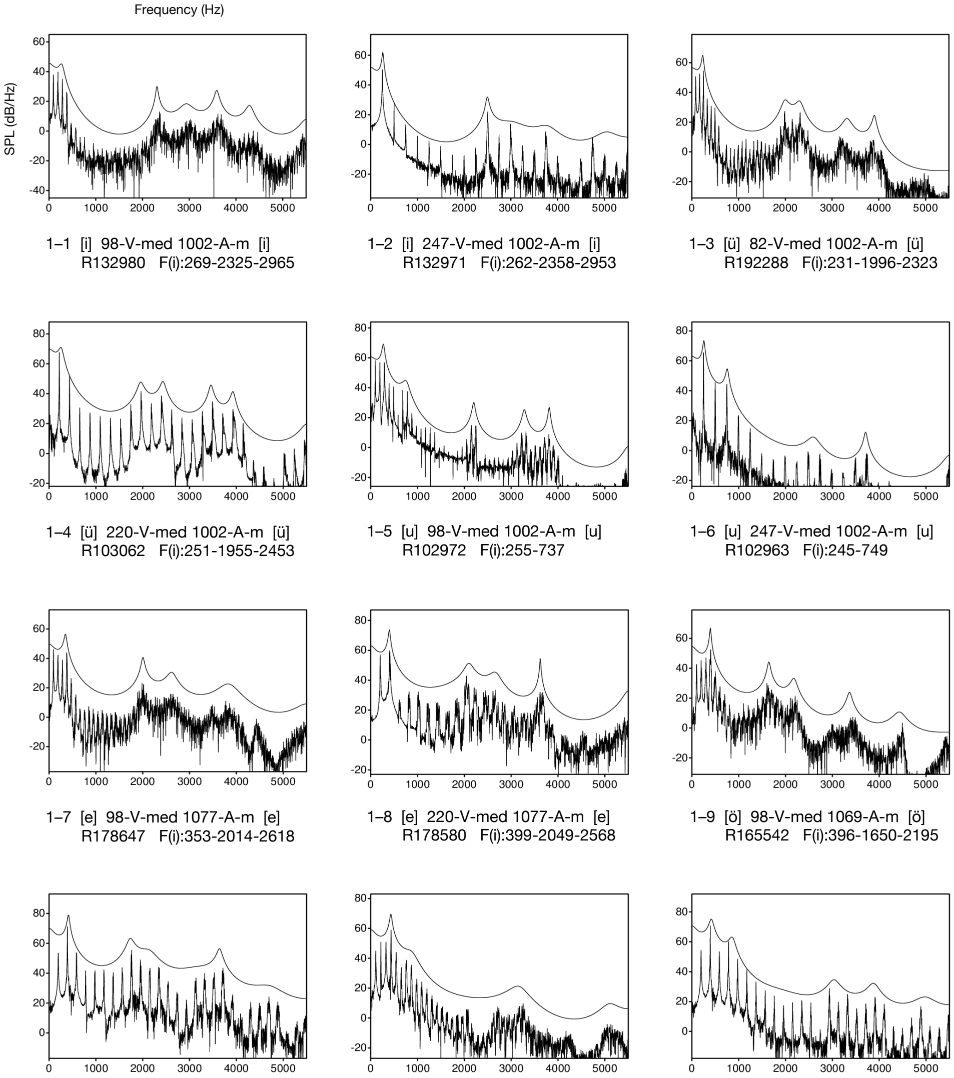
As a result of this inspection, a sample of numerous sound pairs of close and close-mid vowels produced by single speakers was compiled that fulfilled the above selection conditions: These sound pairs were in strong contrast to pronounced spectral differences when including sounds with f_o levels markedly surpassing 250 Hz, as shown in Chapters 2 and 3, since they showed no marked differences in the vowel-related spectral energy distribution and the related spectral maxima despite approximate one-octave differences of f_o. In parallel, no marked vowel quality shifts were found for the resynthesised sounds when the f_o of one sound of a pair was switched with that of the other, neither for an upward nor for a downward f_o switch of one octave or more (author's estimate).

On this basis, an exemplary documentation and illustration of the phenomenon was created for this treatise in the form of one sound pair per vowel, resulting in six exemplary pairs for the vowels /i, y, u/ and /e, ø, o/. For this reduced sound sample, the vowel recognition of the resynthesised sounds applying both f_o levels of a pair was investigated further in a listening test according to the standard procedure of the corpus and involving the five standard listeners. The corresponding recognition results confirmed the estimate of the author. In these terms, the six exemplary sound pairs shown in Figure 1 (compared with the sounds documented in Chapter 3) illustrate that the relation of the lower vowel spectrum to f_o depends on the frequency range of f_o variation.

For references, details of method and results, extended discussion and documentation of results (table including sound links), see the Materials, Chapter M7.4. For a cross-examination of the resynthesis, use the KlattSyn tool in the corpus, taking into account the parameter settings for LPC analysis as given in the Materials.

Figure 1. Sound pairs of the close and close-mid vowels /i, y, u/ and /e, ø, o/ produced by men with f_o variation of approximately one octave ≤ 250 Hz, f_o variation not affecting estimated formant patterns and spectral envelopes. Extracts of Chapter M7.5, Table 1 (see this table for the estimated *F*-patterns and related LPC parameters of acoustic analysis). Six pairs of sounds (intra-speaker comparisons) of the close and close-mid vowels /i, y, u/ and /e, ø, o/ are shown, the sounds produced by men at f_o levels that differ by approximately one octave or more, all f_o levels of comparison being ≤ 250 Hz. Contrary to sound comparisons including sounds at higher f_o levels above 250 Hz, the sounds of the pairs presented exhibit similar vowel-related spectral peak patterns and envelopes despite marked f_o variation, illustrating the nonuniform character of the relation of the (lower) vowel spectrum to f_o with regard to the frequency range of f_o variation. [C-07-05-F01] ↗

Figure 1. Sound pairs of the close and close-mid vowels /i, y, u/ and /e, ø, o/ produced by men with fo variation of approximately one octave ≤ 250 Hz, fo variation not affecting estimated formant patterns and spectral envelopes. [C-07-05-F01]



1–1 [i] 98-V-med 1002-A-m [i]
R132980 F(i):269-2325-2965

1–2 [i] 247-V-med 1002-A-m [i]
R132971 F(i):262-2358-2953

1–3 [ü] 82-V-med 1002-A-m [ü]
R192288 F(i):231-1996-2323

1–4 [ü] 220-V-med 1002-A-m [ü]
R103062 F(i):251-1955-2453

1–5 [u] 98-V-med 1002-A-m [u]
R102972 F(i):255-737

1–6 [u] 247-V-med 1002-A-m [u]
R102963 F(i):245-749

1–7 [e] 98-V-med 1077-A-m [e]
R178647 F(i):353-2014-2618

1–8 [e] 220-V-med 1077-A-m [e]
R178580 F(i):399-2049-2568

1–9 [ö] 98-V-med 1069-A-m [ö]
R165542 F(i):396-1650-2195

1–10 [ö] 196-V-med 1069-A-m [ö]
R165535 F(i):418-1761-2173

1–11 [o] 110-V-med 1069-A-m [o]
R165476 F(i):411-831

1–12 [o] 196-V-med 1069-A-m [o]
R165470 F(i):418-872

7.6 The Role of Vowel Quality With Respect to the Relation of the Lower Vowel Spectrum to f_o

As discussed in the second and third main chapters, the relation of the lower vowel spectrum to f_o – and, with it, formant pattern and spectral shape ambiguity – does not only differ for different frequency ranges of f_o variation but also for different vowel qualities and the individual course of the spectral envelope or the harmonic configuration of sounds of a vowel. These two additional aspects of the nonuniform character of the vowel spectrum are brought into focus in this chapter and the following one.

Concerning the first aspect, when investigating vocalises, the relation of the lower vowel spectrum to f_o proved to be dependent on vowel qualities (see Chapter 2.1), with the most pronounced differences observed when comparing sounds of close and open vowels. Similarly, formant pattern and spectral shape ambiguity occurred far less often for sounds of the open vowel /a/ than for close and close-mid vowel sounds (see Chapters 3.5 and 3.6). In order to exemplify the nonuniform relation of the lower vowel spectrum to f_o concerning vowel quality and its impact on formant pattern and spectral shape ambiguity, a corresponding documentary study was conducted: Inspecting the vocalises of /u/ and /a/ of a man, a woman and a child presented in Chapter 2.1, for each of these vowels and each speaker, two sounds were selected with a difference in their f_o levels of one octave at a minimum and with the higher f_o levels of sound comparison exceeding 400 Hz. As a result, three sound pairs per vowel produced by the three speakers were created.

For these sound pairs, the spectra were visually compared, and the spectral envelope differences < 1.5 kHz were assessed in terms of being marked or marginal. In addition, the relation of the vowel spectrum to f_o was also investigated in vowel resynthesis: All sounds were resynthesised based on both their estimated *F*-patterns and their calculated average f_o. Further, the sound of a pair produced at higher f_o was also resynthesised based on its estimated *F*-pattern but applying the lower f_o level of its opposing sound. The vowel quality of these resynthesised sounds was then tested in a listening test according to the standard procedure of the Zurich Corpus and involving the five standard listeners, with an additional test specification: Each prompt consisted of the original natural sound followed by one of the two resynthesised replicas (separated by a 0.5 sec. pause), and the listeners were asked to label the vowel quality of the second sound.

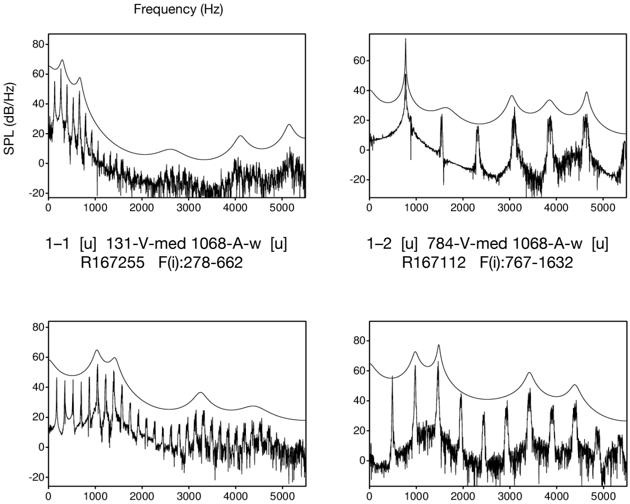
According to the spectral comparison and the vowel recognition results, for all three natural sound pairs of /u/, an increase in f_o levels of one octave or more resulted in a pronounced spectral variation < 1.5 kHz (marked differences in general spectral energy distribution and spectral peaks), and Klatt resynthesis based on the estimated *F*-patterns of the natural reference sounds produced at higher f_o levels but applying the lower f_o level of the opposing sound of the pair in question resulted in a marked vowel quality shift in a close–open direction, the shift including non-adjacent vowel qualities for the sounds of the adult speakers. In contrast, no comparable indication of a pronounced spectral variation or of a distinct vowel quality shift in resynthesis was found for the sounds of /a/.

In these terms, the extracts of the vocalises of the three speakers discussed here exemplify the nonuniform relation of the lower vowel spectrum to f_o concerning vowel quality and its impact on formant pattern and spectral shape ambiguity. Figure 1 illustrates this nonuniform relation for the two sound pairs of /u/ and /a/ produced by the woman.

For references, details of method and results, extended discussion (including the question of estimating *F*-patterns despite middle and higher f_o levels of some of the sounds and the resulting methodological estimation problem) and documentation of results (table including sound links), see the Materials, Chapter M7.6. For a cross-examination of the resynthesis, use the KlattSyn tool in the corpus, taking into account the parameter settings for LPC analysis as given in the Materials.

Figure 1. Sound pairs of /u/ and /a/ produced by a woman illustrating the nonuniform character of the relation of the lower vowel spectrum to f_o. Extract of Chapter M7.6, Table 1 (see sound pairs 2a and 2b in this table). Sounds 1 and 2 = comparison of two sounds of /u/, for which a marked increase in f_o resulted in a pronounced spectral variation < 1.5 kHz, and for which Klatt resynthesis based on the estimated *F*-pattern of the natural reference sound produced at a higher f_o level but applying the lower f_o level of the opposing sound of the pair resulted in a strong vowel quality shift in a close–open direction (according to the labelling majority, the shift was from /u/ to the vowel boundary of /a/ and /ɛ/). Sounds 3 and 4 = corresponding comparison of two sounds of /a/, for which no pronounced spectral variation < 1.5 kHz and no vowel quality shift in resynthesis were found.
[C-07-06-F01] ↗

Figure 1. Sound pairs of /u/ and /a/ produced by a woman illustrating the nonuniform character of the relation of the lower vowel spectrum to fo. [C-07-06-F01]



1–1 [u] 131-V-med 1068-A-w [u]
R167255 F(i):278-662

1–2 [u] 784-V-med 1068-A-w [u]
R167112 F(i):767-1632

1–3 [a] 175-V-med 1068-A-w [a]
R163394 F(i):1041-1416

1–4 [a] 494-V-med 1068-A-w [a]
R163384 F(i):1015-1497

7.7 The Role of the Fine Structure of Spectral Energy Distribution With Respect to the Relation of the Lower Vowel Spectrum to f_o

Besides frequency ranges of f_o and vowel qualities, the fine structure of spectral energy distribution (the individual course of the estimated spectral envelope or the harmonic level configuration) also has an impact on the relation of the lower vowel spectrum to f_o, as was indicated in the investigation of the formant pattern and spectral shape ambiguity phenomenon (see Chapter 3; also consider the vocal effort-related spectral differences discussed in Chapter 5.4). In order to exemplify this impact, too, a corresponding documentary study was conducted: Based on the inspection of voiced sounds of /ɛ/ and /a/ of the Zurich Corpus produced by men and women in nonstyle mode with low or medium vocal effort in V context at calculated $f_o \leq 250$ Hz, and investigating their resynthesis based on their estimated *F*-patterns but applying two f_o levels, the level of the natural sound and a level approximately one octave higher, two sound samples per vowel were compiled. The first sample consisted of natural sounds for which, according to the author's estimate, vowel quality was maintained in resynthesis (using the KlattSyn tool, with default parameters), applying both lower and higher f_o levels. The second sample consisted of sounds for which vowel quality shifted in an open–close direction in resynthesis with increasing f_o levels. All natural sounds compiled were fully recognised (100% vowel recognition rate according to the standard listening test conducted when building up the corpus, matching vowel intention). Natural sounds produced in a lower frequency range of f_o were investigated for two reasons: Firstly, for these sounds, the estimation of *F*-patterns (to which resynthesis related) and spectral envelopes is much less problematic than for sounds produced at middle and higher f_o levels; secondly, in the previous experiments reported in this treatise, no systematic vowel quality shifts were observed due to a one-octave increase in f_o for sounds of these vowels produced in a lower frequency range of f_o.

As a result of this inspection and confirming earlier findings, numerous sounds of /ɛ/ produced at f_o levels ≤ 250 Hz were found for both resynthesis conditions, that is, maintained or shifted vowel quality as a result of an approximate one-octave increase in f_o. However, only a limited number of sounds of /a/ were found for the second resynthesis condition (occurring vowel quality shifts in an open–close direction in resynthesis). On the basis of this observation and as an extract of the above sample, for each of the two vowels and a further limited f_o range

> 160 Hz for the natural sounds, an exemplary documentation of the phenomenon in question was created in the form of a comparison of two natural sounds and their resynthesised replicas with maintained vowel quality for the replicas independent of f_o variation, and two natural reference sounds and their resynthesised replicas with vowel quality shifts for the replicas dependent on f_o variation. For this reduced sample of sounds (four natural sounds per vowel and eight sounds in total), vowel recognition of the resynthesised sounds (lower f_o level = calculated f_o of the natural sounds applied in resynthesis, higher f_o level = 300 Hz; 16 resynthesised sounds in total) was further investigated in a listening test according to the standard procedure of the corpus and involving the five standard listeners, the corresponding results confirming the estimate of the author. The eight exemplary sounds are shown in Figure 1, illustrating the impact of the fine structure of the spectral energy distribution on the relation of the vowel spectrum to f_o.

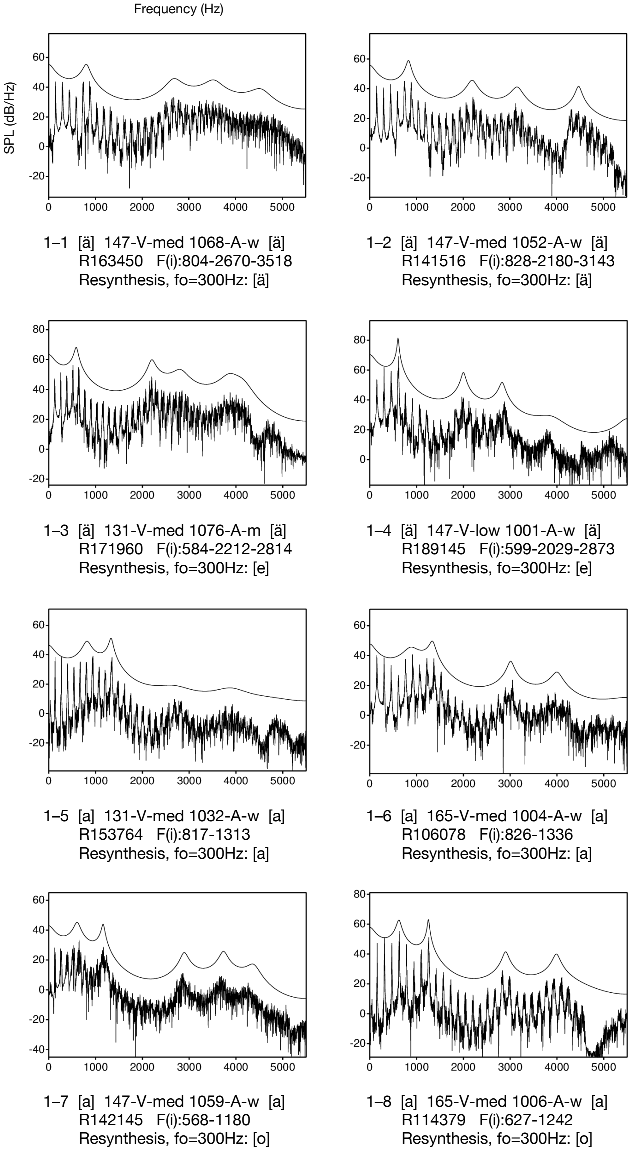
It is noteworthy that, for all presented natural reference sounds, their f_o levels were within a narrow frequency range of 128–157 Hz, and f_o variation in resynthesis was comparable (approximately one octave in a similar frequency range). Thus, the observed difference in perceived vowel quality for the resynthesised sounds with increasing f_o, that is, a vowel quality shift for two sounds but no marked shift for the other two sounds of the same vowel, cannot be attributed to different frequency levels and ranges of f_o of the natural sounds and of f_o variation for their resynthesised replicas. Further, seven of the eight presented natural sounds were produced with medium vocal effort. Consequently, vocal effort variation cannot be considered the main explanation for the vowel recognition differences found for the resynthesised sounds.

In these terms, the selected exemplary sound pairs illustrate that the relation of the vowel spectrum to f_o and its impact on occurring formant pattern and spectral shape ambiguity is further dependent on the fine structure of the spectral energy distribution of a vowel sound.

For references, details of method and results, extended discussion and documentation of results (table including sound links), see the Materials, Chapter M7.7. For a cross-examination of the resynthesis, use the KlattSyn tool in the corpus, taking into account the parameter settings for LPC analysis given in the Materials.

Figure 1. Natural sounds of /ɛ/ and /a/ illustrating the impact of the spectral fine structure on the relation of the lower vowel spectrum to f_o. Extract of Chapter M7.7, Table 1. Sounds 1–4 = sounds of /ɛ/ produced at similar f_o levels with different vowel recognition for their resynthesised replicas applying increased f_o of 300 Hz, that is, maintained vowel quality for the replicas of sounds 1 and 2, and vowel quality shifts in an open–close direction for the replicas of sounds 3 and 4. Sounds 5–8 = corresponding sounds of /a/. (For the recognition results of the synthesised replicas, see Chapter M7.7, Table 1.)
[C-07-07-F01]

Figure 1. Natural sounds of /ε/ and /a/ illustrating the impact of the spectral fine structure on the relation of the lower vowel spectrum to fo. [C-07-07-F01]



1–1 [ä] 147-V-med 1068-A-w [ä]
R163450   F(i):804-2670-3518
Resynthesis, fo=300Hz: [ä]

1–2 [ä] 147-V-med 1052-A-w [ä]
R141516   F(i):828-2180-3143
Resynthesis, fo=300Hz: [ä]

1–3 [ä] 131-V-med 1076-A-m [ä]
R171960   F(i):584-2212-2814
Resynthesis, fo=300Hz: [e]

1–4 [ä] 147-V-low 1001-A-w [ä]
R189145   F(i):599-2029-2873
Resynthesis, fo=300Hz: [e]

1–5 [a] 131-V-med 1032-A-w [a]
R153764   F(i):817-1313
Resynthesis, fo=300Hz: [a]

1–6 [a] 165-V-med 1004-A-w [a]
R106078   F(i):826-1336
Resynthesis, fo=300Hz: [a]

1–7 [a] 147-V-med 1059-A-w [a]
R142145   F(i):568-1180
Resynthesis, fo=300Hz: [o]

1–8 [a] 165-V-med 1006-A-w [a]
R114379   F(i):627-1242
Resynthesis, fo=300Hz: [o]

7.8 The Role of Vocal Effort Variation With Respect to the Relation of the Lower Vowel Spectrum to f_o

In the context of the impact of the spectral fine structure on the relation of the lower vowel spectrum to f_o, finally, vocal effort variation for natural sounds of a vowel must also be taken into consideration (for corresponding previous indications, see Chapters 5.4 and 7.3). To exemplify this further impact, a corresponding documentary study was conducted: Based on sounds of the vowels /e/ and /o/ of men and women already presented in a previous chapter on vocal effort variation (see Chapter 5.4) and supplemented by additional sounds of the Zurich Corpus, for each of the two vowels and each of the two speaker groups, four sounds were compiled and arranged into two sound pairs. The first pair consisted of lower f_o and high vocal effort for the first sound and higher f_o and low vocal effort for the second sound; inversely, the second pair consisted of lower f_o and low vocal effort for the first sound and higher f_o and high vocal effort for the second sound. Lower calculated f_o levels were in the frequency range of 141–170 Hz for the sounds of men and 206–263 Hz for the sounds of women; higher f_o levels were in the frequency range of 319–351 Hz for the sounds of men and 429–441 Hz for the sounds of women; thus, the f_o difference between the two sounds of a sound pair was approximately one octave. The recognition rate for all selected natural sounds was 100% according to the standard listening test conducted when creating the Zurich Corpus, matching vowel intention.

For all sounds, *F*-patterns were estimated according to the standard acoustic analysis of the corpus, including a visual crosscheck based on spectra, spectrograms and formant tracks. The entire estimated *F*-patterns were used for resynthesis (see next paragraph). However, the spectral comparison was limited to F_1, given an identifiable lowest spectral peak: The first formant is commonly considered as a main indicator of vowel-related spectral characteristics < 1.5 kHz, but, as shown, the lower vowel spectrum is strongly related to f_o.

For all natural sounds produced at lower f_o levels (for which LPC analysis is less problematic than for higher f_o), based on the estimated *F*-patterns and using the KlattSyn tool (default parameters), resynthesis was conducted with a step-by-step f_o increase from the f_o level of the natural sound to the higher level of f_o of the opposing natural sound of the pair in question. According to the author's estimate, an open–close vowel quality shift was triggered by a significantly smaller increase in f_o for sounds produced with low vocal effort than for sounds produced with high vocal effort. In an attempt to demonstrate this phenomenon and to involve the standard listeners of the Zurich Corpus for vowel recognition, in a second step, an upper f_o limit of 250 Hz (sounds of men) or 330 Hz (sounds of women) for f_o variation in resynthesis was then chosen as a default for illustration and exemplification, and resynthesis was again conducted, resulting in a sample of 16 resynthesised sounds. Finally, the vowel quality of these resynthesised sounds was tested according to the standard procedure of the corpus and involving the five standard listeners.

As the main result of the spectral inspection, the pairwise comparison of the natural sounds produced with either a high vocal effort at lower f_o or a low vocal effort at higher f_o showed marginal spectral envelope differences in the $F1$ frequency region, with estimated F_1 differences of < 50 Hz. Conversely, very pronounced spectral envelope differences were found in this frequency region for the pairwise comparison of the sounds produced with a low vocal effort at lower f_o and a high vocal effort at higher f_o, with estimated F_1 differences of 332–483 Hz.

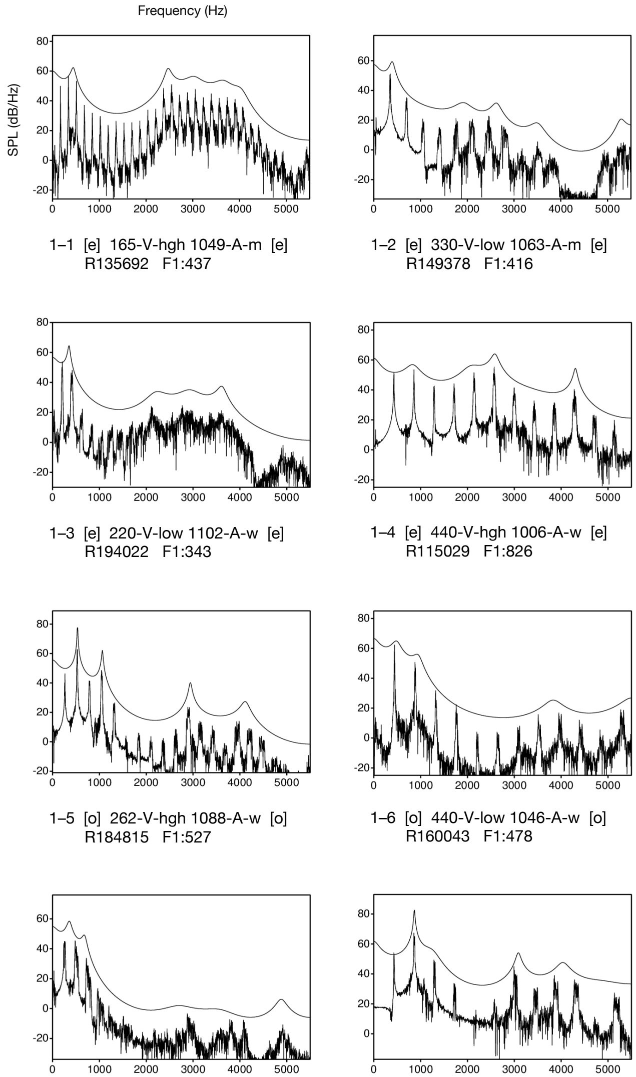
As the main result of investigating vowel recognition, in parallel, the (limited) f_o variation applied in resynthesis had no substantial effect on the recognised vowel quality for the sounds produced with a high vocal effort at lower f_o. Conversely, the effect was pronounced for sounds produced with a low vocal effort at lower f_o.

These two findings highlight the possible impact of marked vocal effort variation on the relation of the lower vowel spectrum to f_o concerning the acoustic representation of vowel quality as a specific aspect of the general impact of the spectral fine structure: The spectra of two sounds, the first produced at lower f_o with high vocal effort and the second produced approximately one octave higher with low vocal effort, can show a similar lower spectral energy distribution, while the spectra of sounds of another pair, the first sound produced at lower f_o with low vocal effort and the second produced approximately one octave higher with high vocal effort, can show very different spectral energy distribution in the lower frequency range. Figure 1 illustrates this phenomenon.

For details of method and results, extended discussion and documentation of results (table including sound links), see the Materials, Chapter M7.8. For a cross-examination of the resynthesis, use the KlattSyn tool in the corpus, taking into account the parameter settings for LPC analysis given in the Materials.

Figure 1. Sound pairs of /e/ and sound pairs of /o/ illustrating the impact of vocal effort variation on the relation of the lower vowel spectrum to f_o. Extract of Chapter M7.8, Table 1 (see Series 1, 2 and 4 in this table). Sounds 1 and 2 of /e/ = comparison of sounds produced by men with a high vocal effort at lower f_o and with a low vocal effort at higher f_o; no indication was found for a substantial spectral envelope difference < 1.5 kHz, and resynthesis of the first sound applying f_o of 250 Hz did not trigger a marked vowel quality shift. Sounds 3 and 4 of /e/ = comparison of sounds produced by women with a low vocal effort at lower f_o and with a high vocal effort at higher f_o; spectral envelope difference < 1.5 kHz proved to be pronounced, and resynthesis of the first sound applying f_o of 330 Hz triggered a marked vowel quality shift in an open–close direction. Sounds 5 and 6, and 7 and 8 = corresponding comparisons for sounds of /o/ produced by women, with f_o levels of 330 Hz in resynthesis.
[C-07-08-F01] ↗

Figure 1. Sound pairs of /e/ and sound pairs of /o/ illustrating the impact of vocal effort variation on the relation of the lower vowel spectrum to fo. [C-07-08-F01]



1–1 [e] 165-V-hgh 1049-A-m [e]
R135692 F1:437

1–2 [e] 330-V-low 1063-A-m [e]
R149378 F1:416

1–3 [e] 220-V-low 1102-A-w [e]
R194022 F1:343

1–4 [e] 440-V-hgh 1006-A-w [e]
R115029 F1:826

1–5 [o] 262-V-hgh 1088-A-w [o]
R184815 F1:527

1–6 [o] 440-V-low 1046-A-w [o]
R160043 F1:478

1–7 [o] 262-V-low 1048-A-w [o]
R130212 F1:382

1–8 [o] 440-V-hgh 1006-A-w [o]
R115027 F1:848

7.9 Conclusion

In this main chapter, different types of nonuniform spectral variation for sounds of a given vowel are revisited, and exemplary documentation is provided. According to earlier indications and the documentation presented here, these variation types can be summed up in two general statements: For natural sounds, the relation of recognised vowel quality and vowel-related spectral peak number and peak structure is nonuniform, and this also holds true for the relation of recognised vowel quality, the lower vowel spectrum and f_o.

Nonuniform relation of recognised vowel quality and vowel-related spectral peak number and peak structure: If sound spectra of given vowels show distinct spectral peaks in the vowel-related frequency range, then the number of these peaks is not constant among the sounds and, up to now, no general and robust method exists for relating different spectral peaks of sounds of a vowel to each other. Above all, as shown, the "centre of gravity" concept (auditory spectral averaging within a frequency range of 3–3.5 Bark) does not account for many of the occurring spectral manifestations in question.

If sounds of a back vowel produced at lower or middle f_o manifesting two distinct spectral peaks < 1.5 kHz are compared with other sounds of that vowel produced at middle or higher f_o manifesting only a single peak, inverse relative spectral maxima and minima can occur, that is, a relative minimum in between the two peaks for the two-peak sounds and, at that minimum frequency, a single spectral maximum for the single-peak sounds. Inversions also occur for sounds of /a/, but they are not systematically related to f_o differences. Cases of inversions were also observed for sounds of front vowels in their upper vowel-specific frequency range > 1.3 kHz, but they often related to marked vocal effort variations, in contrast to the inversions found when comparing sounds of back vowels and /a/. Again, up to now, no general and robust method exists for relating these different spectral peaks of sounds of a vowel to each other.

Further, besides sounds with a varying number of vowel-related spectral peaks, numerous sounds do not exhibit a distinct peak structure in the entire vowel-related frequency range or vowel-related frequency portions. Spectra of this kind show a flat or sloping vowel-related energy distribution.

In these terms, the relation between the recognised vowel quality and the vowel spectrum is nonuniform with regard to vowel-related spectral peak number or peak structure.

Nonuniform relation of recognised vowel quality, the lower vowel spectrum and f_o: Vocalises above all of the close and close-mid vowels showed a marked and systematic relation of the lower spectrum to f_o for pronounced f_o variation above c. 250 Hz (this limit taken as a rough estimation), resulting in formant pattern and spectral shape ambiguity of vowel sounds. However, this is not the case for a variation of f_o below c. 250 Hz. Vocalises of open-mid vowels also showed a relation of the lower spectrum to f_o for pronounced f_o variation but in a less systematic manner, and the relation was weak for vocalises of /a/. Based on the inspection of vocalises of natural vowel sounds, this phenomenological finding was replicated in a vowel (re-)synthesis.

Further investigating the relation between recognised vowel quality, the lower vowel spectrum and f_o for sounds of /a/ and /ɛ/, depending on the spectral energy distribution, pronounced f_o variation had an impact on the acoustic representation of vowel quality for some sounds but not for others, even if vocal effort and the levels and variation extent of f_o of sound production were similar.

Finally, as shown, natural sounds of a vowel produced with a high vocal effort at lower f_o can manifest similar spectral envelopes < 1.5 kHz as those produced with a low vocal effort at higher f_o. Conversely, very pronounced spectral envelope differences can occur for comparisons of sounds of a vowel produced with a low vocal effort at lower f_o and a high vocal effort at higher f_o, given comparable f_o ranges and f_o differences investigated.

In these terms, the relation of the lower spectrum to f_o is nonuniform with regard to frequency ranges and levels of f_o variation, vowel quality (above all, vowel openness) and spectral fine structure (independent or dependent on vocal effort variation) of vowel sounds.

In the Preliminaries, aspects of this nonuniform spectral representation of vowel quality were discussed in the context of a falsification of the prevailing theory that formant patterns and spectral shapes are vowel-specific. Here, based on the new sound sample of the Zurich Corpus, new analyses and new exemplary documentation (supplemented with direct access to the sounds and to tools for sound playback, resynthesis and synthesis, and sound filtering for the purpose of verification and replication), the nonuniform character of the vowel spectrum is discussed from the perspective of a future acoustic theory of the vowel: We conclude that the extent and nonuniform character of spectral variation found for sounds of a given vowel obstructs the formulation of a general concept of relating recognised vowel quality to an average spectral shape even if f_o (and/or pitch) is included in that concept. According to this conclusion, another approach is needed to assess vowel-related acoustic characteristics, and attempts to develop such an approach must take into account the described spectral variation as a part of its most important basis. Below, this matter is addressed in more detail (see the excursus on vowel quality and harmonic spectrum; see also Chapter 10.5).

8 Vowel Recognition of Filtered Vowel Sounds

8.1 Low-Pass Filtering of Vowel Sounds and Related Vowel Recognition

Our many diverse experiences of the recognition and the acoustic representation of vowel quality, as well as our reading of the literature, have led us to assume not only that vowel quality recognition is related to the pitch level of the sounds but also that the vowel sound is a kind of foreground–background phenomenon (see also the Preliminaries, p. 81; to avoid misunderstandings concerning the term, see the corresponding note in the Introduction and the excursus on vowel quality and harmonic spectrum). This chapter approaches this phenomenon by testing vowel recognition of low-pass (LP) filtered vowel sounds.

In the literature, it has been shown that LP sound filtering or LP filtering-like sound manipulation does not result in a general loss of recognised vowel quality but that it often results in a front–back shift for sounds of front vowels and in an open–close shift for sounds of back vowels and /a/. However, the studies so far have not investigated LP sound filtering and vowel recognition in a systematic way, including all long vowels of a language, different phonation types, different f_o levels of voiced sounds and stepwise filtering out higher frequencies of the sounds. Therefore, a corresponding experiment was conducted: Based on the inspection of the Zurich Corpus, sound samples of a man, a woman and a child were selected that included recognised sounds of all long Standard German vowels produced in V context with voiced, whispered and breathy phonation. The voiced sounds were produced in nonstyle mode with a medium vocal effort at the f_o levels of 131–262–523 Hz (man) and 262–523 Hz (woman and child). As a result, a sample of 104 natural reference sounds was created (40 sounds of the man, 32 sounds of each the woman and the child). All sounds of this sample were LP filtered with CFs of 2640–2370–2100–1840–1570–1310–1050–790–530 Hz. As a consequence, and taking the harmonic structure of all spectra into account, statistical F_2 as reported in the literature for long Standard German vowels were LP filtered at a CF of 1840 Hz for the investigated sounds of /i, e/, a CF of 1310 Hz for sounds of /y, ø, ɛ/, a CF of 1050 Hz for sounds of /a/ and a CF of 530 Hz for sounds of /o, u/. As a result, a sample of 936 filtered sounds was created (360 sounds of the man, 288 sounds of each the woman and the child). Finally, vowel recognition of these sounds was assessed according to the standard procedure of the corpus and involving the five standard listeners.

The results of the experiment provided two main indications. Firstly, for all sounds of front vowels, LP filtering caused a general vowel quality shift in a front–back direction. For most sounds of the close front vowels, this shift was preceded by an unrounded–rounded shift, and for some sounds of the front vowels (above all for sounds of the close-mid front vowels), an additional open back to close back shift was observed. For all sounds of /a/, LP filtering caused shifts in an open–close direction. The same held true for half of the sounds of /o/. Secondly, in the course and within the limits of these general shift directions, depending on vowel qualities, multiple single vowel quality shifts occurred for LP filtering of single natural reference sounds, including some variation regarding the order of single quality shifts for different reference sounds of a vowel. This variation was likely related to phonation types, f_0 levels of the sounds and the course of the harmonic envelope. Table 1 shows the general shift directions observed and, within the limits of these directions, some variations of occurring single vowel quality shifts. Figures 1–7 illustrate the two main findings based on selected series of natural reference sounds and filtered sounds created thereof. Illustrations are given for all vowels investigated except /u/ (no systematic shifts occurred for sounds of this vowel).

The general phenomenon that the vowel quality of the sounds investigated was not lost but changed when higher spectral frequency ranges were deleted supported the notion that vowel sounds are describable in terms of a relation of lower spectral similarity (background) and subsequent spectral difference (foreground; see the Preliminaries, p. 81): If the effect of LP filtering is not looked at in terms of stepwise decreasing CFs but in terms of stepwise increasing CFs, sounds produced as different vowels can be perceived as similar in vowel quality up to a given CF level (related to the qualities in question as well as to the spectral energy distribution of the individual sounds compared for the frequency range up to this level), and the sounds only differ if higher spectral frequencies of the natural reference sounds above that CF level are included.

In the context of the present experiment, special attention should again be paid to the evidence that vowel recognition is not based on vowel-related model patterns of spectral peaks: If, for example, sounds of front unrounded vowels were LP filtered with stepwise decreasing CFs below the frequency range of the second spectral peak of the original sounds (or the frequency range of the related statistical F_2), in many cases, the filtered sounds were at first still recognised as front vowels (but rounded) and then, subsequently, as back vowels (for illustration, see below, Figures 1 and 2). Thus, comparable to unmanipulated natural sounds documented in Chapter 7.3, a pattern of two spectral peaks proved not to be a precondition of vowel recognition, and front vowel qualities of sounds with only one manifest lower peak in the spectrum, the higher peaks of the natural reference sounds being LP filtered, could be recognised clearly. At the same time, this finding indicated that vowel recognition does not stand in direct relation to sound production: Above all, the shifts from front unrounded to front rounded to back vowels in LP filtering found for the sounds of this study demonstrated in a paradigmatic way that the actual resonance characteristics of vowel sound production – and with them, the actual articulator positions – cannot, *in general,* be recognised based on a radiated vowel sound.

The variation regarding individual vowel quality shifts occurring within the limits of the general shift directions indicated that f_o, as well as the individual course of the spectral envelope and/or the individual configuration and frequency distance of the harmonics of the unfiltered reference sounds, has to be taken into account when considering the effect of LP filtering or LP filter-like sound manipulations on vowel recognition. Further, note that some filtered sounds occurred whose vowel recognition either did not accord with the general shift directions (rare in number) and/or differed from the intended vowel quality of the unfiltered natural sounds of /o, u/. This may at least in part be understood as being a result of artefacts of LP filtering, an aspect that also has to be taken into account when considering the effect of LP filtering or filter-like sound manipulations. Note also that an estimation of *F*-patterns was methodologically unsubstantiated for many filtered vowel sounds.

For references, extended background information, details of experimental design, method and results, extended discussion, two separate appendices including a discussion of a study by Stumpf (1926) and longer citations by Ito et al. (2001), and for documentation of results (tables including sound links), see the Materials, Chapter M8.1.

Table 1. General vowel quality shift directions and single quality shifts for LP filtered sounds (summary). Simplified summary of the vowel recognition results as given in Chapter M8.1, Table 2; results are given in relation to the intended and recognised vowel quality of the unfiltered natural reference sounds. Columns 1 and 2 = sounds and vowel quality shift directions (V = intended and recognised vowel quality of the unfiltered natural reference sounds, with 13 reference sounds per vowel investigated; N = number of natural reference sounds for which shifts as given in the table occurred). Columns 3–6 = vowel quality shift directions in LP filtering in reference to the quality and number of the unfiltered natural reference sounds (c–o front = close to open front shift direction; ur–r front = unrounded to rounded front shift direction; front–back = front to back shift direction; o–c back = open to close back shift direction). Columns 7–15 = vowel quality of the unfiltered sounds and occurring vowel quality shifts due to LP filtering. Colour code: Grey = main shift directions for sounds of a vowel; dark blue = vowel quality matching the quality of the unfiltered reference sound; light blue = front qualities differing from the quality of the unfiltered reference sound; light red = open, open-mid and close-mid back qualities differing from the quality of the unfiltered reference sound; dark red = close back quality differing from the quality of the unfiltered reference sound. Note that, for filtered sounds of /e/, the order of vowel qualities in shifts involving the recognition of the close-mid front rounded vowel /ø/ and the open-mid front unrounded vowel /ɛ/ was /ɛ/ before /ø/ (see the qualities marked with "*").
[C-08-01-T01]

Figures 1–7 illustrate the main findings of vowel quality shifts due to LP filtering according to the labelling majority: For each of the vowels /i, e, y, ø, ɛ, a, o/ (in this order), two series of a natural reference sound and selected filtered sounds are shown, illustrating general vowel quality shift directions and, within the limits of these directions, some variation of CF levels causing the shifts and/or of sound-specific individual quality shifts. All figures show sound compilations as extracts of the entire sound sample investigated, documented in Chapter M8.1, Table 1.

Figure 1. Examples of vowel quality shifts for LP-filtered sounds of /i/. Sounds 1–4 = a natural sound of /i/ produced by the man with voiced phonation at an intended f_o of 131 Hz and three LP-filtered sounds with CFs of 2640–1840–790 Hz. Vowel recognition of natural and filtered sounds = /i–i–y–u/. Sounds 5–8 = a natural sound of /i/ produced by the woman with voiced phonation at an intended f_o of 262 Hz and three LP-filtered sounds with CFs = 2640–2100–1050 Hz. Vowel recognition of natural and filtered sounds = /i–i–y–u/.
[C-08-01-F01]

Figure 2. Examples of vowel quality shifts for LP-filtered sounds of /e/. Sounds 1–5 = a natural sound of /e/ produced by the woman with breathy phonation at a calculated f_o of 266 Hz and four LP-filtered sounds with CFs = 2640–2100–1310–530 Hz. Vowel recognition of natural and filtered sounds = /e–e–ø–o–u/. Sounds 6–10 = a natural sound of /e/ produced by the woman with voiced phonation at an intended f_o of 262 Hz and four LP-filtered sounds with CFs = 2640–1840–1050–530 Hz. Vowel recognition of natural and filtered sounds = /e–e–ø–o–o/.
[C-08-01-F02]

Figure 3. Examples of vowel quality shifts for LP-filtered sounds of /y/. Sounds 1–4 = a natural sound of /y/ produced by the child with voiced phonation at an intended f_o of 262 Hz and three LP-filtered sounds with CFs = 2640–2100–1050 Hz. Vowel recognition of natural and filtered sounds = /y–y–y–u/. Sounds 5–8 = a natural sound of /y/ produced by the woman with whispered phonation and three LP-filtered sounds with CFs = 2640–1570–530 Hz. Vowel recognition of natural and filtered sounds = /y–y–ø–u/.
[C-08-01-F03] ↗

Figure 4. Examples of vowel quality shifts for LP-filtered sounds of /ø/. Sounds 1–5 = a natural sound of /ø/ produced by the woman with breathy phonation at a calculated f_o of 268 Hz and four LP-filtered sounds with CFs = 2640–1840–1050–530 Hz. Vowel recognition of natural and filtered sounds = /ø–ø–ø–o–u/. Sounds 6–10 = a natural sound of /ø/ produced by the child with voiced phonation at an intended f_o of 262 Hz and four LP-filtered sounds with CFs = 2640–1840–1050–530 Hz. Vowel recognition of natural and filtered sounds = /ø–ø–ø–o/ and vowel confusion for the last sound (with individual assignments by the five listeners = e, e, ø, o, o).
[C-08-01-F04] ↗

Figure 5. Examples of vowel quality shifts for LP-filtered sounds of /ɛ/. Sounds 1–5 = a natural sound of /ɛ/ produced by the woman with voiced phonation at an intended f_o of 262 Hz and four LP-filtered sounds with CFs = 2640–1310–790–530 Hz. Vowel recognition of natural and filtered sounds = /ɛ–ɛ–ɔ–o–u/. Sounds 6–11 = a natural sound of /ɛ/ produced by the woman with whispered phonation and five LP-filtered sounds with CFs = 2640–2370–1570–1050–530 Hz. Vowel recognition of natural and filtered sounds = /ɛ–ɛ–ɛ–a–ɔ–u/.
[C-08-01-F05] ↗

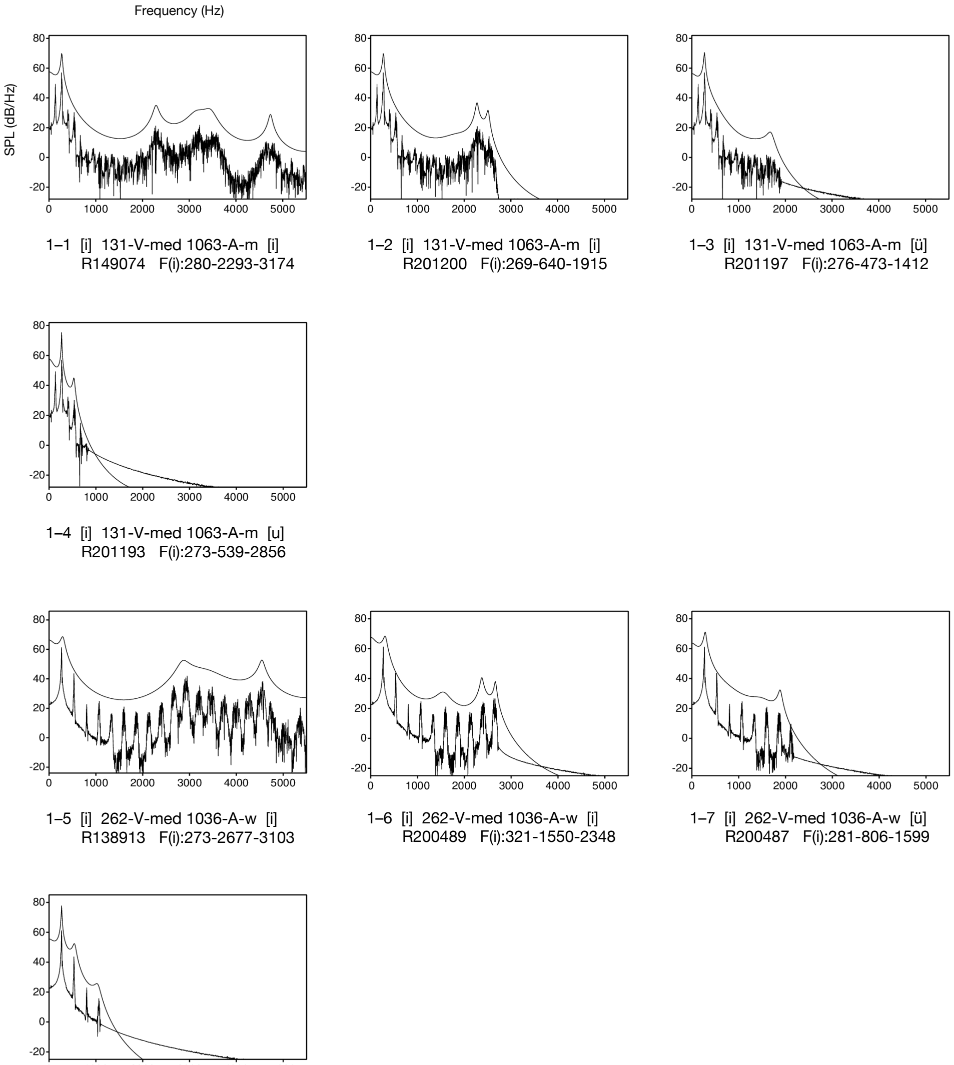
Figure 6. Examples of vowel quality shifts for LP-filtered sounds of /a/. Sounds 1–4 = a natural sound of /a/ produced by the woman with voiced phonation at an intended f_o of 262 Hz and three LP-filtered sounds with CFs = 1570–790–530 Hz. Vowel recognition of natural and filtered sounds = /a–a–ɔ–u/. Sounds 5–10 = a natural sound of /a/ produced by the child with breathy phonation at a calculated f_o of 276 Hz and five LP-filtered sounds with CFs = 1570–1310–1050-790–530 Hz. Vowel recognition of natural and filtered sounds = /a–a–a–(a-ɔ)–o–u/, with no labelling majority for the fourth sound of the series (recognised within the /a-ɔ/ boundary).
[C-08-01-F06] ↗

Figure 7. Examples of vowel quality shifts for LP-filtered sounds of /o/. Sounds 1–4 = a natural sound of /o/ produced by the child with breathy phonation at a calculated f_o of 330 Hz and three LP-filtered sounds with CFs = 1050–790–530 Hz. Vowel recognition of natural and filtered sounds = /o–o–o–u/. Sounds 5–8 = a natural sound of /o/ produced by the man with voiced phonation at an intended f_o of 131 Hz and three LP-filtered sounds with CFs = 1050–790–530 Hz. Maintained vowel recognition of /o/ for all sounds (according to the labelling majority), with the last sound somewhat closer than the preceding sounds (author's estimate).
[C-08-01-F07] ↗

Table 1. General vowel quality shift directions and single vowel quality shifts for LP filtered sounds (summary). [C-08-01-T01]

Sounds and vowel quality shift directions						Single vowel quality shifts								
V	N	c–o (front)	ur–r (front)	front–back	o–c (back)	i	y	e	ø	ɛ	a	ɔ	o	u
i	13	1	10	13	1	i	y							u
						i	y						o	u
						i		e						u
						i								u
e	13	1	13	13	6			e	ø				o	u
								e	ø				o	
								e	ø					u
								e	ø*	ɛ*		ɔ		u
y	13	4	0	13	1		y	e	ø				o	u
							y		ø					u
							y							u
ø	13	–	–	13	6				ø				o	u
									ø					u
									ø				o	
ɛ	13	–	–	13	13					ɛ	a	ɔ	o	u
										ɛ	a	ɔ		u
										ɛ	a		o	u
										ɛ	a	ɔ	o	
										ɛ	a	ɔ		
										ɛ		ɔ	o	u
a	13	–	–	–	13						a	ɔ	o	u
											a	ɔ		u
											a		o	u
											a			u
											a	ɔ	o	
											a	ɔ		
											a		o	
o	6	–	–	–	6								o	u

Figure 1. Examples of vowel quality shifts for LP-filtered sounds of /i/.
[C-08-01-F01]



1–1 [i] 131-V-med 1063-A-m [i]
R149074 F(i):280-2293-3174

1–2 [i] 131-V-med 1063-A-m [i]
R201200 F(i):269-640-1915

1–3 [i] 131-V-med 1063-A-m [ü]
R201197 F(i):276-473-1412

1–4 [i] 131-V-med 1063-A-m [u]
R201193 F(i):273-539-2856

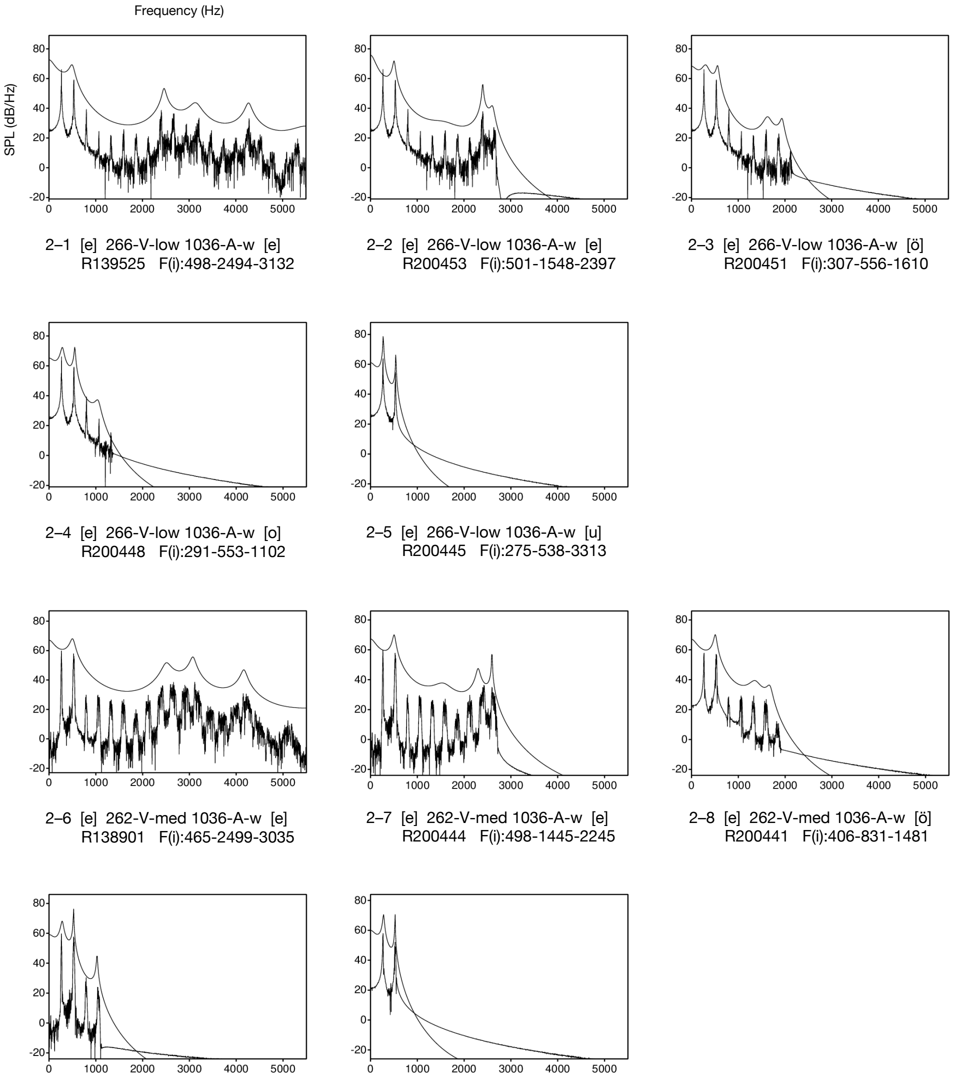
1–5 [i] 262-V-med 1036-A-w [i]
R138913 F(i):273-2677-3103

1–6 [i] 262-V-med 1036-A-w [i]
R200489 F(i):321-1550-2348

1–7 [i] 262-V-med 1036-A-w [ü]
R200487 F(i):281-806-1599

1–8 [i] 262-V-med 1036-A-w [u]
R200483 F(i):266-532-1054

Figure 2. Examples of vowel quality shifts for LP-filtered sounds of /e/.
[C-08-01-F02]



2–1 [e] 266-V-low 1036-A-w [e]
R139525 F(i):498-2494-3132

2–2 [e] 266-V-low 1036-A-w [e]
R200453 F(i):501-1548-2397

2–3 [e] 266-V-low 1036-A-w [ö]
R200451 F(i):307-556-1610

2–4 [e] 266-V-low 1036-A-w [o]
R200448 F(i):291-553-1102

2–5 [e] 266-V-low 1036-A-w [u]
R200445 F(i):275-538-3313

2–6 [e] 262-V-med 1036-A-w [e]
R138901 F(i):465-2499-3035

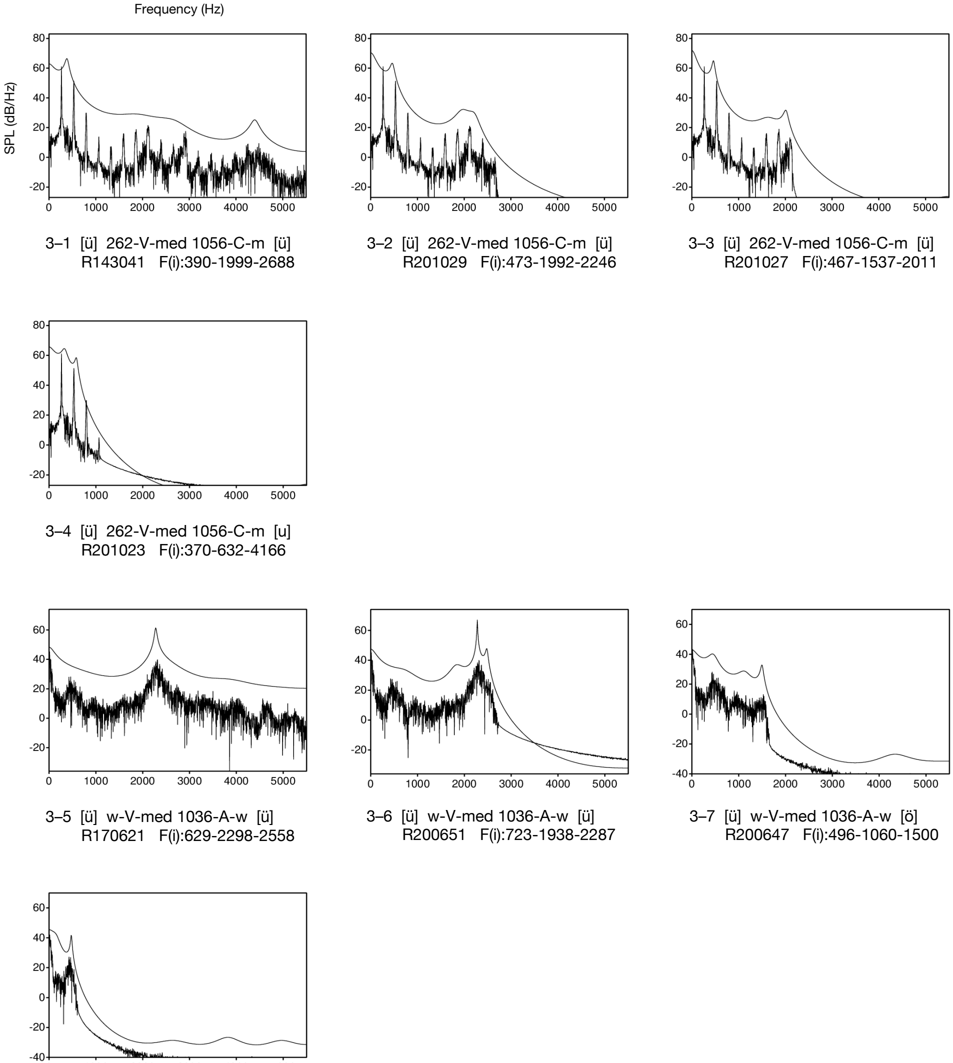
2–7 [e] 262-V-med 1036-A-w [e]
R200444 F(i):498-1445-2245

2–8 [e] 262-V-med 1036-A-w [ö]
R200441 F(i):406-831-1481

2–9 [e] 262-V-med 1036-A-w [o]
R200438 F(i):289-535-1025

2–10 [e] 262-V-med 1036-A-w [o]
R200436 F(i):285-535-3297

Figure 3. Examples of vowel quality shifts for LP-filtered sounds of /y/. [C-08-01-F03]



3–1 [ü] 262-V-med 1056-C-m [ü]
R143041 F(i):390-1999-2688

3–2 [ü] 262-V-med 1056-C-m [ü]
R201029 F(i):473-1992-2246

3–3 [ü] 262-V-med 1056-C-m [ü]
R201027 F(i):467-1537-2011

3–4 [ü] 262-V-med 1056-C-m [u]
R201023 F(i):370-632-4166

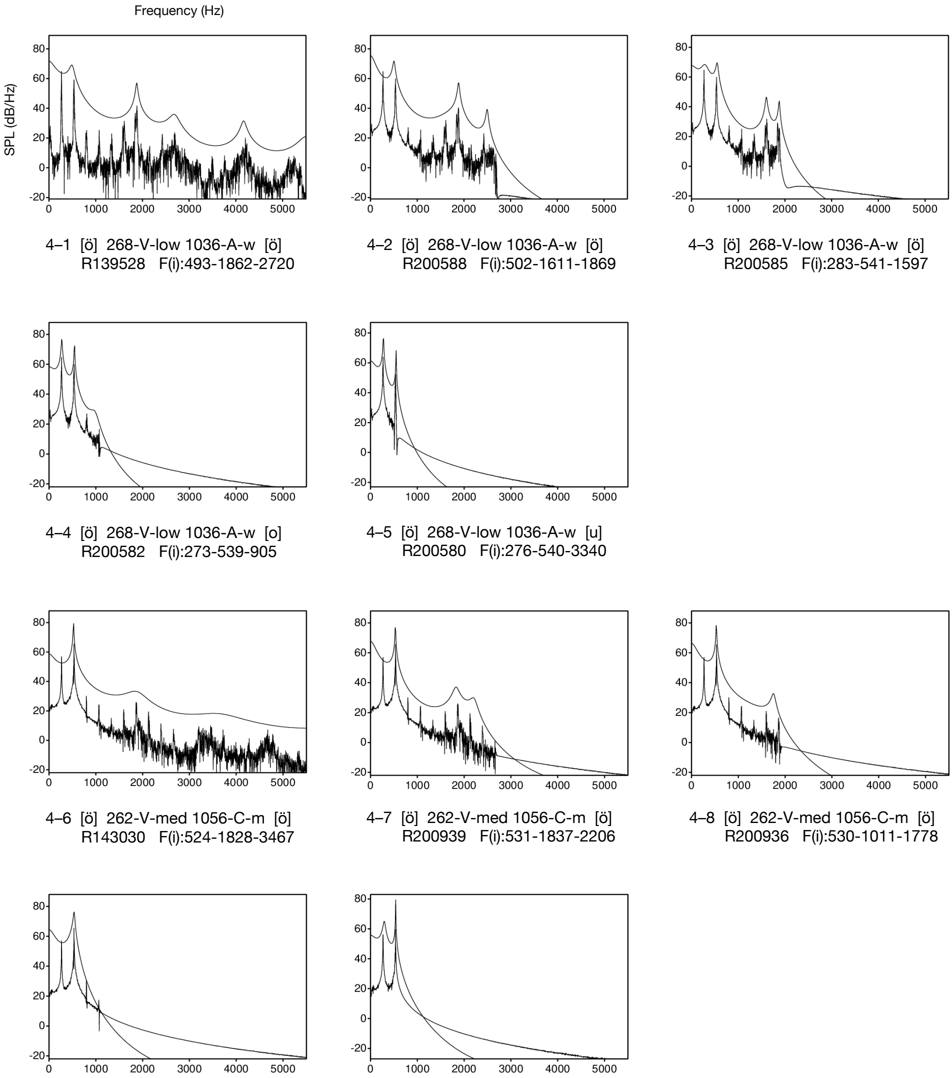
3–5 [ü] w-V-med 1036-A-w [ü]
R170621 F(i):629-2298-2558

3–6 [ü] w-V-med 1036-A-w [ü]
R200651 F(i):723-1938-2287

3–7 [ü] w-V-med 1036-A-w [ö]
R200647 F(i):496-1060-1500

3–8 [ü] w-V-med 1036-A-w [u]
R200643 F(i):237-682-2875

Figure 4. Examples of vowel quality shifts for LP-filtered sounds of /ø/.
[C-08-01-F04]



4–1 [ö] 268-V-low 1036-A-w [ö]
R139528 F(i):493-1862-2720

4–2 [ö] 268-V-low 1036-A-w [ö]
R200588 F(i):502-1611-1869

4–3 [ö] 268-V-low 1036-A-w [ö]
R200585 F(i):283-541-1597

4–4 [ö] 268-V-low 1036-A-w [o]
R200582 F(i):273-539-905

4–5 [ö] 268-V-low 1036-A-w [u]
R200580 F(i):276-540-3340

4–6 [ö] 262-V-med 1056-C-m [ö]
R143030 F(i):524-1828-3467

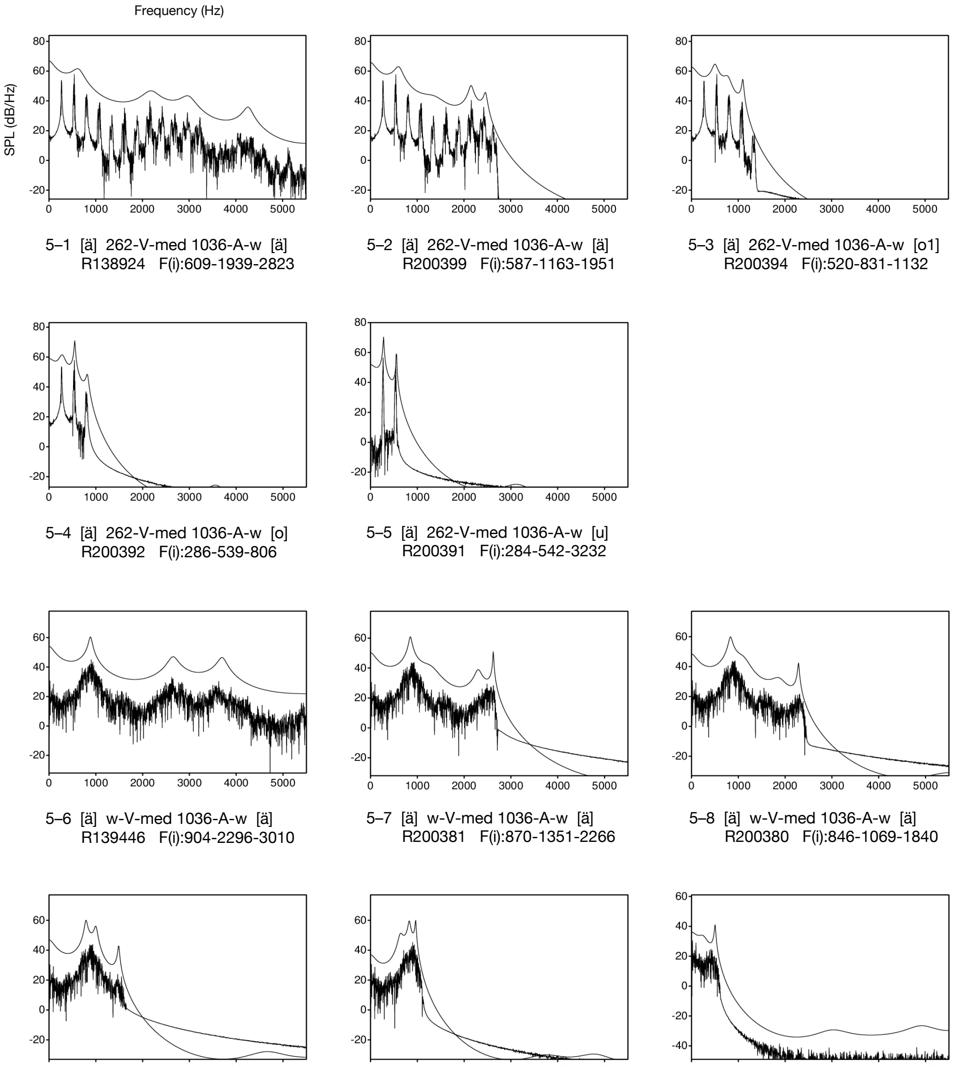
4–7 [ö] 262-V-med 1056-C-m [ö]
R200939 F(i):531-1837-2206

4–8 [ö] 262-V-med 1056-C-m [ö]
R200936 F(i):530-1011-1778

4–9 [ö] 262-V-med 1056-C-m [o]
R200933 F(i):525-568-4575

4–10 [ö] 262-V-med 1056-C-m [–]
R200931 F(i):315-540-3960

Figure 5. Examples of vowel quality shifts for LP-filtered sounds of /ε/.
[C-08-01-F05]



5–1 [ä] 262-V-med 1036-A-w [ä]
R138924 F(i):609-1939-2823

5–2 [ä] 262-V-med 1036-A-w [ä]
R200399 F(i):587-1163-1951

5–3 [ä] 262-V-med 1036-A-w [o1]
R200394 F(i):520-831-1132

5–4 [ä] 262-V-med 1036-A-w [o]
R200392 F(i):286-539-806

5–5 [ä] 262-V-med 1036-A-w [u]
R200391 F(i):284-542-3232

5–6 [ä] w-V-med 1036-A-w [ä]
R139446 F(i):904-2296-3010

5–7 [ä] w-V-med 1036-A-w [ä]
R200381 F(i):870-1351-2266

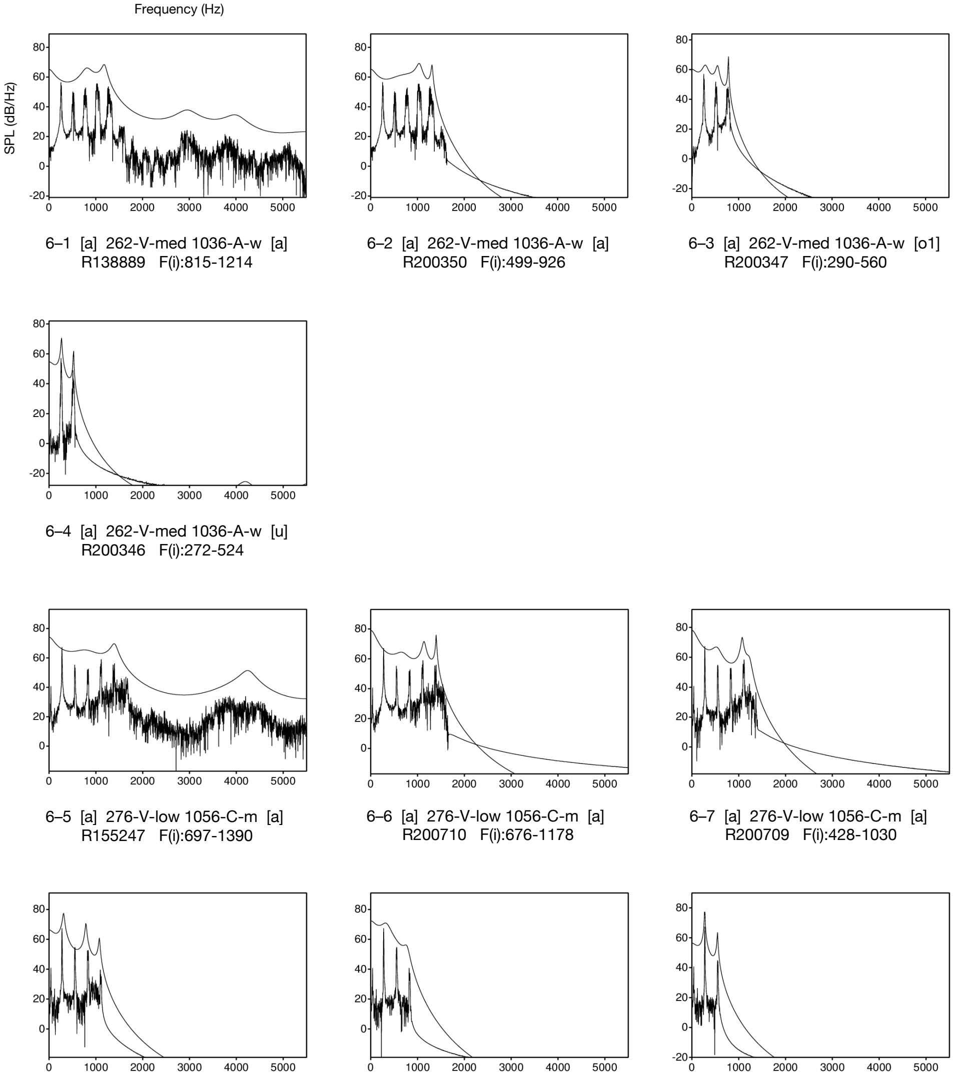
5–8 [ä] w-V-med 1036-A-w [ä]
R200380 F(i):846-1069-1840

5–9 [ä] w-V-med 1036-A-w [a]
R200377 F(i):809-1004-1462

5–10 [ä] -V-med 1036-A-w [o1]
R200375 F(i):624-849-969

5–11 [ä] -V-med 1036-A-w [u]
R200373 F(i):291-492-2867

Figure 6. Examples of vowel quality shifts for LP-filtered sounds of /a/.
[C-08-01-F06]



6–1 [a] 262-V-med 1036-A-w [a]
R138889 F(i):815-1214

6–2 [a] 262-V-med 1036-A-w [a]
R200350 F(i):499-926

6–3 [a] 262-V-med 1036-A-w [o1]
R200347 F(i):290-560

6–4 [a] 262-V-med 1036-A-w [u]
R200346 F(i):272-524

6–5 [a] 276-V-low 1056-C-m [a]
R155247 F(i):697-1390

6–6 [a] 276-V-low 1056-C-m [a]
R200710 F(i):676-1178

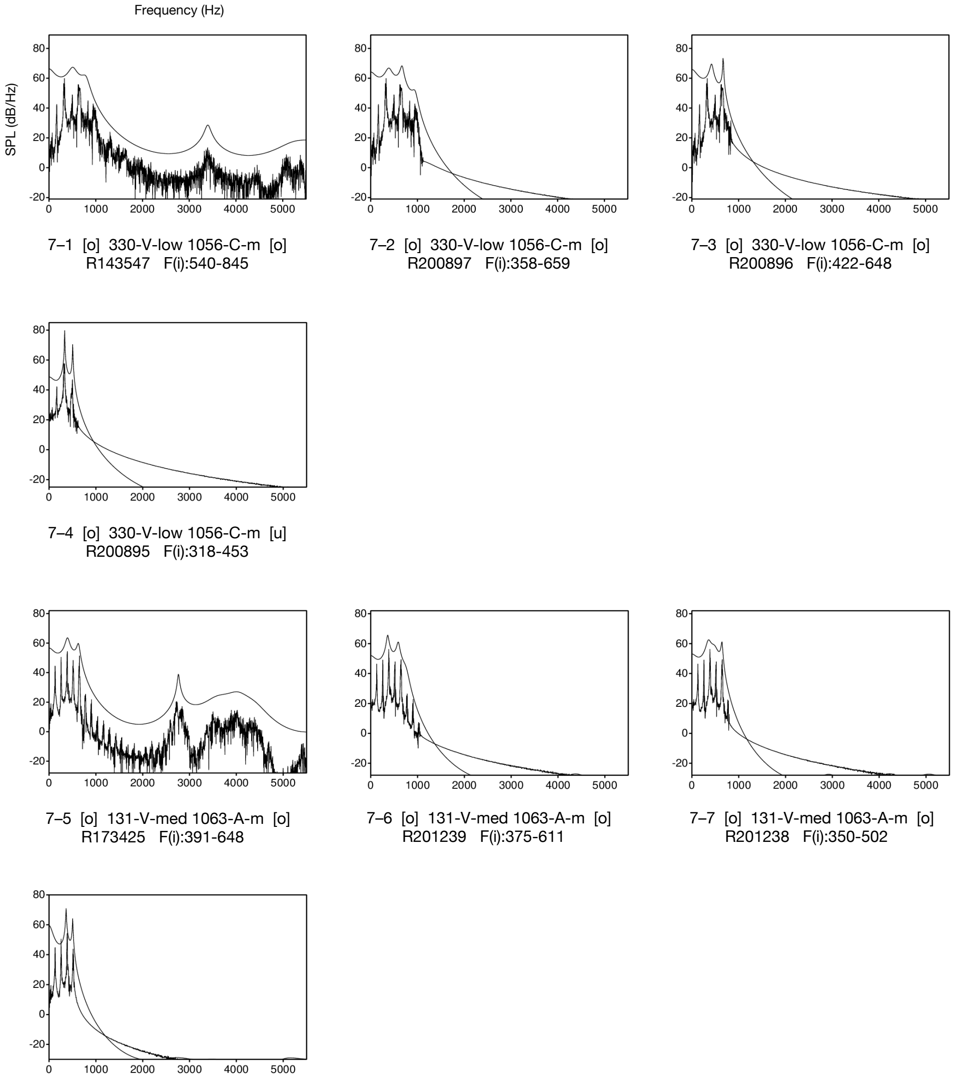
6–7 [a] 276-V-low 1056-C-m [a]
R200709 F(i):428-1030

6–8 [a] 276-V-low 1056-C-m [-]
R200708 F(i):306-782

6–9 [a] 276-V-low 1056-C-m [o]
R200707 F(i):313-505

6–10 [a] 276-V-low 1056-C-m [u]
R200706 F(i):276-541

Figure 7. Examples of vowel quality shifts for LP-filtered sounds of /o/.
[C-08-01-F07]



7–1 [o] 330-V-low 1056-C-m [o]
R143547 F(i):540-845

7–2 [o] 330-V-low 1056-C-m [o]
R200897 F(i):358-659

7–3 [o] 330-V-low 1056-C-m [o]
R200896 F(i):422-648

7–4 [o] 330-V-low 1056-C-m [u]
R200895 F(i):318-453

7–5 [o] 131-V-med 1063-A-m [o]
R173425 F(i):391-648

7–6 [o] 131-V-med 1063-A-m [o]
R201239 F(i):375-611

7–7 [o] 131-V-med 1063-A-m [o]
R201238 F(i):350-502

7–8 [o] 131-V-med 1063-A-m [o]
R201237 F(i):372-500

8.2 High-Pass Filtering of Vowel Sounds and Related Vowel Recognition

Similar to the question of LP filtering of vowel sounds, there is no systematic study investigating vowel recognition of high-pass (HP) filtered sounds in the literature that includes all long vowels of a language, different phonation types, different levels of f_o and stepwise increasing CFs. However, studies on the matter investigating voiced vowel sounds of single speakers produced at a given f_o level indicated (i) vowel quality-specific effects of HP filtering in general, and (ii) in particular, initial close–open and subsequent reverted open–close shifts for HP-filtered sounds of front vowels with stepwise increasing CFs as well as recognisable vowel sounds with HP-filtered spectral energy below c. 2 kHz associated with back–front shifts for filtered sounds of back vowels. Against this background, and to broaden the experiment described in Chapter 6.4, a corresponding HP sound filtering experiment was conducted so as to address vowel quality recognition for sounds for which vowel-related lower spectral frequency ranges are filtered.

For the experiment, the sound sample of the long Standard German vowels of the previous experiment was again used and enlarged by additional voiced sounds of the Zurich Corpus produced by the three speakers in nonstyle mode with medium vocal effort in V context at intended f_o levels of 220–330–440–659 Hz; the sounds selected had obtained the highest recognition rate for the production parameters in question (phonation type, f_o levels) in the standard listening test conducted when creating the corpus. As a result, the enlarged speaker-specific subsamples consisted of one whispered and one breathy sound and of voiced sounds produced at intended f_o levels of 131–220–262–330–440–523–659 Hz for the man (nine sounds per vowel and a total of 72 sounds) and 220–262–330–440–523–659 Hz for the woman and the child (eight sounds per vowel and a total of 64 sounds each). In these terms, an overall sample of 200 natural sounds was investigated.

All sounds of this sample were HP filtered with CFs of 440–660–990–1320 Hz. (Note that the intended f_o of 659 Hz and the CF of 660 Hz are considered equal frequency levels; note also that the f_o levels of breathy sounds are integrated into the scale of intended f_o levels of the voiced sounds.) As a result, a sample of 800 filtered sounds (288 sounds related to the original reference sounds of the man and 256 sounds related to the original reference sounds of each the woman and the child) was created for the vowel recognition test. Notably, for the close vowels investigated, a CF of 440 Hz surpassed the average F_1 of the sounds produced at lower f_o levels as given in many formant statistics for adult speakers. Similarly, for the sounds of the close-mid vowel, a CF of 660 Hz also surpassed the statistical average F_1. For filtered sounds with a CF of 990 or 1320 Hz, no spectral energy represented assumed F_1 differences as given in formant statistics for different vowel qualities. On this basis, concerning vowel recognition, the effect of deleting spectral energy in the frequency region of *F*1 of a vowel could be investigated, in combination with the relation of f_o of the sounds and CFs applied. The same held true for *F*2 of back vowels. Finally, the vowel recognition of the filtered sounds was assessed in a listening test according to the standard procedure of the Zurich Corpus and involving the five standard listeners.

Within the limits of varying vowel recognition rates, vowel boundary recognition and marked changes in sound timbre for the filtered sounds, four main results were obtained.

Firstly, HP filtering of the frequency region of *F*1, generally assumed to be vowel-related, in many cases did not result in a sound for which vowel quality was lost: According to the labelling majority of the vowel recognition test, for most of the HP-filtered sounds with CFs up to 990 Hz, the intended vowel quality either was maintained or shifted to another quality. If it shifted, an initial close–open shift direction was found for most sounds of /i, y, e/ and for some of the sounds of /ø, o, u/, and an initial open–close shift direction was found for a few sounds of /ɛ/ and /a/.

Secondly, for the majority of the sounds of the close front vowels /i, y/ and some of the sounds of the close back vowel /u/ produced at intended f_o of up to 330 Hz and HP filtered with a CF of 440 Hz, the vowel quality shifted and differed from the quality of the unfiltered sounds produced at intended f_o of 440–523–659 Hz (equal to or above the CF applied). A similar effect was found for some of the sounds of the close-mid vowels /e, ø, o/ produced at intended f_o of up to 523 Hz and filtered with a CF of 660 Hz when compared with the unfiltered sounds produced at intended f_o of 659 Hz. Thus, in general terms, the effect of HP filtering of natural close and close-mid vowels depended on the f_o level of the sounds. This finding was to be expected from the many other indications reported that the lower frequency range of the vowel spectrum of natural sounds is related to f_o.

Thirdly, with increasing CFs, initial close–open shifts for filtered sounds of close and close-mid front vowels were, in many cases, reverted back to the intended vowel qualities of the unfiltered sounds (above all for natural sounds of close front vowels) or even inverted from close-mid to close vowels. Thus, it was again demonstrated that numerous natural sounds of close and close-mid front vowels remained recognisable even if the entire frequency range of statistical *F*1 of all vowels of a language was HP filtered, that is, energy in the lower frequency range commonly assumed to be vowel quality-specific was not a general precondition for vowel recognition.

Fourthly, HP filtering sounds with a CF of 1320 Hz caused a back–front shift for some sounds of /o, u/ in strong contrast to the assumed vowel-related resonances of vowel production not being represented in the sound spectra.

Looking at the general shift directions found, if shifts occurred, the effect of HP filtering proved to differ in relation to openness: An initial close–open shift direction was found for close and close-mid vowels and, conversely, an initial open–close shift direction was found for open-mid and open vowels. Further, within the limits of the general shift directions and the role of f_o in this type of filter experiment, CFs and associated vowel quality shifts also varied to some extent for sounds of the same vowel. Hence, once again, the vowel qualities investigated and the individual spectral energy distribution of single sounds of a vowel also have to be accounted for when interpreting and generalising the results.

Table 1 shows the general shift directions and, within the limits of these directions, some sound-specific variation regarding occurring single vowel quality shifts related to the applied CFs and to vowel qualities of the natural reference sounds and their f_o levels. Figures 1 to 7 illustrate the above main findings.

In these terms, and in line with the indications given in the literature, the results of the experiment confirmed vowel quality-specific effects of HP filtering, initial close–open and subsequent reverted open–close shifts above all for HP-filtered sounds of close and close-mid front vowels with stepwise increasing CFs, and recognisable vowel sounds with HP-filtered spectral energy below c. 1.5 kHz associated with back–front shifts for filtered sounds of back vowels. In addition, the results also indicated an HP filter effect that is related to the individual energy distribution of a sound of a vowel. All these findings supported anew the thesis that the vowel sound is a kind of foreground–background phenomenon.

For references, extended background information, details of experimental design, method and results and their documentation (tables including sound links), see the Materials, Chapter M8.2.

Table 1. General vowel quality shift directions and single quality shifts for HP-filtered sounds (simplified summary). Simplified summary of the vowel recognition results as given in Chapter M8.2, Table 1; results are given in relation to the intended and recognised vowel quality of the unfiltered natural reference sounds. Columns 1–3 = sounds (V = intended and recognised vowel quality of the unfiltered natural reference sounds; fo = range of intended f_o of the natural reference sounds, in Hz; CF = CF applied, in Hz). Columns 4–8 = shift directions and related number of natural reference sounds, for which HP filtering caused the shifts (c–o = close–open shift direction; ns = no shift; o–c = open–close shift direction; b–f = back–front shift direction; m = miscellaneous for vowel boundary recognition or vowel confusion). Column 9 = recognised vowel quality of the natural reference sound (V ref; repetition of Column 1). Columns 10–18 = confusion matrix of the vowel recognition results for the HP-filtered sounds (sounds per vowel quality or quality boundaries or vowel confusion; vb = vowel boundary recognition; vc = vowel confusion; fr = front vowel qualities; N = total number of the natural reference sounds investigated). Columns 19–23 = number of reverted or inverted shifts per CF (sounds of front vowels) or back–front shifts (sounds of back vowels; for details see text). Colour code: Dark blue = recognised vowel quality matching the quality of the unfiltered reference sound; light blue = vowel quality shift (with the exception of shifts from unrounded to rounded or vice versa); purple = inverted vowel quality shift from a close-mid to a close vowel quality; red = back–front vowel quality shift.
[C-08-02-T01]

Figure 1. Examples of maintained vowel quality for HP-filtered sounds. A whispered sound of /i/ and a voiced sound of /a/ are shown, each followed by the HP-filtered sounds created thereof with CFs of 440–660–990–1320 Hz, illustrating maintained vowel recognition for all filtered sounds. Extract of Chapter M8.2, Table 2 (see Series 1 in this table).
[C-08-02-F01] ↗

Figure 2. Examples of initial close–open and subsequent reverted open–close vowel quality shifts for HP-filtered sounds of close and close-mid front vowels. Extract of Chapter M8.2, Table 2 (see Series 2 in this table). Two voiced sounds of /i/ and /e/ and the HP-filtered sounds created thereof with CFs of 440–660–990–1320 Hz are shown, illustrating initial close–open and subsequent reverted open–close shifts.
[C-08-02-F02] ↗

Figure 3. Examples of initial close–open and subsequent back–front shifts for HP-filtered sounds of close-mid and close back vowels. Extract of Chapter M8.2, Table 2 (see Series 2 in this table). Two voiced sounds of /o/ and /u/ and the HP-filtered sounds created thereof with CFs of 440–660–1320 Hz (sound of /o/) and 440–1320 Hz (sound of /u/) are shown, illustrating initial close–open and subsequent back–front shifts.
[C-08-02-F03] ↗

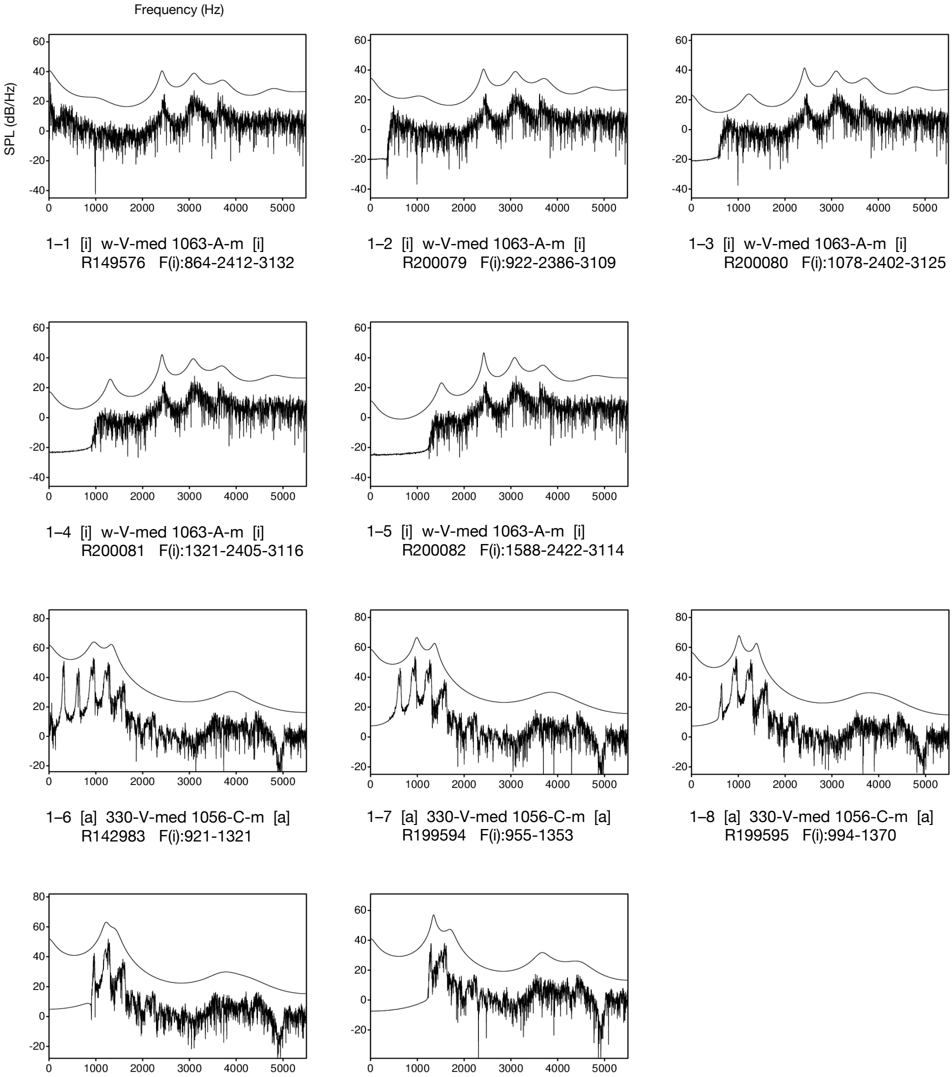
Figure 4. Examples of sounds for which the impact of HP sound filtering on vowel recognition depended on the f_o (and pitch) of sound production. Extract of Chapter M8.2, Table 2 (see Series 3 in this table, sounds of /e/ and /u/). Sounds 1–3 = comparison of a voiced sound of /e/ produced at an f_o below 660 Hz and the filtered sound created thereof applying a CF of 660 Hz (sounds 1 and 2) with a voiced sound of /e/ produced at an f_o of c. 660 Hz, unaffected by this filtering (sound 3). For the first natural sound and its filtered version, HP filtering caused a close–open shift from /e/ to /ɛ/, in contrast to the recognised vowel quality /e/ of the third sound shown. Sounds 4–6 = comparison of a voiced sound of /u/ produced at an f_o below 440 Hz and the filtered sound created thereof applying a CF of 440 Hz (sounds 1 and 2) with a voiced sound of /u/ produced at an f_o of c. 440 Hz, unaffected by this filtering (sound 3). For the first natural sound and its filtered version, HP filtering caused a close–open shift from /u/ to /o/, in contrast to the recognised vowel quality /u/ of the third sound shown. Illustration of the relation of HP sound filtering and f_o concerning their impact on vowel recognition.
[C-08-02-F04]

Figure 5. Further examples of recognisable sounds of front vowels with the entire frequency range of statistical *F*1 of all vowels being HP filtered. Extract of Chapter M8.2, Table 2 (see Series 4 in this table). 12 HP-filtered whispered and voiced sounds of /i/ and /y/, HP filtered with a CF of 1320 Hz. According to the labelling majority of the vowel recognition test, the intended vowel quality of the unfiltered natural reference sounds was maintained for all filtered sounds shown. Illustration of recognisable sounds of close front vowels with the entire frequency range of statistical *F*1 of all vowels being HP filtered.
[C-08-02-F05]

Table 1. General vowel quality shift directions and single quality shifts for HP filtered sounds (simplified summary). [C-08-02-T01]

Sounds			Vowel recognition					Vowel recognition (details)														
			Shift directions					V	Confusion matrix									Rev./inv. shifts				
V	fo	CF	c–o	ns	o–c	–	m	ref	i	y	e	ø	ε/ə	a	vb	vc	N	ε/ə	ö	e	y	i
i	< 440	440	11	3			2	i	3		11				2		16					
	< 660	660	15	2			5		2		12		3		1	4	22					
	all	990	6	12			7		12		3		3		3	4	25					8
	all	1320	1	20			4		20		1					4	25					8
y	< 440	440	15	–			1	y			3	12				1	16					
	< 660	660	13	2			7			2	1	8	4			7	22				2	
	all	990	9	9			7			9		6	3			7	25				6	
	all	1320	2	19			4			19		2				4	25				10	
e	< 440	440	–	16				e			16						16					
	< 660	660	10	11			1				11		10			1	22					
	all	990	14	3			8				3		14		2	6	25			1		
	all	1320	4	7	3		11		3		7		4			11	25			6		3
ø	< 440	440	–	16				ø			1	15					16					
	< 660	660	3	14			5					14	3		1	4	22					
	all	990	7	7	1		10			1		7	5	2	1	9	25		1		1	
	all	1320	2	4	7		12			7		4	1	1		12	25		1		6	
			Shift directions					V	confusion matrix											–		
V	fo	CF	c–o	ns	o–c	–	m	ref	e	ε/ə	a	ɔ	o	u	vb	vc	N					
/ε/	< 440	440		11	2		3	ε	2	11					1	2	16					
	< 660	660		21			1			21					1		22					
	all	990		24			1			24					1		25					
	all	1320		13	3		9		3	13					1	8	25					
a	< 440	440		14	2			a			14	2					16					
	< 660	660		15	3		4				15	3			3	1	22					
	all	990		25							25						25					
	all	1320		15	5		5			5	15					5	25					
			Shift directions					V	confusion matrix									back–front shifts				
V	fo	CF	c–o	ns	o–c	b-f	m	ref	a	ɔ	o	u	–	fr	vb	vc	N	ε/ə	ö	e	y	i
o	< 440	440		16				o			16						16					
	< 660	660	11	9			2			11	9					2	22					
	all	990	3	5	2		15			3	5	2				15	25					
	all	1320	4	–	2	6	13		4			2		6		13	25	1			4	1
u	< 440	440	5	8			3	u			5	8			3		16					
	< 660	660	6	15			1				6	15				1	22					
	all	990	7	16			2		1	2	4	16				2	25					
	all	1320	1	–		7	17		1					7		17	25				7	

Figure 1. Examples of maintained vowel quality for-HP filtered sounds. [C-08-02-F01]



1–1 [i] w-V-med 1063-A-m [i]
R149576 F(i):864-2412-3132

1–2 [i] w-V-med 1063-A-m [i]
R200079 F(i):922-2386-3109

1–3 [i] w-V-med 1063-A-m [i]
R200080 F(i):1078-2402-3125

1–4 [i] w-V-med 1063-A-m [i]
R200081 F(i):1321-2405-3116

1–5 [i] w-V-med 1063-A-m [i]
R200082 F(i):1588-2422-3114

1–6 [a] 330-V-med 1056-C-m [a]
R142983 F(i):921-1321

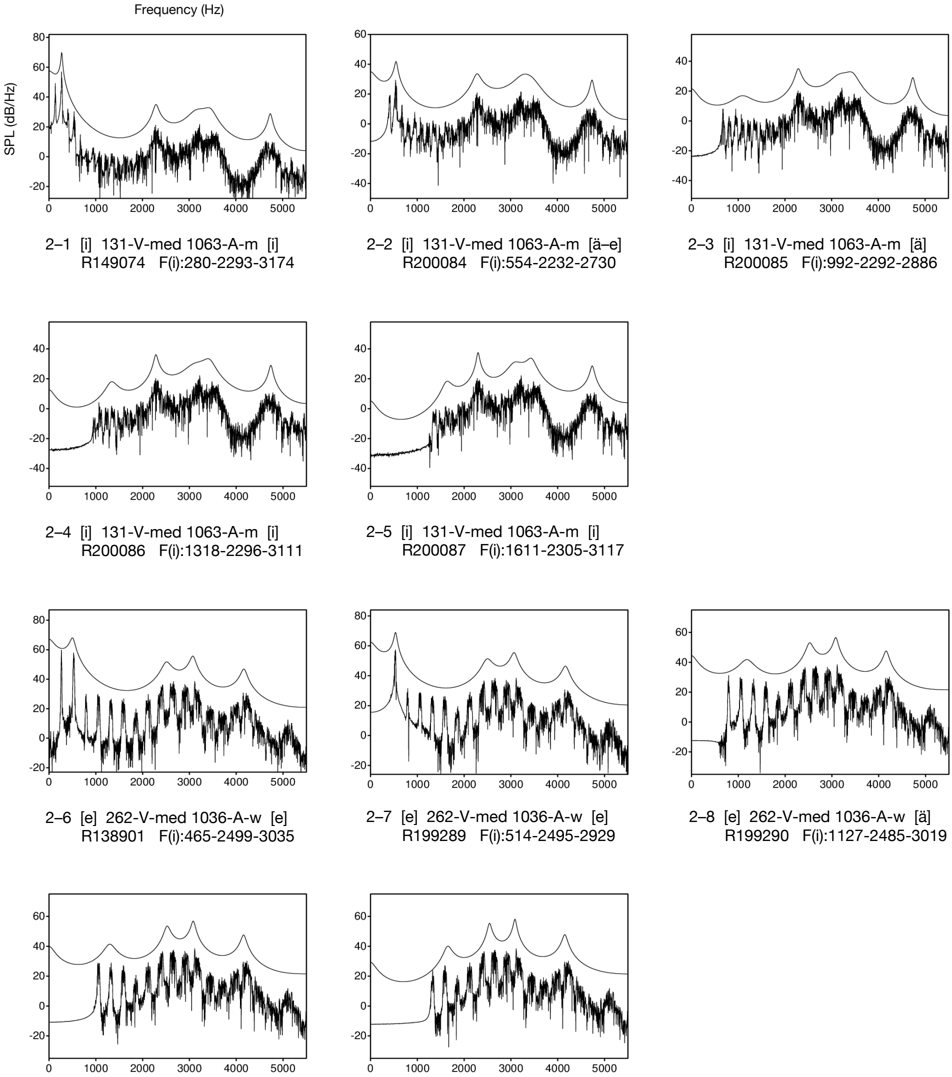
1–7 [a] 330-V-med 1056-C-m [a]
R199594 F(i):955-1353

1–8 [a] 330-V-med 1056-C-m [a]
R199595 F(i):994-1370

1–9 [a] 330-V-med 1056-C-m [a]
R199596 F(i):1219-1481

1–10 [a] 330-V-med 1056-C-m [a]
R199597 F(i):1442-1822

Figure 2. Examples of initial close–open and subsequent reverted open–close vowel quality shifts for HP-filtered sounds of close and close-mid front vowels.
[C-08-02-F02]



2–1 [i] 131-V-med 1063-A-m [i]
R149074 F(i):280-2293-3174

2–2 [i] 131-V-med 1063-A-m [ä–e]
R200084 F(i):554-2232-2730

2–3 [i] 131-V-med 1063-A-m [ä]
R200085 F(i):992-2292-2886

2–4 [i] 131-V-med 1063-A-m [i]
R200086 F(i):1318-2296-3111

2–5 [i] 131-V-med 1063-A-m [i]
R200087 F(i):1611-2305-3117

2–6 [e] 262-V-med 1036-A-w [e]
R138901 F(i):465-2499-3035

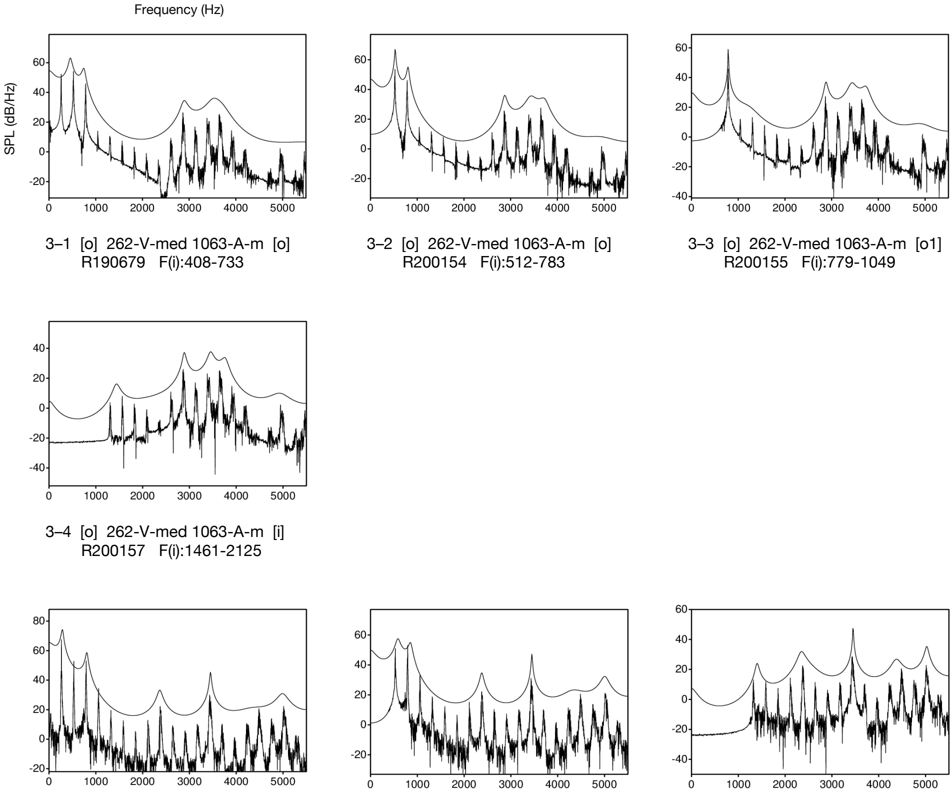
2–7 [e] 262-V-med 1036-A-w [e]
R199289 F(i):514-2495-2929

2–8 [e] 262-V-med 1036-A-w [ä]
R199290 F(i):1127-2485-3019

2–9 [e] 262-V-med 1036-A-w [ä–e]
R199291 F(i):1223-2497-3034

2–10 [e] 262-V-med 1036-A-w [e]
R199292 F(i):1526-2498-3021

Figure 3. Examples of initial close–open and subsequent back–front shifts for HP-filtered sounds of close-mid and close back vowels. [C-08-02-F03]



3–1 [o] 262-V-med 1063-A-m [o]
R190679 F(i):408-733

3–2 [o] 262-V-med 1063-A-m [o]
R200154 F(i):512-783

3–3 [o] 262-V-med 1063-A-m [o1]
R200155 F(i):779-1049

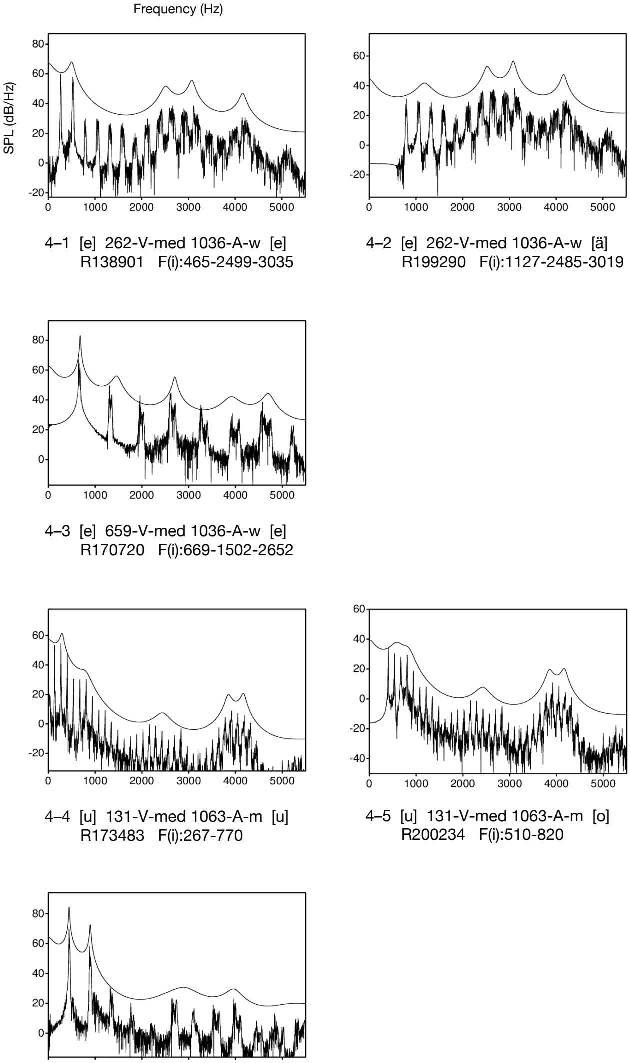
3–4 [o] 262-V-med 1063-A-m [i]
R200157 F(i):1461-2125

3–5 [u] 262-V-med 1063-A-m [u]
R173474 F(i):272-783

3–6 [u] 262-V-med 1063-A-m [o]
R200254 F(i):564-814

3–7 [u] 262-V-med 1063-A-m [ü]
R200257 F(i):1396-2232

Figure 4. Examples of sounds for which the impact of HP sound filtering on vowel recognition depended on the fo (and pitch) of sound production. [C-08-02-F04]



4–1 [e] 262-V-med 1036-A-w [e]
R138901 F(i):465-2499-3035

4–2 [e] 262-V-med 1036-A-w [ä]
R199290 F(i):1127-2485-3019

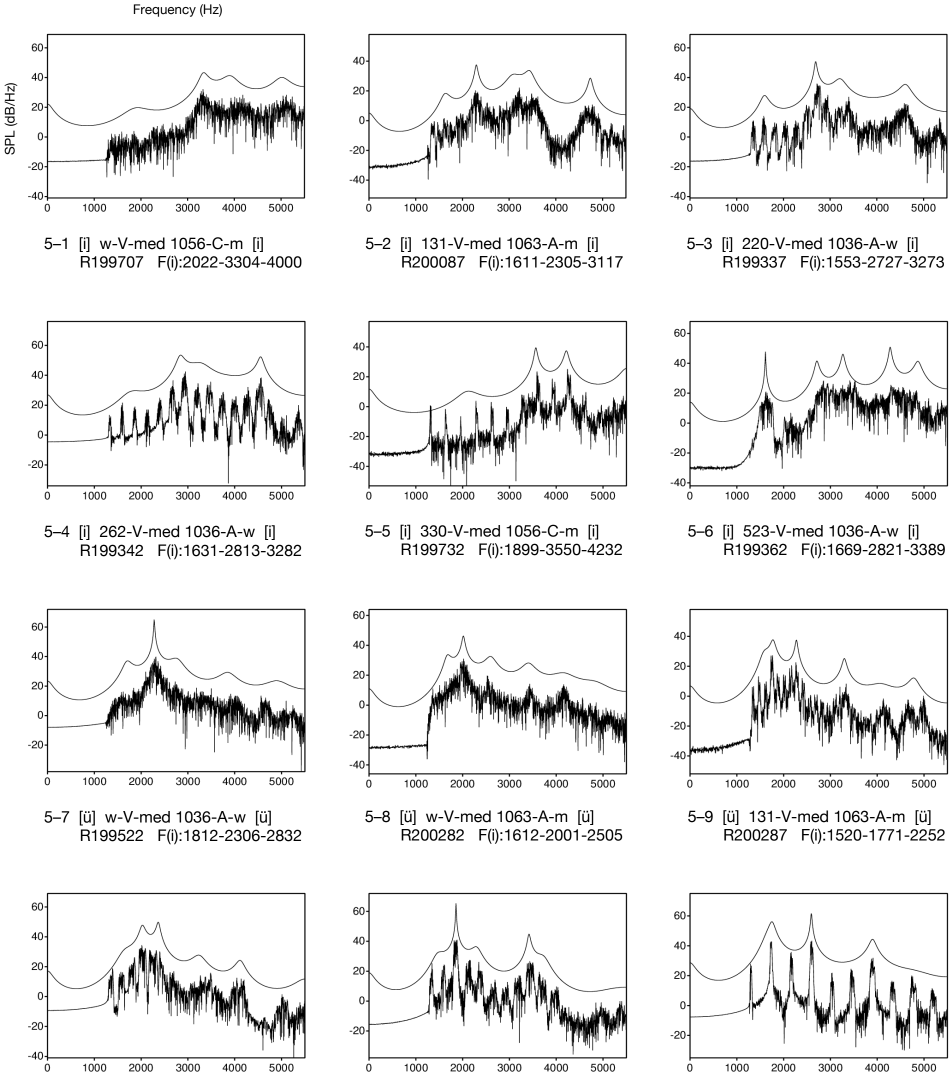
4–3 [e] 659-V-med 1036-A-w [e]
R170720 F(i):669-1502-2652

4–4 [u] 131-V-med 1063-A-m [u]
R173483 F(i):267-770

4–5 [u] 131-V-med 1063-A-m [o]
R200234 F(i):510-820

4–6 [u] 440-V-med 1063-A-m [u]
R149134 F(i):456-910

Figure 5. Further examples of recognisable sounds of front vowels with the entire frequency range of statistical F1 of all vowels being HP-filtered. [C-08-02-F05]



5–1 [i] w-V-med 1056-C-m [i]
R199707 F(i):2022-3304-4000

5–2 [i] 131-V-med 1063-A-m [i]
R200087 F(i):1611-2305-3117

5–3 [i] 220-V-med 1036-A-w [i]
R199337 F(i):1553-2727-3273

5–4 [i] 262-V-med 1036-A-w [i]
R199342 F(i):1631-2813-3282

5–5 [i] 330-V-med 1056-C-m [i]
R199732 F(i):1899-3550-4232

5–6 [i] 523-V-med 1036-A-w [i]
R199362 F(i):1669-2821-3389

5–7 [ü] w-V-med 1036-A-w [ü]
R199522 F(i):1812-2306-2832

5–8 [ü] w-V-med 1063-A-m [ü]
R200282 F(i):1612-2001-2505

5–9 [ü] 131-V-med 1063-A-m [ü]
R200287 F(i):1520-1771-2252

5–10 [ü] 220-V-med 1063-A-m [ü]
R200297 F(i):1526-2021-2347

5–11 [ü] 262-V-med 1063-A-m [ü]
R200307 F(i):1408-1859-2284

5–12 [ü] 440-V-med 1063-A-m [ü]
R200317 F(i):1652-1815-2573

8.3 Conclusion

According to the indications given in the literature and the results of the two experiments presented here, LP or HP filtering of vowel sounds does not, in general, cause a loss of recognised vowel quality, but it often causes vowel quality shifts whose general directions are predictable. This indication is interpreted here as supporting the notion that vowel sounds are a kind of perceptual and acoustic foreground–background phenomenon.

If LP filtering is looked at from the perspective of stepwise increasing CFs, natural sounds produced as back or as front vowels are indicated to be recognised as back vowels initially, and back–front differentiation only occurs with CF levels above c. 1.3 kHz, followed by rounded–unrounded differentiation for CF levels above c. 2–2.3 kHz. (Note that, when filtered from the perspective of stepwise increasing CFs, most of the sounds initially produced as close and close-mid front unrounded vowels shifted from back to front rounded vowels before they were finally recognised as front unrounded vowels. However, some sounds manifested a direct back to front unrounded shift.) Thus, the sound characteristics of natural sounds of back and front vowels, relevant for vowel recognition, are perceptually and acoustically partly similar, which is mirrored in the lower part of the vowel spectrum, and they differ only in relation to the subsequent difference in middle and higher frequencies. Likewise, the vowel-related sound characteristics of natural sounds of close front rounded and unrounded vowels are often perceptually and acoustically partly similar, mirrored in the lower and middle part of the vowel spectrum, and they differ only in relation to the subsequent difference in higher frequencies. The same holds true for the sound characteristics of close-mid front rounded and unrounded vowels.

If HP filtering is also looked at from the perspective of stepwise increasing CFs, natural sounds produced as close front vowels are indicated to often initially shift to close-mid or even open-mid vowels before these shifts may then be reverted. Thus, sounds of close and close-mid (and sometimes open-mid) vowels can have a similar higher part of the vowel spectrum, and they then differ only in relation to the preceding difference in lower frequencies. Furthermore, occurring cases of back–front vowel quality shifts for natural sounds of /o, u/ when HP filtered with a CF of 1320 Hz indicated a possible similarity of the higher part of the vowel spectrum for some sounds of back and front vowels.

Within the limits of the general shift directions found and the role of f_0 in LP and HP sound filtering experiments, CFs and associated vowel quality shifts varied for different vowels. In addition, the individual shifts also varied to some extent for sounds of the same vowel. Both findings indicate that the relation between lower, middle and higher spectral energy does not follow a simple rule, but rather it concerns the entire course of a given spectral energy distribution. (Therefore, again, an investigation of sounds of a limited set of long vowels of a language produced at a single f_0 level does not allow for a generalisation of the results for sounds of other vowels or sounds produced at other f_0 levels.)

In the context of interpreting the effect of vowel sound filtering and discussing the vowel as a kind of perceptual and acoustic foreground–background phenomenon, the finding that spectral peak patterns were not a prerequisite for vowel recognition is central and completes the demonstration given in Chapters 7.3 and 7.4. Most importantly, numerous sounds of close front unrounded vowels were recognised as front rounded vowels, with no manifest peak > 1 kHz in the sound spectrum when LP filtered with CFs in the frequency range of 2100–1530 Hz. Furthermore, the vowel quality of numerous sounds of front vowels was still recognised successfully even if the spectral energy below 1 kHz was lacking completely due to HP filtering. Therefore, to repeat, the foreground–background phenomenon cannot generally be related to spectral peaks or formants.

All this leads to the assumption that, even though vocal tract resonances are engaged in vowel production, the listener's ear does not recognise the resonance pattern of sound production in an unmediated way, but it relies on an actual perceptual and acoustic foreground–background relation, represented in the actual energy distribution of the entire course of a sound spectrum. Both the consequences concerning the classification of spectra in relation to vowel quality recognition and the assumption that vowel recognition does not relate to production resonances in a direct and unmediated manner will be addressed in more detail in the following excursus and the following main chapter.

It is in these terms that the vowel sound is understood here as a kind of perceptual and acoustic foreground–background phenomenon: The vowel sound cannot be described in a simple manner as a phenomenon of resonance patterns or spectral envelope shapes that are external to individual sounds. The vowel sound has to be approached as a phenomenon of a sound-internal relation of energy distribution, whose character is not yet disclosed but is indicated to also relate to pitch.

Excursus – Vowel Quality and Harmonic Spectrum

Introduction

In the Preliminaries, the main line of argument was based on the investigation of natural vowel sounds and the related spectral representation of vowel quality in terms of reviewing the prevailing acoustic theory of formants. Besides both the lack of a methodological basis for formant estimation for all recognisable natural vowel sounds and the systematic divergence of empirical findings from predictions of formant theory, the main conclusion was that the "[…] prevailing theory is falsified because, for a substantial portion of vowel sounds, the opposite of what the theory claims to be true actually applies: In many cases, given a variation of fundamental frequency, vowel sounds with very different formant patterns allow for a perception of the same vowel quality, while vowel sounds with similar formant patterns allow for a perception of different vowel qualities." (p. 76)

In the context of this rejection of a formant pattern as a given spectral peak or shape pattern that would represent vowel quality, the question was brought up of whether the vowel spectrum should not be looked at with regard to its course and the corresponding relation of lower and higher spectral energy. Based on our observational knowledge, we speculated that the vowel-related spectral characteristics of voiced sounds of two vowels produced at similar f_o may indeed be described as a relation of maximal spectral similarity and subsequent spectral difference: "[…] any single sound of a vowel compared with sounds of another vowel (given similar fundamental frequencies of the sounds) is assumed to be describable in terms of a relation of maximal spectral similarity and subsequent – related – spectral difference: For a (lower) frequency range, the harmonic spectrum of the single sound of the first vowel of comparison can resemble some other harmonic spectra of the second vowel, but if the maximum of this frequency range of possible resemblance is reached, its spectrum differs from all the spectra of the second vowel sharing the maximal similarity, while still resembling some other spectra of the first vowel." (p. 81)

In the present treatise, however, the line and focus of argument has shifted from a critical review of the prevailing acoustic theory of the vowel to observation- and experiment-based statements about the relation between recognised vowel quality and spectral representation in terms of indices for a future acoustic theory. With this, the focus has shifted from fundamental frequency to pitch (or to a comparable perceptual reference to a sound pattern repetition over time), from a kind of spectral foreground–background relation of vowel sounds – valued in an acoustic perspective to the vowel sound being a perceptual phenomenon of sound pattern recognition which includes not only a referencing to pitch (or to the above alternative) but also a perceptual weighting of the energy configuration within a repeating sound pattern –, and from formant patterns and spectral shapes to a broader field of possible vowel-related spectral characteristics which, in part, cannot be described within existing frameworks of spectral peak patterns or spectral envelopes. In this context, a terminological clarification is needed, and the thesis of maximal spectral similarity and subsequent and related spectral difference for sounds of two vowels has to be exposed and discussed anew.

The vowel sound as a kind of perceptual and acoustic foreground–background phenomenon

As demonstrated, LP filtering of vowel sounds does not generally impair or corrupt vowel quality recognition, but it causes vowel quality shifts whose general directions are indicated to be predictable: Unfiltered natural sounds of two vowel qualities are recognised categorically as different qualities, but when the sounds are LP filtered with decreasing CFs, the quality difference may change or it may give way directly so that only one of the two vowels is recognised for all sounds of comparison. This perceptual phenomenon indicates that, when looked at from an acoustic perspective of the spectral representation of vowel quality, sounds of two vowels can have similar lower spectral energy up to a given frequency limit, and only above this limit will the energy configuration be different in relation to the two qualities in question.

Having anticipated this indication in the Preliminaries based only on viewing vowel sound spectra and not on conducting LP filtering experiments, we have named it “a kind of spectral foreground–background relation of vowel quality representation” in terms of “a relation of maximal spectral similarity and subsequent – related – spectral difference” (see above, the vowel spectrum looked at from the perspective of increasing spectral frequency). In the context of the shift in argument, mentioned above, and because of the new experiments conducted, we concluded that the phenomenon in question does not only concern the spectral representation of vowel quality but also, and foremost, the perceptual process of vowel recognition. In consequence, here, we discuss the vowel sound and its recognised quality as “a kind of perceptual and acoustic foreground–background phenomenon”.

On the one hand, it is attractive to use the expression “foreground–background” for both the perceptual relation of sounds of two vowels observed in LP filtering and, as a consequence, the possible spectral similarity and subsequent spectral difference between these filtered and unfiltered sounds. Further, the expression characterises well the reported cases of sounds (observed in several of our experiments) that were recognised as a front vowel by some listeners but as a back vowel by other listeners. It may even be the case that vowel recognition differences among listeners, such as either front unrounded or rounded vowel recognition (also observed in several of our experiments), may also be related to a kind of foreground–background character of the vowel sound in terms of some kind of sound energy weighting which may differ to some degree among listeners. Finally, the reported vowel quality shifts for HP-filtered natural sounds may also be related to such a characteristic of the vowel sound.

On the other hand, the expression “foreground–background” has drawbacks: Here, the expression only stands for and names an ensemble of observations and experimental findings without a theoretical basis, and the conclusion that the vowel sound is indeed the result of a recognition process involving a foreground–background valuation of a sound pattern repeated over time is not proven here in a definitive way. Furthermore, as the Introduction notes, the expression does not refer to foreground–background relations as in auditory scene analysis. Finally, using this expression, other perceptual phenomena may spontaneously be associated without evidence that they are comparable to vowel sounds. To give an example, it is not yet made evident here that vowel sounds are comparable to visual phenomena such as ambiguous or reversible figures (although the demonstration of double-vowel and double-pitch may be considered within such a comparison).

Thus, the question of whether the expression “foreground–background phenomenon” is appropriate for the vowel sound as such is not yet answered. However, because we did not consider any other expression more indicative of both the perceptual phenomenon of vowel quality shifts in LP and HP filtering and the acoustic classification approach (note this parallelism) according to a concept of spectral similarity and subsequent vowel-related spectral difference, we use the expression in this treatise, even if in a temporary and provisory manner. Future theory building and related empirical investigation will clarify the issue.

Reciprocal maximal spectral similarity and subsequent spectral difference between sounds of different vowels – a hypothesis

Concerning the thesis of vowel-related maximal spectral similarity and subsequent spectral difference for quasi-periodic sounds of two vowels, because of the shift of perspective and argument in this treatise and because of the substantial methodological and empirical effort required for an extensive investigation of the hypothesis, no study of a large sound sample was conducted for presentation and discussion here. However, the hypothesis may be of interest both for the valuation of findings of several experiments reported here and for future theory building. Therefore, in this excursus, the hypothesis is outlined anew. This outline includes an initial assumption in terms of the rejection of the theory that average spectral templates can represent vowel qualities, some details of earlier attempts for such templates, a revised formulation of the hypothesis of maximal spectral similarity and subsequent spectral difference of sounds of different vowels, including the condition of reciprocity of maximal similarity which was not formulated as a criterium in the Preliminaries, and exemplary sound compilations and graphic illustrations.

Initial assumption: In the Preliminaries, based on the documented examination of the vowel-related spectrum, we have assumed that no single average spectrum of sounds of a vowel can serve as a template, even if the template were related to f_o:

“Given that a [natural] voiced vowel sound is produced in isolation and that it exhibits a quasi-constant periodic spectral characteristic, and given its unambiguous perception as belonging to a specific vowel quality, then its average harmonic spectrum, measured for the entire duration of the respective sound, is said to be vowel specific. For a frequency range concerning the physical representation of all vowels of a language, a series of harmonics quasi-identical in number, frequencies and levels can only be found for other sounds of the same vowel but not for other sounds of any other vowel.

“At first glance, such a statement seems trivial. But it is not.

“To say that a harmonic spectrum of a vowel sound is specific for the perceived vowel quality – given the above conditions for the sounds under investigation – is not to say that all sounds of a vowel have very similar spectra of this kind. As shown, large spectral variations can be found for the sounds of one vowel, particularly if the vocal effort is varied during sound production, sounds of different speaker groups are compared and different speaking and singing modes and styles, including stage voices, are also considered.

“Therefore, an attempt to directly assess the spectral difference related to a perceptual difference of two vowels simply by calculating an average harmonic spectrum for all sounds of one vowel at a given fundamental frequency and comparing it with the similarly averaged harmonic spectrum of the other vowel may, in many cases, not result in a clear spectral difference.” (pp. 80–81)

Earlier attempts at classifying harmonic spectra according to spectral distances when compared to vowel-related spectral templates: However, earlier studies did attempt to define vowel-related average spectra as vowel quality-related templates and classify individual harmonic spectra by calculating the spectral distance to these templates. This will be discussed here, referring to the two studies of de Cheveigné and Kawahara (1999) and Hillenbrand and Houde (2003).

The spectral shape or envelope of a vowel sound is, in most cases, derived through some kind of smoothing operation. However, as de Cheveigné and Kawahara and Hillenbrand and Houde discuss in detail, the smoothing operation is unproblematic only for lower levels of f_o, while an estimation of spectral envelopes for sounds produced at middle or higher f_o lacks methodological substantiation. Because of this, these authors propose to relate to unsmoothed harmonic spectra or narrow band spectra of individual vowel sounds and compare them with a set of smoothed vowel-related spectral templates in an attempt to classify the unsmoothed spectra according to vowel recognition.

According to the approach of de Cheveigné and Kawahara, for each vowel quality, a single spectral envelope is assumed to be available as a template. The calculated harmonic spectrum of an individual vowel sound is then compared with the available templates of the vowels investigated by calculating the spectral distance at the harmonic frequencies of the individual sound. Finally, vowel classification relates to the smallest distance found for comparison with one of the templates. This classification procedure was tested on a sample of synthesised sounds of the five Japanese vowels /a, e, i, o, u/ produced at f_o of 20–300 Hz (with increasing f_o in 1-Hz steps), with the sounds being compared with five vowel-related spectral envelope templates at the frequencies of the harmonics of the sounds, the templates based on the synthesis filters. Apparently, the study provided good classification results, although detailed results were not given.

In a subsequent investigation, pursuing the comparison of unsmoothed harmonic spectra (narrow band spectra) of individual vowel sounds with vowel-related smoothed spectral templates, Hillenbrand and Houde developed an approach that addresses two limitations of the above study: Instead of comparing harmonic spectra and vowel-related spectral templates only at harmonic frequencies of the individual sounds, the comparison concerned all frequency bands investigated, and instead of testing the procedure on synthesised sounds of five vowels /a, e, i, o, u/, the procedure was tested on a large sample of naturally spoken utterances of 12 American English vowels. Further differentiations concerned separate templates for children, women and men (the templates created by averaging the narrow band spectra of sounds of a vowel spoken by a panel of speakers of a given speaker group) and, instead of comparing a single average spectrum of an individual sound with single vowel templates, spectral sequences were compared, accounting for the dynamic course of the harmonics of a natural sound. Classification accuracy was found to be ≥ 90%.

As mentioned, the motivation in these studies to compare smoothed vowel-related spectral templates with harmonic spectra or narrow band spectra of individual sounds resulted from the methodological problem of *F*-pattern and spectral shape estimation. The basic idea underlying the general approach was to replace *F*-patterns with (ideal or averaged) spectrally smoothed vowel templates as references, one template per vowel or per vowel and speaker group, and to compare harmonic spectra or narrow band spectra of individual sounds with the templates in order to classify them according to the minimal spectral distance to a template. Yet, such an approach has shortcomings. Above all, concerning the spectral representation of vowel quality of natural vowel sounds, it does not account for three basic aspects of the vowel spectrum: its general relation to f_o (independent of the speakers and speaker groups), its possible variation extent (mentioned as a limitation in the study of de Cheveigné and Kawahara) and its non-uniform character (above all with regard to f_o ranges and vowel qualities) – not to mention the lacking consideration of the differentiation between f_o and pitch and the fact that the vowel spectra of many types of manipulated or synthesised vowel sounds differed greatly from the vowel spectra of unmanipulated natural vowel sounds. Based on the evidence given in this treatise, it is expected that systematic vowel confusion would have occurred in both studies if the applied f_o range for the investigation of natural sounds produced by speakers of a given speaker group and also for sound synthesis had been extended up to 500 Hz or higher. Additionally, if a variation of phonation type and vocal effort had been taken into account as well, the confusion would have proven to be even more pronounced.

However, as also indicated in the Preliminaries, we assume that even an attempt at a classification of individual sound spectra of a vowel in comparison with vowel quality-related spectral templates related to f_o (and/or pitch) will fail if no strong restrictions are imposed on the examined sound sample. We assume this mainly because of two findings: the extent of spectral variation observed for natural sounds of a vowel resulting from the variation of sound production (above all concerning phonation, vocal effort, production style and speaker-specific characteristics) and the extent of spectral variation observed for manipulated natural and synthesised sounds of a vowel. In light of these findings, we propose an alternative approach to assess the spectral difference between sounds of different vowels with quasi-periodic sound characteristics, discussed in the following section.

Reciprocal maximal spectral similarity and subsequent spectral difference between sounds of different vowels: Comparing the harmonic spectra of natural sounds of two different vowels, according to the hypothesis formulated in the Preliminaries and presented here in more detail, the vowel-related difference can be assessed by investigating the occurring maximal and reciprocal spectral similarity and their subsequent spectral difference.

Given that
(i) for all (long) vowel qualities of a language, sounds with quasi-periodic and steady-state sound characteristics (monophthongs) are compared with each other,
(ii) the sounds investigated are produced in isolation or extracted as sound nuclei from words or syllables and manifest quasi-static spectral characteristics,
(iii) the sounds compared with each other are produced at similar f_o levels,
(iv) vowel recognition rate of the sounds is high,
(v) vowel recognition does not directly relate to sound duration,
(vi) recognised vowel quality is maintained in resynthesis based on estimated average harmonic spectra of the natural sounds,
(vii) the number of investigated sounds of a single vowel quality represents a sufficient degree of possible spectral variation within the vowel-related frequency range,
the following is said to apply: If the harmonic spectrum of a single sound – here termed reference sound – of vowel quality A is compared with all of the spectra of other sounds of the same vowel quality and all of the spectra of sounds of vowel quality B, there is a frequency limit above which the spectrum of the reference sound diverges from any

spectrum of the sounds of the second vowel B, but not from any spectrum of the sounds of the first vowel A. If, in reverse, a sound of vowel quality B with spectral similarity up to the frequency limit mentioned is, in turn, taken as a reference sound and compared with both the spectra of other sounds of the same vowel quality B and the spectra of all sounds of vowel quality A and, if the frequency limit of spectral similarity remains unchanged, that is, reciprocal to the first comparison, then the subsequent spectral difference is predicted to be vowel-specific: If this reciprocal comparison condition is fulfilled, the spectral difference will then consist of higher levels for the harmonics succeeding the frequency range of spectral similarity for one vowel quality (A or B) when compared with the harmonics of the sounds of the other quality (B or A).

In these terms, from an acoustic perspective, sounds of two vowels (given the above conditions) are assumed to manifest either a reciprocal or a non-reciprocal relation of maximal spectral similarity and subsequent spectral difference. If the similarity is maximal and reciprocal, the subsequent spectral difference is vowel-related: Spectral energy succeeding the range of similarity is higher for the sounds of one vowel than for the sounds of the other. In parallel and from a perceptual perspective, if sounds of two vowels manifesting reciprocal maximal spectral similarity are LP filtered with a CF of the upper level of the frequency range of similarity, this will result in a single recognised vowel quality for the corresponding sounds of comparison. Only a subsequent stepwise increase in CF (according to the increase in harmonic number) will result in the recognition of two different vowel qualities.

If the spectral similarity of two sounds of two vowels is not maximal in reciprocal terms, then the sounds of one of these vowels and (at least some) sounds of a third vowel will result in a reciprocal relation of maximal spectral similarity and subsequent spectral difference, and LP filtering will produce corresponding effects.

This concept of assessing vowel-related spectral differences is formulated in an abstract and ideal way and, evidently, leaves many questions concerning the concrete empirical verification or falsification unanswered. However, because the approach is not investigated further empirically here and future theory building may create a more direct approach to investigating the relation between recognised vowel quality and acoustic sound characteristics, this presentation and discussion of the hypothesis is limited to a general outline. It originated from an extensive examination of vowel spectra, it interprets the findings of the LP filter study, it illustrates the assumed foreground–background character of the vowel sound, and it aims at serving future theory building in terms of a knowledge-based formulation of how the spectra of sounds of different vowels with quasi-periodic and steady-state sound characteristics differ from each other.

In the following sections, examples are presented that illustrate reciprocal maximal spectral similarity and subsequent spectral difference, as well as non-reciprocal similarity and extension of comparison. The natural sounds used for the illustration are extracted from the Zurich Corpus and were fully recognised in the standard listening test conducted when creating the corpus (100% vowel recognition rate matching vowel intention).

A first example and illustration of sound comparison: In order to illustrate this method of assessing a vowel-related spectral difference only after the assessment of a reciprocal maximal spectral similarity for sounds of two vowels, the corresponding classification procedure and first illustration are given here based on a sample of six voiced or breathy sounds, four sounds of /o/ and two sounds of /u/ produced by men in V context with different vocal efforts at calculated f_o of c. 150 Hz. The sounds are shown in Figure 1 and Table 1, Series 1a. For demonstration purposes, these sounds were edited: Middle 1 sec. sound nuclei were extracted from the entire sounds produced, and a fade in/out of 0.03 sec. was applied. (See the corresponding comment in the online sound archive for the unedited reference sounds.) The average sound spectrum was calculated for the entire duration of the 1 sec. sound nucleus. The levels of the harmonics were assessed based on the spectra.

Sounds of /o/ and /u/ were selected because, in Standard German, there is no third quality of long vowels between them, and all of the other long vowel qualities are either more open or more front. Therefore, for all sounds of /o/ and /u/, there are sound configurations with reciprocal maximal spectral similarity and subsequent spectral difference according to the hypothesis (given that only long vowels are investigated). At the same time, for f_o levels of the sounds < 500 Hz, the spectral similarity and subsequent spectral difference separating the sounds of these two vowels can be demonstrated for a very limited frequency range < 1 kHz, involving only a few harmonics. (For higher f_o levels, according to the author's estimate, the first two harmonic levels are vowel-related, that is, the levels of $H2$ are higher for sounds of /o/ than for sounds of /u/.) Sounds produced at f_o below 200 Hz were selected because, despite the very limited frequency range < 1 kHz, the number of harmonics within this range is still high enough to demonstrate the possible variability of the harmonic level configurations that occur for sounds of a single vowel produced at lower or middle f_o levels, hindering a simple assessment of two averaged harmonic spectra related to the two qualities, even if these average spectral templates were related to f_o.

In an attempt to represent the possible variation of the harmonic level configuration for the sounds of the two vowels, in the sample, sounds produced with low vocal effort were opposed to sounds produced with medium and high vocal effort. Four sounds of /o/ were selected because the discussion of the classification procedure focuses on a single reference spectrum of a sound of /o/ compared with other sounds of /o/ and /u/.

A first visual inspection of the frequency range < 1 kHz of all harmonic spectra compared indicates that the spectra related to the sounds of /o/ produced with low vocal effort more closely resemble the spectrum of the sound of /u/ produced with low vocal effort (see Figure 1, comparing sounds 2 and 3 with sound 1) than the spectra of the other sounds of /o/ produced with medium or high vocal effort (see Figure 1, comparing sounds 2 and 3 with 5 and 6). Moreover and most importantly, the spectrum related to the sound of /u/ produced with high vocal effort more closely resembles the spectra of the sounds of /o/ produced with medium or high vocal effort than the spectra of the other sounds of /o/ and /u/ (see Figure 1, comparing sounds 5 and 6 with sound 4). Thus, the configurations of the spectra of the two vowels (their harmonic envelope) overlap for the frequency range < 1 kHz. This is illustrated in Figure 2, Chart A, in which the vowel-related harmonic level configurations of all sounds are compared with each other (for illustration, the levels were slightly adjusted to simplify the graphic configuration shown and to illustrate the main idea): For a frequency range of up to 1 kHz, the configuration of the sound of o1 in the figure resembles the configurations of u1 and o2 more than o3 and o4. Conversely, the configuration for u2 resembles the configurations of o3 and o4 more than that of u1.

As a consequence, if the harmonic spectra of the natural sounds of /o/ produced with low vocal effort are compared with many spectra of sounds of /u/ produced with different vocal efforts and representing the entire range of possible spectral variation for that second vowel – the variation here represented by u1 and u2 – there is no general vowel-related spectral separation. Accordingly, in the example illustrated in Figure 2, the spectrum of o1 “underruns” the spectrum of u2 but “overruns” the spectrum of u1, which is almost true for o2, too. This is illustrated in Chart B. Thus, there is no possible pair of spectral templates for the two vowels in reference to which the sounds of /o/ compared with the template of /o/ would always result in a smaller spectral distance than when compared with the template of /u/, and vice versa.

But if the two harmonic spectra of the sounds of /o/ produced with low vocal effort are only compared with those spectra of sounds of /u/ for which the frequency range of the resemblance of the harmonic levels is the largest with regard to the low harmonic numbers (from *H*1 upwards) found for sounds of the two vowels in question (at a given f_o), here represented by u1, then there is a frequency limit above which the harmonic levels of the sounds of /o/ will markedly surpass the levels of the sounds of /u/ (see Chart C). In other terms, if a harmonic spectrum of a sound of /o/ (here represented by o1) is compared with those harmonic spectra of /u/ for which the lower spectral similarity for all sounds of these two vowels is maximal in reciprocal terms (here represented by u1), then there is indeed a vowel-related subsequent spectral difference in that the subsequent harmonic levels of the sound of /o/ must resemble other sounds of /o/ with the same reciprocal maximal spectral similarity (here represented by o2), and they must differ from the corresponding harmonic levels of all sounds of /u/ with that similarity (see Chart D). And if this is the case for the comparison of one sound of /o/, it is the case for all other sounds of /o/ compared with sounds of /u/ fulfilling the comparison conditions: The energy of the harmonics succeeding the reciprocal maximal spectral similarity will diverge in a vowel-specific direction, here in terms of higher levels for sounds of /o/ than /u/ (compare Charts D and E). Such is the hypothesis of reciprocal maximal spectral similarity and subsequent spectral difference for sounds of two vowels.

Note that the frequency extension of reciprocal maximal spectral similarity (the number of harmonics with marginal level differences) for different comparisons of sounds of two vowels may somewhat differ, as is indicated here in Charts D and E: For the comparison of sounds produced with low vocal effort (see Chart D), the harmonic spectra of o1 and u1 differ only slightly for the first three harmonics before progressively deviating more strongly from each other for *H*4 and the higher harmonics. The same holds true for o2 and u1. However, for the comparison of sounds produced with medium and high vocal effort (see Chart E), the harmonic spectra of o3, o4 and u2 are similar for the first two harmonics only, and the spectra of /o/ markedly deviate from the spectrum of /u/ for *H*3. This indicates an upper-frequency limit of reciprocal maximal spectral similarity for the sounds illustrated in Chart

D up to c. 450 Hz but only up to c. 300 Hz for the sounds illustrated in Chart E, although sounds of the same two vowels are compared.

In sum, for different comparisons of single sounds of two vowels, the frequency extension of reciprocal maximal spectral similarity may vary to some degree, but the spectral difference of the sounds with regard to the harmonics succeeding reciprocal maximal spectral similarity is vowel-related in that the harmonics manifest higher levels for sounds of one vowel than sounds of the other. Accordingly, LP filtering of sounds of /o/ will always shift to a recognised quality of /u/, but the CF related to that shift may vary to some degree for different sounds. Conversely, LP filtering of sounds of /u/ will never result in a shift to a recognised quality of /o/ (filtering artefacts excluded).

In Series 1e–1h in Table 1, for each of the four sounds of /o/, LP-filtered variants with CFs of 350–500–650 Hz are presented, that is, reducing the resulting spectra to two, three or four harmonics. (Series 1e and 1f are related to the sound comparison shown in Chart D, and Series 1g and 1h are related to the sound comparison shown in Chart E.) The filtered sounds of a series are preceded by the respective sound of /u/ of comparison, and they are followed by the unfiltered sound of /o/. Indeed, when listening to the four sound series (from the first sound of /u/ of each series to the subsequent filtered and unfiltered sounds of /o/), a pronounced and unambiguous /u/–/o/ shift is recognisable for a CF of 500 Hz for o1 (above *H*3; author's estimate), but this shift is already pronounced for a CF of 350 Hz for o3 and o4 (above *H*2). For the sound o2, a somewhat intermediate transition from /u/ to /o/ is indicated (for verification, use the SpecFilt tool implemented in the Zurich Corpus and listen in both directions of the sound order).

A second example: With increasing openness or frontness of vowel sounds, the vowel-related frequency range increases, too, and so does the range of reciprocal maximal spectral similarity of sounds of different vowels. At the same time, if for some sounds of vowel A of comparison, a reciprocal maximal spectral similarity occurs in comparison with sounds of vowel B, this may not hold true for other sounds of vowel A. This is illustrated in the second example given here, based on a sample of four voiced sounds of /i, y, u/ produced by women in V context with different vocal efforts at a calculated f_o of c. 255 Hz (see Table 1, Series 2a). For demonstration purposes, the sounds in this example were edited and analysed similarly to those in the previous example. In addition, the harmonic levels in the frequency range of 2.2–3 kHz of the sound of /y/ were adjusted (manual BP filtering using the filter functionality of Adobe Audition) to produce a very similar spectrum up to 3 kHz when compared with one of the sounds of /i/. (For the original unedited sound of /y/, see sound 101029 in the Zurich Corpus.)

A visual inspection of the spectra of the two sounds of /i/ and the sound of /y/ indicates that one spectrum of /i/ related to the sound produced with high vocal effort is quasi-identical to the spectrum of /y/ up to c. 3 kHz, but this is not the case for the second spectrum of /i/ (see Series 2b). Although not demonstrated here, it is assumed that the comparison of the first sound of /i/ and the sound of /y/ represents a case of reciprocal maximal spectral similarity (see Series 2c) and that no such reciprocal similarity can be found for the second sound of /i/ compared with all occurring sounds of /y/. For the second sound of /i/, reciprocal maximal spectral similarity requires a comparison with sounds of /u/ (see Series 2d).

If this spectral examination relates to actual sound characteristics relevant for vowel recognition, then LP filtering of the higher frequency range for the first sound of /i/ will result in a (first) recognised vowel quality shift towards /y/, and LP filtering of the higher frequency range for the second sound of /i/ will result in a (first) recognised vowel quality shift towards /u/. This is indeed the case: If the first sound of /i/ in comparison with the sound of /y/ is LP filtered with stepwise decreasing CF = 2700–2500–2300 Hz, that is, from the assumed upper-frequency limit of reciprocal maximal spectral similarity of the two sounds down to their still common prominent spectral energy at c. 2200 Hz, the recognised vowel quality of the sound of /i/ shifts to /y/, as shown in Series 2e in Table 1 (author's estimate; for verification, use the SpecFilt tool implemented in the Zurich Corpus, with sufficient volume). Yet, if the second sound of /i/ is LP filtered with the same decreasing CFs, the recognised vowel quality of this sound shifts to /u/, as shown in Series 2f.

Because of sound configurations of this type, it was said that sounds of two vowels (given the above conditions of comparison) manifest either a reciprocal or a non-reciprocal relation of maximal spectral similarity, and it was further predicted that if the relation is reciprocal, the spectral difference is vowel-related, and if the relation is non-reciprocal, sounds of one of these vowels and sounds of a third vowel will result in a reciprocal relation of maximal spectral similarity and subsequent spectral difference. Note that based on only these four sounds presented in this second example, it is expected that any sound of /i/ for which a reciprocal maximal spectral similarity to sounds of /y/ is found will show higher levels of the harmonics that succeed the frequency range of said spectral similarity, and that the same holds true for the sound of /i/ for which a reciprocal maximal spectral similarity to sounds of /u/ is found.

Two additional examples: A third and fourth additional example further illustrate the spectral similarity–difference hypothesis.

As a third example, for a common level of calculated *f*ₒ of c. 240 Hz, Series 3 in Table 1 shows the spectra of two sounds of /e/, one sound of /ø/ and one sound of /o/, the sounds produced by children, women and men in V context with different vocal efforts (entire sounds; acoustic analysis according to the standard procedure of the Zurich Corpus). Note that for the second sound of /e/, the harmonic levels in the frequency range of 1–3 kHz were lowered (manual BP filtering) to produce a similar spectrum up to 3 kHz compared to the sound of /o/. (For the unedited original sound of /e/, see sound 187540 in the Zurich Corpus; no further sound editing was applied for the other three sounds.) As was the case for the previous comparison of sounds of /i, y, u/, a visual inspection of the harmonic spectra indicates no general vowel-related similarity for the spectra in terms of a single vowel-related spectral template for /e/. For the first sound of /e/, the harmonic level configuration *H*1–*H*8 resembles the configuration of /ø/ (compare sounds 1 and 2 of the series). Still, for the second sound of /e/, no sounds of /ø/ with a similar configuration of harmonic levels, including frequencies of > 1 kHz, were found in the Zurich Corpus. Instead, when compared with sounds of /o/, sounds with similar harmonic level configurations *H*1–*H*10 occurred, as is illustrated in the sound series (compare sounds 3 and 4 of the series). Accordingly, LP filtering of the first sound of /e/ of the series with stepwise decreasing CFs from 3 kHz downwards results in a recognised vowel quality shift towards /ø/ and subsequently to /o/, but LP filtering of the second sound of /e/ results in a direct recognised vowel quality shift towards /o/ (author's estimate; for verification, use the SpecFilt tool implemented in the Zurich Corpus).

As a fourth example, for a common level of calculated *f*ₒ of c. 190 Hz, Series 4 in Table 1 shows the spectra of two sounds of /a/ and two sounds of /ɔ/, the sounds produced by women and men in V context with different vocal efforts (entire sounds; acoustic analysis according to the standard procedure of the Zurich Corpus). A visual inspection of the frequency range < 1.5 kHz of all four harmonic spectra indicates that, once again, there is no general vowel-related similarity for the spectra. There is no template pair of harmonic level configurations or harmonic envelopes for /a/ and /ɔ/, in reference to which the sounds could be classified according to their spectral distance. Rather, the harmonic level configuration *H*1–*H*4 of the first sound of /a/ of the series resembles the configuration of the first sound of /ɔ/ (compare sounds 1 and 2 of the series, both sounds produced with low vocal effort), as is the case for *H*1–*H*5 of the second sounds of /a/ and /ɔ/ (compare sounds 3 and 4 of the series, sounds produced with medium and high vocal effort). However, for both sounds of /a/, the levels of the succeeding harmonics markedly surpass the levels of the sounds of /ɔ/ of comparison.

For the sounds in this example, LP filtering of both sounds of /a/ with stepwise decreasing CFs from 1.5 kHz downwards results in a recognised vowel quality shift to /ɔ/ (author's estimate; for verification, use the SpecFilt tool implemented in the Zurich Corpus).

Open questions and relativisations

The hypothesis of reciprocal maximal spectral similarity and subsequent vowel-related spectral difference as a prediction for the vowel-related spectra difference when comparing sounds of two vowels is outlined here for sounds with quasi-periodic and steady-state characteristics only because the hypothesis relates to the harmonic spectrum. The question of applying the hypothesis to sounds with non-periodic characteristics or marked dynamic spectral characteristics is left open here. However, the LP filtering can be applied to all sounds of all phonation types (including whispered sounds) and independent of the spectral characteristics being steady-state or dynamic, and the LP filtering results given in Chapter 9.1 indicate mostly phonation-independent vowel quality shifts. Notably, LP filtering vowel sounds is an easy and direct empirical approach to verify or falsify the hypothesis discussed here.

Furthermore, the hypothesis is outlined assuming "marginal" differences of configurations of harmonic levels for parts of spectral similarity of sounds and "substantial" level differences for parts of spectral dissimilarity. The question of the details of harmonic analysis and the definition of a limit separating "marginal" and "substantial" harmonic level differences related to vowel quality recognition is again left open here, as is also the case for other methodological aspects that would have to be laid out in detail for a comprehensive empirical study.

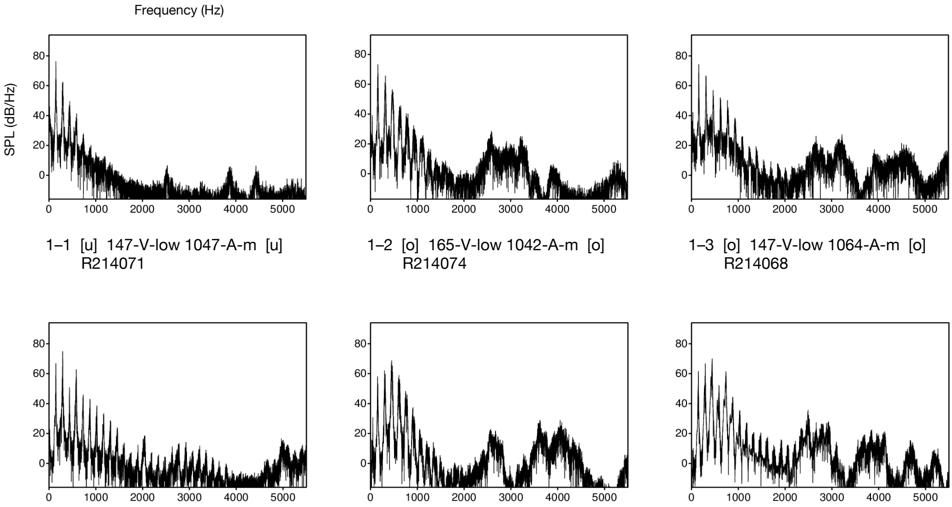
When investigating vowel sounds produced at f_o of > 700 Hz, we have observed sounds with no assessable difference in the harmonic spectrum but with different recognised vowel qualities. Yet, lacking a corresponding (re-)synthesis tool enabling the production of high-pitched vowel sounds with a high sound quality comparable to natural sounds, we could not investigate the matter. However, in this context, it is important to note that the approach of classifying vowel spectra according to reciprocal maximal similarity and subsequent spectral difference is proposed here under the condition that resynthesis based on a single estimated average harmonic spectrum of a natural sound does not affect recognised vowel quality.

Figure 1. Reciprocal maximal spectral similarity and subsequent spectral difference between sounds of different vowels: Comparison of four sounds of /o/ and two sounds of /u/. For details, see the text. The sounds are also shown in Table 1, Series 1a.
[C-E2-F01]

Figure 2. Reciprocal maximal spectral similarity and subsequent spectral difference between sounds of different vowels: Schematic illustration based on the sounds of Figure 1. For details, see the text.
[C-E2-F02]

Table 1. Reciprocal maximal spectral similarity and subsequent spectral difference between sounds of different vowels: Sound examples and spectral illustration. Column 1 = sound series and sound links (S/L). Columns 2 and 3 = vowel qualities of comparison (V1 and V2). Column 4 = calculated f_o (fo, approximate values given in Hz). Column 5 = vocal effort of sound production for the sounds compared (VE, with l = low, m = medium, h = high). Column 6 = Sound comparison and illustrated phenomena (for details, see the text). Column 7 = related parts of Figure 1. Note that LPC analysis for the edited sounds of Series 1 and 2 is not given in order to focus on the harmonic spectrum.
[C-E2-T01]

Figure 1. Reciprocal maximal spectral similarity and subsequent spectral difference between sounds of different vowels: Comparison of four sounds of /o/ and two sounds of /u/. [C-E2-F01]



1–1 [u] 147-V-low 1047-A-m [u]
R214071

1–2 [o] 165-V-low 1042-A-m [o]
R214074

1–3 [o] 147-V-low 1064-A-m [o]
R214068

1–4 [u] 147-V-hgh 1060-A-m [u]
R214069

1–5 [o] 147-V-med 1063-A-m [o]
R214072

1–6 [o] 147-V-hgh 1045-A-m [o]
R214070

Figure 2. Reciprocal maximal spectral similarity and subsequent spectral difference between sounds of different vowels: Schematic illustration based on the sounds of Figure 1. [C-E2-F02]

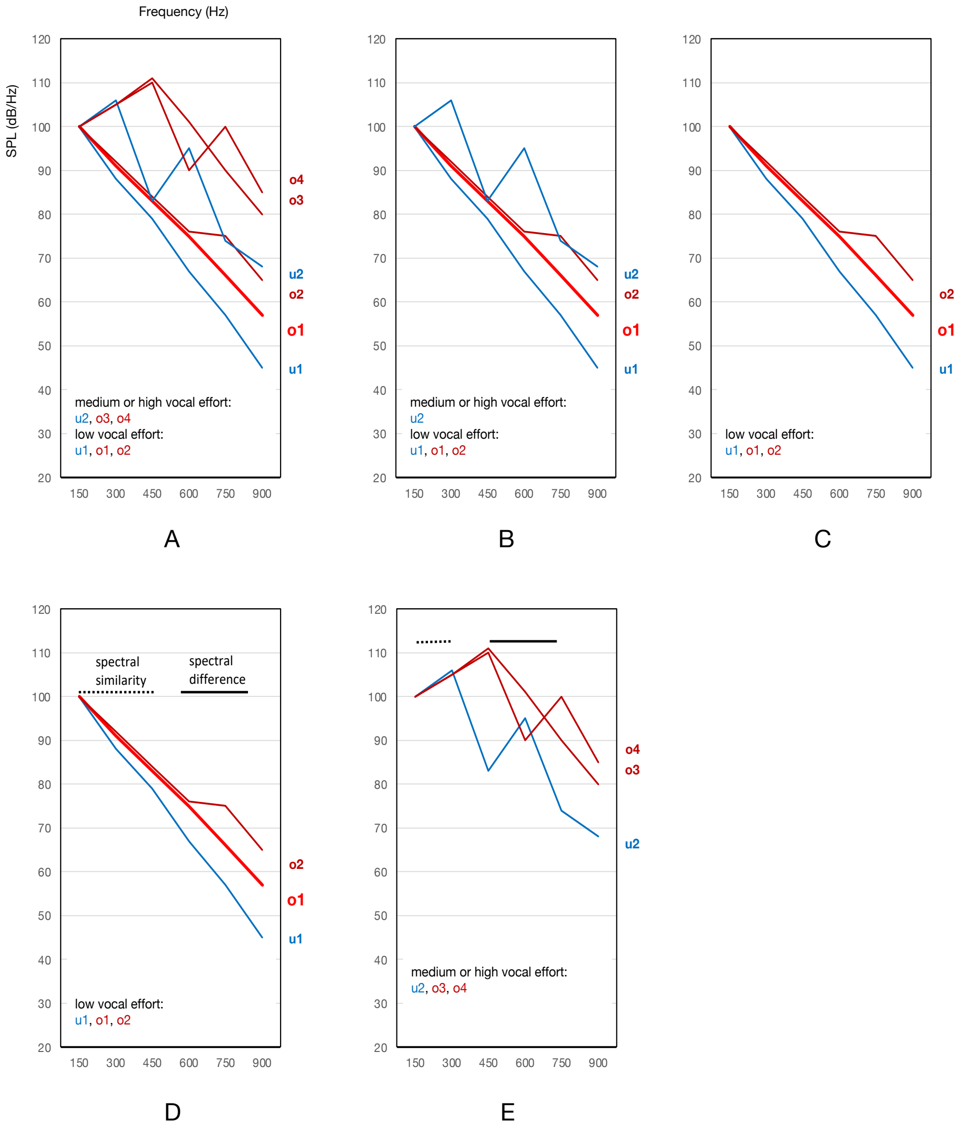


Table 1. Reciprocal maximal spectral similarity and subsequent spectral difference between sounds of different vowels: Sound examples and spectral illustration. [C-E2-T01]

S/L		V1	V2	fo	VE	Sound comparison and illustrated phenomena	Fig. 2
1	a ↗	o	u	150	l m h	Four edited natural sounds of /o/ and two sounds of /u/ (for the reference sounds, see the comment in the online sound archive).	A
	b ↗				l h	Extract of Series 1a: Two sounds of /o/ and two sounds of /u/.	B
	c ↗				l	Extract of Series 1a: Two sounds of /o/ and one sounds of /u/.	C
	d ↗				m h	Extract of Series 1a: Two sounds of /o/ and one sounds of /u/.	E
	e ↗				l	Extract of Series 1d: Sound of /u/, three LP filtered variants of a sound of /o/ with CFs of 350, 500 and 600 Hz, and the unfiltered sound of /o/ (see u1 and o1 in Figure 1d).	D
	f ↗				l	Extract of Series 1d: Sound of /u/, three LP filtered variants of a sound of /o/ with CFs of 350, 500 and 600 Hz, and the unfiltered sound of /o/ (see u1 and o2 in Figure 1d).	D
	g ↗				m h	Extract of Series 1e: Sound of /u/, three LP filtered variants of a sound of /o/ with CFs of 350, 500 and 600 Hz, and the unfiltered sound of /o/ (see u2 and o3 in Figure 1e).	E
	h ↗				h	Extract of Series 1e: Sound of /u/, three LP filtered variants of a sound of /o/ with CFs of 350, 500 and 600 Hz, and the unfiltered sound of /o/ (see u2 and o4 in Figure 1e).	E
2	a ↗	i	y u	255	l h	Two sounds of /i/, one sound of /y/ and one sound of /u/ (edited sounds).	
	b ↗		y		l h	Extract of Series 2a: Comparison of both sounds of /i/ with the sound of /y/.	
	c ↗		y		l h	Extract of Series 2a: Comparison of the first sound of /i/ with the sound of /y/, with maximal reciprocal spectral similarity.	
	d ↗		u		l	Extract of Series 2a: Comparison of the second sound of /i/ with the sound of /u/, with maximal reciprocal spectral similarity.	
	e ↗		y		l h	Related to Series 2c: Sound of /i/, unfiltered and LP filtered with CF = 2700–2500–2300 Hz, and unfiltered sound of /y/.	
	f ↗		u		l	Related to Series 2d: Sound of /i/, unfiltered and LP filtered with CF = 2700–2500–2300 Hz, and unfiltered sound of /u/.	
3	↗	e	ø o	240	m h	Two sounds of /e/, one sound of /ø/ and one sound of /o/ (for sound editing, see text).	
4	↗	a	ɔ	190	l m h	Two sounds of /a/ and two sounds of /ɔ/ (unedited sounds).	

9 Resonance Characteristics of Vowel Sound Production and Their Detectability in the Acoustic Analysis of Radiated Sound

9.1 Questioning the Direct Relation Between Resonances of Vowel Sound Production and Estimated Resonance Characteristics of Radiated Sound

According to the prevailing theory, vowel sounds generally mirror the resonance pattern of sound production in a direct way. However, objections brought forward in the literature (above all concerning limitations of formant and spectral shape estimation) and the many phenomena discussed in this treatise give reason to question such a direct mirroring.

Conceptually, from a physical perspective, the effect of a resonance pattern is quasi-independent of the source sound it transforms. Thus, within a purely physical concept of resonances, the observation that vowel-related spectral characteristics in general and the spectral envelope in particular (if its estimation is methodologically substantiated) of natural vowel sounds relate to the f_o of sound production – and that the spectral envelope is, therefore, an ambiguous representation of vowel quality – is hard to comprehend. Further experimental findings presented in this treatise, indicating that the observed relation of the spectral envelope of natural vowel sounds to the f_o of sound production is to be explained by the fact that the vowel spectrum is related to pitch (or to its alternative), accentuate the above statement: Pitch is not an acoustic characteristic and, therefore, the relation of the vowel spectrum to pitch cannot be understood within a primarily physical model such as the source–filter model of sound production.

The temptation to assume that, for natural sounds of a single speaker, the resonances of the vocal tract are directly adapted (related) to the pitch and f_o of the source sound within the process of sound production is confronted with both the observable variation extent and the nonuniformity of vowel quality-specific spectral characteristics, with and without f_o variation: The complexity level of adaptation that articulation would have to undergo in order to embrace the variation extent and nonuniformity of the vowel spectrum is so high that no speculation on a systematic source–filter interaction and filter adaptation to pitch and f_o within the existing concept of the prevailing source–filter theory should be asserted until a thorough experimental investigation of vowel sound production is carried out including an extensive variation of production parameters.

Methodologically, there is no basis for formant measurement for all recognisable vowel sounds.

Experimentally, as has been shown, recognition is often not based on consistent vowel-related patterns of spectral peaks that would directly reflect the resonance patterns of sound production. Also, recognisable vowel sounds can be synthesised either without any spectral peak structure and/or without a spectral fine structure that would allow for assessing a spectral envelope.

In these terms, the discovery of a vowel–pitch relation, the observable variation extent and nonuniformity of vowel-related spectral characteristics, the lack of methodological substantiation of formant measurement for all recognisable vowel sounds, the lack of evidence of formants being a perceptual cue for vowel recognition, the finding that vowel recognition for filtered sounds differed from the intended vowel qualities of the unfiltered sounds and that, then, they did not relate to all actual, vowel-related resonances of sound production, as well as the finding that synthesised vowel sounds produced outside the framework of the prevailing source–filter model are recognisable all stand against the understanding of specific vocal tract resonance configurations being always directly and imperatively mirrored in the radiated vowel sound. This leads to the assumption that the vocal tract resonance configuration is reflected in the produced vowel sound in a mediated way, a topic that has to be addressed and clarified in future research.

Although there is a long-standing and controversial debate on the relation between production and perception, an objection to understanding the vowel spectrum as directly mirroring the resonance pattern of sound production still seems provocative and difficult to accept. We will return to this matter in the next chapter. In this chapter, the discussion is limited to the exposition of the above counterarguments and to the examination and documentation of an additional spectral aspect that emerged in the course of analysing the spectra of natural vowel sounds and creating the documentation provided in Chapters 2.2 and 2.3 (vowel sounds produced at high f_o levels), 5.1 (breathy vowel sounds) and 7.3 (flat or sloping vowel-related spectral portions). While investigating such voiced and breathy sounds, we often observed noise and noise peaks manifest in the spectra parallel to the harmonic series. The noise peaks could be interpreted as a direct indication of a resonance characteristic of sound production. However, in numerous cases, one or several of these noise peaks did not correspond to the course and the peaks of the harmonic spectrum. To document this observation, a corresponding study was conducted: Based on the inspection of natural voiced and breathy sounds of the Zurich Corpus produced in V context, three sound samples were compiled, with all selected sounds being fully recognised in the standard listening test conducted when creating the corpus (100% vowel recognition rate matching vowel intention; note that further production parameters not explicitly given below were disregarded). The initial investigation and subsequent selection for exemplary documentation were focused on sounds for which the spectra manifested a contrast between the peak structure of noise and the peak structure of the harmonic spectrum (the relative energy maxima within the harmonic configuration or the frequencies and frequency distance between the harmonics) for a frequency range or a part of that range that is usually assumed as vowel-related.

The first sample consisted of sounds of the eight long Standard German vowels produced with breathy phonation or voiced phonation and low vocal effort at intended f_o ranging from 98–392 Hz. Sounds with these characteristics were investigated because they often exhibit a dominant first harmonic and subsequent sloping and/or flat parts of the harmonic spectrum, either for the entire vowel-related frequency range or a substantial part of that range. Therefore, if the noise spectrum manifested a more differentiated peak structure substantially above $H1$, a corresponding contrast could be demonstrated for the harmonic spectrum not directly mirroring the resonances indicated by the manifest noise.

The second sample consisted of sounds of the eight long Standard German vowels produced at intended f_o in the range of 440–587 Hz, manifesting a flat or sloping harmonic spectrum for the entire vowel-related frequency range or for a part of it, with peaks of noise occurring in that frequency range. Sounds with these characteristics were investigated for a similar reason to that mentioned above: If the spectrum manifested noise peaks but no corresponding peaks in the harmonic spectrum, the contrast in question could be demonstrated again.

The third sample consisted of sounds of the corner vowels /u, a, i/ produced at intended f_o of 784–880 Hz, whose spectra showed a noise peak either on a frequency level substantially below the frequency level of $H1$ or noise peaks in between $H1$ and $H2$ and/or $H2$ and $H3$. Sounds with these characteristics were investigated because of the high frequency level of $H1$ and the large frequency distance between the harmonics in the spectrum: If the spectrum exhibited a noise peak on a frequency level substantially below the level of $H1$ or a noise peak in between the low harmonics separated by a large frequency distance, the contrast in question could be demonstrated anew.

For the sounds of the three samples, during sound selection, occurring contrasts between peak structures of noise and harmonics were analysed, described accordingly, classified in terms of different types of spectral contrasts and interpreted with regard to possible indications of similarities or differences of *R*-patterns of sound production and estimated *P*-patterns or *F*-patterns of the produced sounds. Note in this context that interpreting occurring noise peaks as direct indications of resonances of sound production is a hypothetical approach, and if resonances of production are referred to in this chapter, then it is only in this hypothetical sense. The reflection and documentation put forward here only aim at serving as a basis for future experimental designs and clarification.

As a result of the inspection and analysis of the sound corpus and the subsequent sound selection, for the first compilation of breathy and voiced sounds produced with low vocal effort, four types of incongruent noise peaks and energy maxima of harmonics were observed:

A = Noise in the spectrum of sounds of back vowels and /a/ indicated two lower resonances < 1.5 kHz related to sound production; the harmonic spectrum did not exhibit a corresponding distinct spectral double-peak structure with corresponding frequency levels.

B = Noise in the spectrum of sounds of /a/ indicated a resonance in the frequency range of c. 1–1.5 kHz related to sound production (frequency range of statistical F_2); the harmonic spectrum did not exhibit a corresponding distinct second peak at a corresponding frequency level.

C = Noise in the spectrum of sounds of front vowels indicated a lower resonance < 1 kHz related to sound production; the harmonic spectrum did not exhibit a corresponding distinct peak at the corresponding frequency level.

D = Noise in the spectrum of sounds of front vowels indicated resonances > 1 kHz related to sound production in a frequency range usually considered vowel-related; the harmonic spectrum did not exhibit a corresponding distinct and marked peak structure with corresponding frequency levels of pronounced relative energy maxima of the harmonics.

Figure 1 illustrates these four types of spectral contrasts for sounds of the first sound sample.

For the second sample of sounds with flat or sloping harmonic spectra, three types of incongruent noise peaks and energy maxima of harmonics were observed:

E = Noise in the spectrum of sounds of /u/ indicated two lower resonances < 1.5 kHz related to sound production; the frequency level of *H*1 was equal to the frequency level of the first resonance, or it occurred in between the two lower resonances, or it was equal to the frequency level of the second resonance; *H*2 manifested a markedly lower level than *H*1 and was substantially above the frequency of the second indicated resonance.

F = Noise in the spectrum of sounds of /o/ indicated two resonances < 1.5 kHz related to sound production; the frequency distance between the first two or three harmonics was large, and one or both indicated lower resonances occurred in between the frequency levels of the lower harmonics.

G = Noise in the spectrum of sounds of a vowel (all vowels except /u/) indicated two or three peaks related to sound production, the peaks being in a frequency range or in parts of that range usually assumed as vowel-related; the harmonic spectrum in this frequency range or a part of it was either flat or sloping.

For the third sample of sounds produced at high f_o levels, three further types of incongruent noise peaks and energy maxima of harmonics were observed:

H = Noise in the spectrum of sounds of /u/ indicated two lower resonances < 1.5 kHz related to sound production; the frequency level of *H*1 was near or equal to the frequency level of the second indicated resonance and, therefore, the first indicated resonance was not represented in the harmonic spectrum (compare with type E).

I = Noise in the spectrum of sounds of /a/ indicated two resonances < 1.5 kHz related to sound production (in some cases close in frequencies); the frequency distance between *H*1 and *H*2 was large, and one or both of the resonances occurred in between the frequency levels of the lower harmonics (compare with type F).

J = Noise in the spectrum of sounds of /i/ indicated one lower resonance in the range of c. 450–550 Hz related to sound production; the frequency of *H*1 was equal to or above 750 Hz; therefore, the first indicated resonance was not represented in the harmonic spectrum.

Figure 2 illustrates the six types of spectral contrasts for sounds of the second and third sound samples.

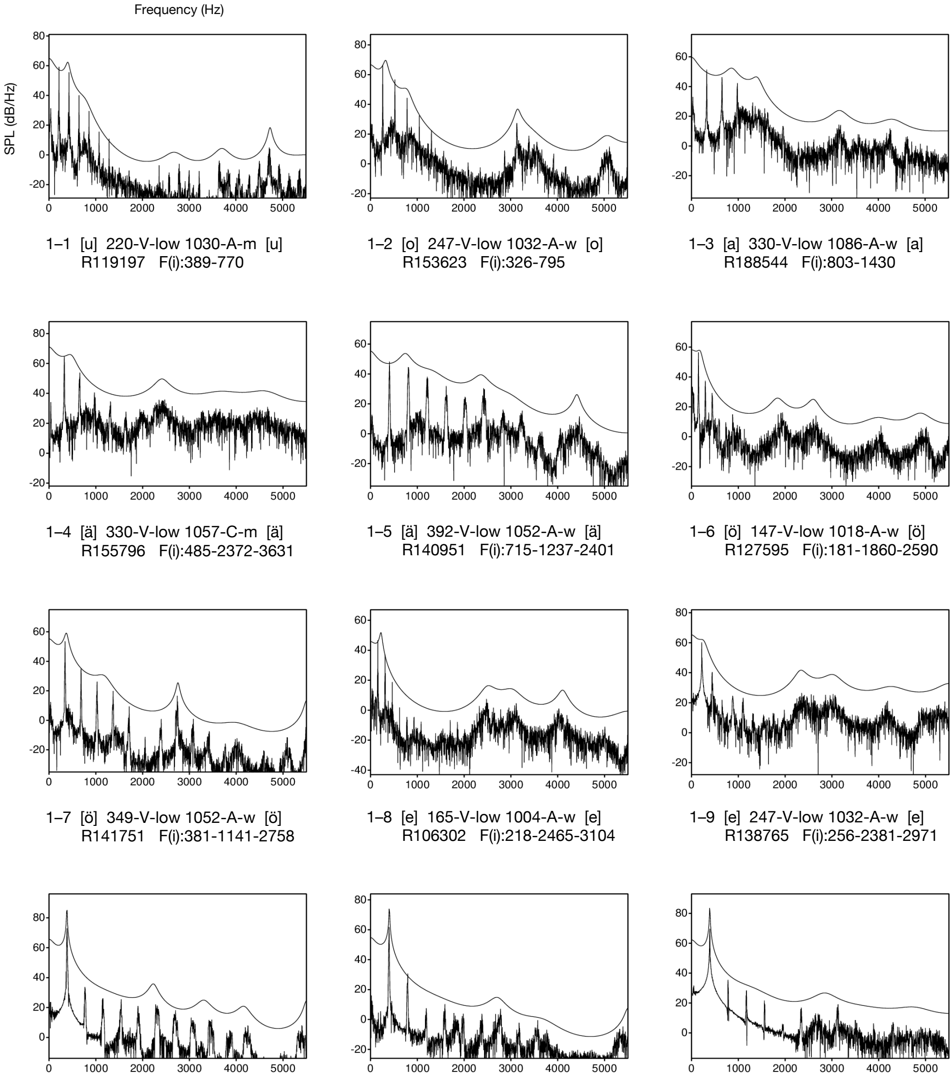
In these terms, numerous sound spectra were found and are documented here for which the peak structure of noise stood in considerable contrast to relative spectral maxima of harmonics or to their frequencies, for a frequency range usually considered vowel-related. If the noise peaks of the documented sounds indeed indicated the actual resonances of sound production, this would support the thesis that the harmonic spectrum does not, *in general,* mirror the resonances of sound production in a direct (unmediated) way. Consequently, numerous cases of vowel sounds with marked differences between vowel-related *R*-patterns of sound production and estimated *F*-patterns and/or patterns of spectral energy maxima would have to be expected to occur. This observation and reflection is transferred into a synthesis experiment in the next chapter.

For references, extended background information, details of experimental design, method and results and their documentation (table including sound links), see the Materials, Chapter M9.1.

Figure 1. Examples of sound spectra manifesting incongruent energy maxima of noise and harmonics according to the types A to D. Extract of Chapter M9.1, Table 1 (see Series 1–8 in this table). Sounds 1–3 = examples of types A or A combined with B. Sounds 4–12 = examples of types C or C combined with D or D.
[C-09-01-F01] ↗

Figure 2. Examples of sound spectra manifesting incongruent energy maxima of noise and harmonics according to the types E to J. Extract of Chapter M9.1, Table 1 (see Series 9–19 in this table). Sounds 1–6 = examples of types E, F and G. Sounds 7–12 = examples of types H, I and J.
[C-09-01-F02] ↗

Figure 1. Examples of sound spectra manifesting incongruent energy maxima of noise and harmonics according to the types A to D. [C-09-01-F01]



1–1 [u] 220-V-low 1030-A-m [u]
R119197   F(i):389-770

1–2 [o] 247-V-low 1032-A-w [o]
R153623   F(i):326-795

1–3 [a] 330-V-low 1086-A-w [a]
R188544   F(i):803-1430

1–4 [ä] 330-V-low 1057-C-m [ä]
R155796   F(i):485-2372-3631

1–5 [ä] 392-V-low 1052-A-w [ä]
R140951   F(i):715-1237-2401

1–6 [ö] 147-V-low 1018-A-w [ö]
R127595   F(i):181-1860-2590

1–7 [ö] 349-V-low 1052-A-w [ö]
R141751   F(i):381-1141-2758

1–8 [e] 165-V-low 1004-A-w [e]
R106302   F(i):218-2465-3104

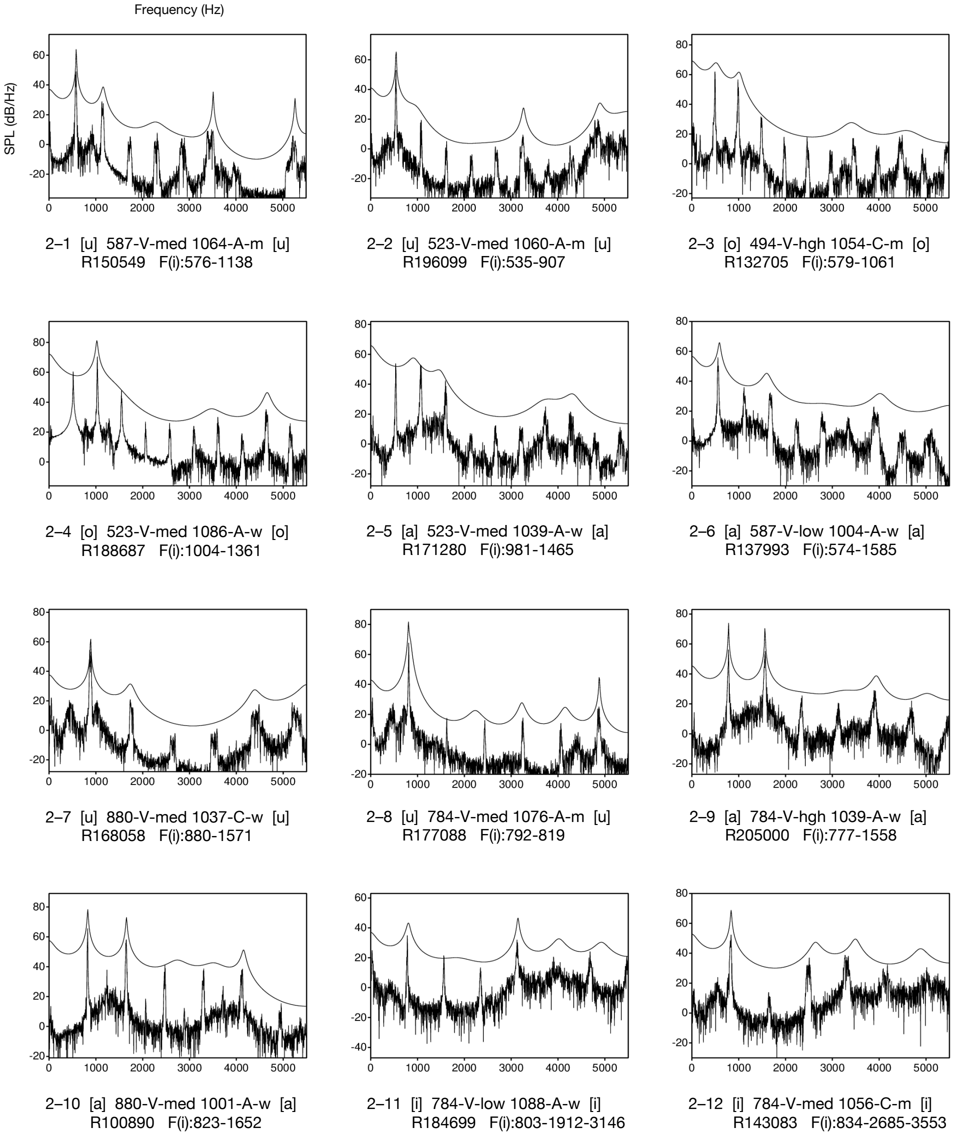
1–9 [e] 247-V-low 1032-A-w [e]
R138765   F(i):256-2381-2971

1–10 [ü] 392-V-low 1051-A-m [ü]
R153207   F(i):385-1360-2346

1–11 [ü] 392-V-low 1052-A-w [ü]
R140981   F(i):399-1328-2638

1–12 [i] 392-V-low 1018-A-w [i]
R127555   F(i):395-1144-2877

Figure 2. Examples of sound spectra manifesting incongruent energy maxima of noise and harmonics according to the types E to J. [C-09-01-F02]



2–1 [u] 587-V-med 1064-A-m [u]
R150549 F(i):576-1138

2–2 [u] 523-V-med 1060-A-m [u]
R196099 F(i):535-907

2–3 [o] 494-V-hgh 1054-C-m [o]
R132705 F(i):579-1061

2–4 [o] 523-V-med 1086-A-w [o]
R188687 F(i):1004-1361

2–5 [a] 523-V-med 1039-A-w [a]
R171280 F(i):981-1465

2–6 [a] 587-V-low 1004-A-w [a]
R137993 F(i):574-1585

2–7 [u] 880-V-med 1037-C-w [u]
R168058 F(i):880-1571

2–8 [u] 784-V-med 1076-A-m [u]
R177088 F(i):792-819

2–9 [a] 784-V-hgh 1039-A-w [a]
R205000 F(i):777-1558

2–10 [a] 880-V-med 1001-A-w [a]
R100890 F(i):823-1652

2–11 [i] 784-V-low 1088-A-w [i]
R184699 F(i):803-1912-3146

2–12 [i] 784-V-med 1056-C-m [i]
R143083 F(i):834-2685-3553

9.2 Resonance Patterns of Sound Production That Differ From Estimated Formant Patterns and Characteristics of the Harmonic Spectrum of Radiated Sounds

The observation that the spectra of natural voiced and breathy vowel sounds indicated noise peaks that sometimes did not correspond to the characteristics of the harmonic spectrum led to the question of whether vowel sounds can be produced by means of a vowel synthesis based on resonance or filter patterns that cannot be detected in the acoustic analysis of radiated sounds.

In the context of the methodological problems of *F*-pattern and spectral shape estimation, in the literature, a possible contrast between a resonance or filter pattern of sound production and its detection from the radiated sound is discussed above all concerning f_o levels of voiced sounds: As already mentioned, with increasing f_o, the resonance or filter curve is progressively undersampled, and sampling is poor for f_o levels above c. 300 Hz. Thus, for sounds at these middle or higher f_o levels, the detection of resonance characteristics of sound production is often methodologically unsubstantiated. However, the sampling problem is not uniform but depends on whether or not the harmonics of a sound spectrum match the resonance frequencies of sound production. To give an example for sounds of the vowel /a/: If two sounds are produced at equal f_o of 500 Hz but with two different *R*-patterns of 1150–1350–3000 Hz and 1000–1500–3000 Hz, respectively, estimated F_1 will markedly differ from R_1 of sound production for the first sound but not for the second, even if the frequency distance between the harmonics is the same. This is a consequence of *H*1–*H*2 matching *R*1–*R*2 for the second but not for the first sound.

Following this reflection and further developing the experimental design, the question of detectability of resonance or filter characteristics of radiated sounds as characteristics of sound production was addressed in a vowel synthesis model experiment: An attempt was made to synthesise two voiced-like sounds at an equal f_o level but based on two different *R*-patterns in such a way that

- the two radiated sounds manifested similar harmonic spectra and similar estimated *F*-patterns;
- for one sound, part of the measured *F*-pattern markedly deviated from the *R*-pattern of synthesis;
- for the other sound, the entire measured *F*-pattern corresponded with the *R*-pattern of synthesis;
- for both sounds, the harmonic spectra and estimated *F*-patterns were comparable to spectra of natural vowel sounds as documented in the Zurich Corpus.

If two sounds can be synthesised with two different *R*-patterns in such a way that the resulting harmonic spectra and measured *F*-patterns of the radiated sounds are similar, then cases occur for which resonances of production cannot be unambiguously detected on the basis of the radiated and perceived sounds. If, in addition, sound pairs of this kind are recognised as vowel sounds, then cases of *R*-patterns of sound production that are undetectable in the radiated sounds are relevant for an acoustic theory of the vowel.

Because of two experiences undergone in the course of preliminary synthesis attempts, two preconditions concerning f_o ranges and vowel qualities were established: The design of the present experiment was based on the observations reported for sounds produced at f_o of ≥ 400 Hz, limiting the upper f_o in synthesis to 700 Hz, and it addressed sounds of /u, o, a/ only. (For the rationale, see Chapter M9.2.) Within this limitation of f_o ranges and vowel qualities, in the first step, vowel synthesis based on various configurations of R_1–R_2 patterns and f_o levels was investigated by the author by means of a trial-and-error approach, attempting configurations according to the further developed experimental design described above. Bandwidths of these two lower resonances and spectral tilt were set individually for each *R*-pattern to bring the harmonic spectra and estimated formant frequencies of a sound pair close to each other. Based on the experiences of this first investigation, in a second step, eight exemplary pairs of configurations of *R*-patterns and f_o levels were created for final synthesis, acoustic analysis and a vowel recognition test, fulfilling the above conditions of the further developed experimental design. (For numerical details, including the higher resonances also used in synthesis, see Chapter M9.2, Tables 1 and 2.)

For every single configuration of a pair, three sounds were produced using a Klatt synthesiser in cascade mode (sound duration of 1 sec., including fade-in/fade-out of 0.05 sec.): Sound 1 was synthesised as a voiced-like sound with voiced source and lower levels of breathiness and aspiration; sound 2 was synthesised as a whispered-like sound with noise as the source and higher levels of breathiness and aspiration; sound 3 was again synthesised as a voiced-like sound with a voiced source but with higher levels of breathiness and aspiration. Sounds 1 and 2 of a single configuration were investigated concerning acoustic analysis and vowel recognition. Sound 3 and its spectrum only served as a graphic illustration of the contrast between resonance characteristics of sound production (in most cases visible based on the noise related to breathiness and aspiration in the spectrum) and the harmonic spectrum of the sound produced. As a result, for the eight pairs of *R*-patterns and related f_o levels, a total sample of 48 synthesised sounds was created, of which 32 sounds (16 voiced-like and 16 whispered-like sounds) were compiled for acoustic analysis and a vowel recognition test, and 16 additional sounds were used for the documentary purpose only. For an illustration of the experimental design, see Figures 1–3.

When designing the experiment, whispered-like sounds were included for two reasons: Firstly, in most cases, a synthesis with a noise source results in a sound that approximately reflects the resonances of production in the sound spectrum in terms of noise peaks at frequencies corresponding to the *R*-pattern of synthesis. Consequently, the spectral similarity or dissimilarity of two synthesised sounds based on a single *R*-pattern but with two different sources, periodic and noise, can be demonstrated graphically. Secondly, the vowel recognition for both types of sounds can be tested and compared with each other.

Acoustic analysis was conducted for the synthesised sounds according to the standard procedure of the Zurich Corpus, including a cross-check of the calculated *F*-patterns based on sound spectra, spectrograms and formant tracks. Vowel recognition of the sounds was tested in a listening test according to the standard procedure of the Zurich Corpus (forced choice, all long Standard German vowels and schwa, no vowel boundaries) and involving the five standard listeners.

According to the results of the acoustic analysis of the voiced-like sound pairs, for all eight pairs investigated, the first sound of a pair showed estimated F_1 or F_2 or F_1–F_2 markedly deviating from R_1 or R_2 or R_1–R_2 of synthesis, the differences between F_1 and R_1 being in the range of 131–328 Hz and the differences between F_2 and R_2 being in the range of 94–251 Hz. Conversely, the second sound of a pair showed corresponding F_1–F_2 and R_1–R_2, the differences between F_1 and R_1 being in the range of 0–33 Hz and the differences between F_2 and R_2 being in the range of 4–44 Hz. Finally, a comparison of the estimated *F*-patterns of both voiced-like sounds of a pair showed that these patterns also matched, the differences for F_1 being in the range of 6–22 Hz and the differences for F_2 being in the range of 0–13 Hz. Likewise, the harmonic spectra of the two sounds also corresponded with each other.

According to the results of the acoustic analysis of the whispered-like sound pairs, in contrast to the voiced-like sounds, the difference of *R*-patterns of sound production with noise as the source was mostly reflected in a corresponding difference of the estimated *F*-patterns when comparing the two sounds of a pair.

According to the labelling majority of the vowel recognition test, the recognised vowel qualities of the voiced-like sounds of a pair corresponded to each other and to the qualities of the natural voiced sounds imitated in their spectral characteristics for six of the eight series. The same held true for the remaining two series concerning a vowel boundary of the natural voiced sounds imitated. In contrast, for all sounds of the two back vowels investigated and the corresponding pairs of configurations of *R*-patterns, the vowel quality of at least one whispered-like sound differed from that of the related voiced-like sound. Moreover, for five of the six pairs of whispered-like sounds recognised as back vowels, the vowel qualities also differed according to the two different *R*-patterns of sound production compared.

In these terms, exemplary cases of voiced-like sound pairs of a vowel could be synthesised and are demonstrated here for which the vowel-related production resonances of one sound could be detected based on the radiated and perceived sound but could not be detected for the other sound, with equal f_o levels and quasi-equal harmonic spectra and estimated *F*-patterns for the two sounds. Thus, two different *R*-patterns can result in voiced or voiced-like sounds of a vowel with similar harmonic spectra and similar estimated *F*-patterns. Thereby, for the first sound of a pair, the differences between R_1 and/or R_2 and F_1 and/or F_2 equalled or exceeded the differences for estimated statistical average F_1 and F_2 as often given in the literature for sounds of the two adjacent back vowels /o/ and /u/ produced by adults.

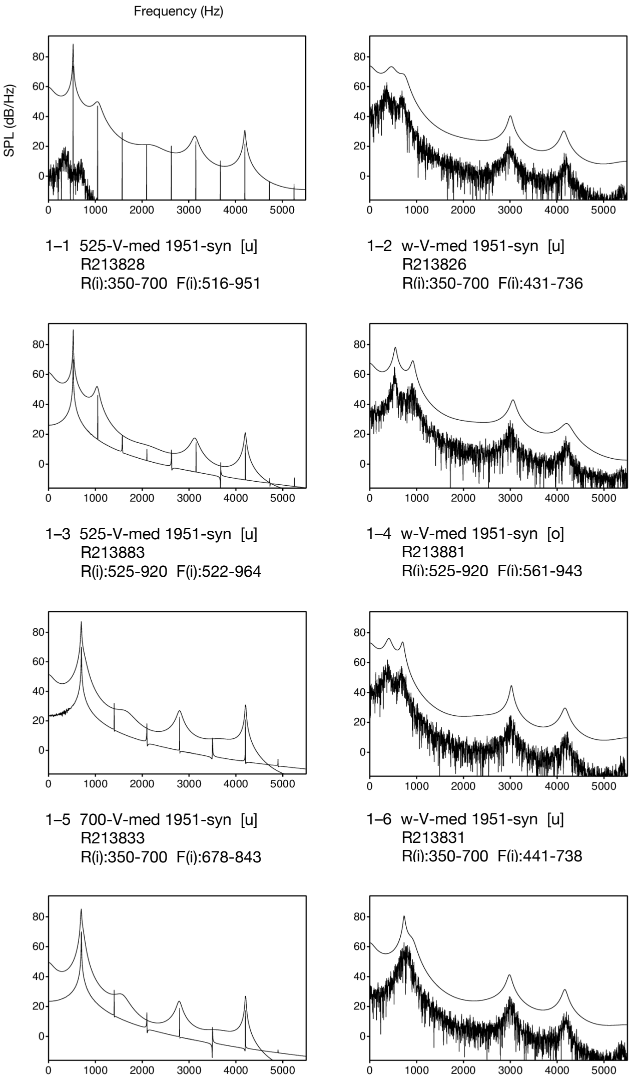
For references, details of experimental design, method and results, an extended discussion, including some relativisations and indications for future research as well as additional observations on pitch levels of whispered-like sounds, and for documentation of results (tables including sound links), see the Materials, Chapter M9.2.

Figure 1. Comparison of synthesised voiced-like and whispered-like vowel sounds based on two different *R*-patterns, whereby the voiced-like sound pairs were recognised as /u/. Extract of Chapter M9.2, Tables 1–3 (see Series 2 and 3 in these tables). Sounds 1–4 = first sound comparison. Sounds 1 (voiced-like) and 2 (whispered-like) were synthesised based on R_1–R_2 of 350–700 Hz (for higher resonances and estimated $F_{(i)}$, see Chapter M9.2) and sounds 3 (voiced-like) and 4 (whispered-like) were synthesised based on R_1–R_2 of 525–920 Hz. f_o of synthesis was 525 Hz. The voiced-like sounds (sounds 1 and 3) were recognised as the same vowel /u/ and manifested similar harmonic spectra and similar estimated *F*-patterns despite the differences in the *R*-patterns of sound production. In contrast, the whispered-like sounds (sounds 2 and 4) were recognised as different vowels /u/ and /o/ and manifested different spectral envelopes related to the different *R*-patterns. (Note also the marked pitch difference for the two whispered-like sounds; for details, see Chapter M9.2.) Sounds 5–8 = second comparison of voiced-like and whispered-like sounds, with comparable findings. Synthesis was based on R_1–R_2 of 350–700 Hz and 700–840 Hz, respectively, and f_o of synthesis was 700 Hz. Whispered-like sounds were recognised as /u/ and /ɔ/.
[C-09-02-F01] ↗

Figure 2. Comparison of synthesised voiced-like and whispered-like vowel sounds based on two different *R*-patterns, whereby the voiced-like sound pairs were recognised as corresponding to the vowel boundaries of /ɔ–o/ and /o–u/. Extract of Chapter M9.2, Tables 1–3 (see Series 5 and 6; note the corresponding estimated $F_{(i)}$ in these tables). Sounds 1–4 = first comparison of pairs of voiced-like and whispered-like sounds, with comparable findings as described in Figure 1. Synthesis was based on R_1–R_2 of 350–1000 Hz and 570–1000 Hz, respectively, and f_o of synthesis with a voiced-like source was 500 Hz. The voiced-like sounds (sounds 1 and 3) were recognised as corresponding to the vowel boundary of /ɔ–o/. The whispered-like sounds were recognised as /u/ and as corresponding to the vowel boundary of /ɔ–o/. Sounds 5–8 = second comparison of pairs of voiced-like and whispered-like sounds, with comparable findings. Synthesis was based on R_1–R_2 of 700–800 Hz and 500–1000 Hz, respectively, and f_o of synthesis with a voiced-like source was 500 Hz. The voiced-like sounds (sounds 5 and 7) were recognised as corresponding to the vowel boundary of /o–u/. The whispered-like sounds were recognised as /ɔ/ and /o/.
[C-09-02-F02] ↗

Figure 3. Comparison of synthesised voiced-like and whispered-like vowel sounds based on two different *R*-patterns, whereby the voiced-like sound pairs were recognised as /a/. Extract of Chapter M9.2, Tables 1–3 (see Series 7 and 8; note the corresponding estimated $F_{(i)}$ in these tables). Sounds 1–4 = first comparison of pairs of voiced-like and whispered-like sounds, with comparable findings as described in Figure 1 except for vowel recognition of the whispered-like sounds, which were recognised as /a/ in accordance with the voiced-like sounds. Synthesis was based on R_1–R_2 of 1150–1350 Hz and 1000–1500 Hz, respectively, and f_o of synthesis with a voiced-like source was 500 Hz. Sounds 5–8 = second comparison of pairs of voiced-like and whispered-like sounds, with comparable findings as for sounds 1–4. Synthesis was based on R_1–R_2 of 1200–1300 Hz and (again) 1000–1500 Hz, respectively, and f_o of synthesis with a voiced-like source was 500 Hz.
[C-09-02-F03] ↗

Figure 1. Comparison of synthesised voiced-like and whispered-like vowel sounds based on two different *R*-patterns, whereby the voiced-like sound pairs were recognised as /u/. [C-09-02-F01]



1–1 525-V-med 1951-syn [u]
R213828
R(i):350-700 F(i):516-951

1–2 w-V-med 1951-syn [u]
R213826
R(i):350-700 F(i):431-736

1–3 525-V-med 1951-syn [u]
R213883
R(i):525-920 F(i):522-964

1–4 w-V-med 1951-syn [o]
R213881
R(i):525-920 F(i):561-943

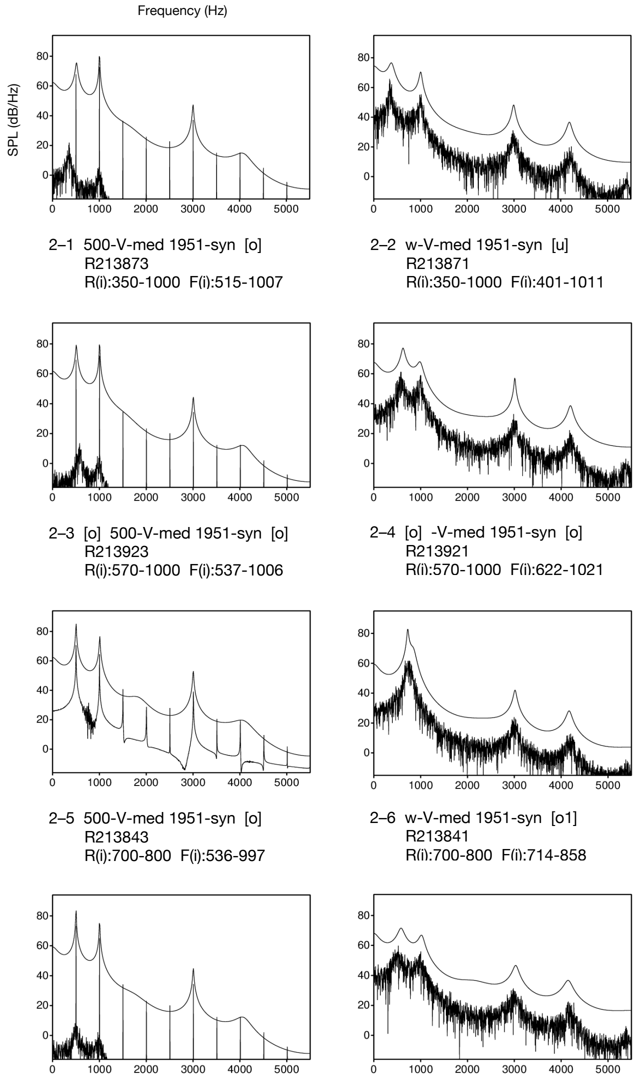
1–5 700-V-med 1951-syn [u]
R213833
R(i):350-700 F(i):678-843

1–6 w-V-med 1951-syn [u]
R213831
R(i):350-700 F(i):441-738

1–7 700-V-med 1951-syn [u]
R213888
R(i):700-840 F(i):709-824

1–8 [u] w-V-med 1951-syn [o1]
R213886
R(i):700-840 F(i):709-824

Figure 2. Comparison of synthesised voiced-like and whispered-like vowel sounds based on two different *R*-patterns, whereby the voiced-like sound pairs were recognised as corresponding to the vowel boundaries of /ɔ–o/ and /o–u/. [C-09-02-F02]



2–1 500-V-med 1951-syn [o]
R213873
R(i):350-1000 F(i):515-1007

2–2 w-V-med 1951-syn [u]
R213871
R(i):350-1000 F(i):401-1011

2–3 [o] 500-V-med 1951-syn [o]
R213923
R(i):570-1000 F(i):537-1006

2–4 [o] -V-med 1951-syn [o]
R213921
R(i):570-1000 F(i):622-1021

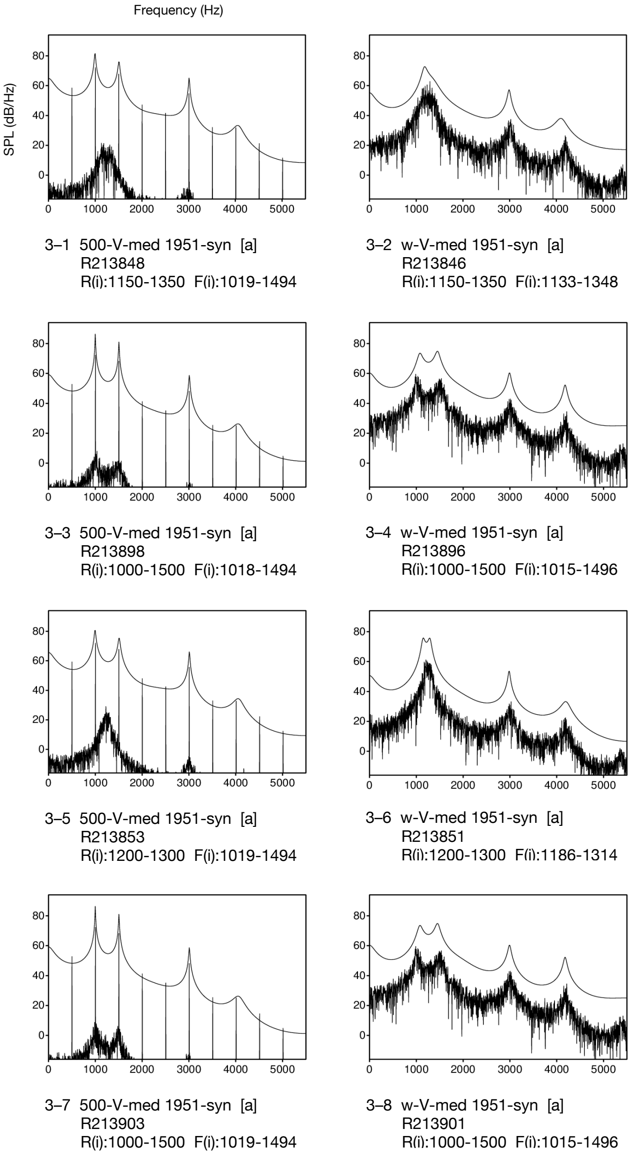
2–5 500-V-med 1951-syn [o]
R213843
R(i):700-800 F(i):536-997

2–6 w-V-med 1951-syn [o1]
R213841
R(i):700-800 F(i):714-858

2–7 500-V-med 1951-syn [o]
R213893
R(i):500-1000 F(i):516-1013

2–8 [o] -V-med 1951-syn [o]
R213891
R(i):500-1000 F(i):580-1024

Figure 3. Comparison of synthesised voiced-like and whispered-like vowel sounds based on two different *R*-patterns, whereby the voiced-like sound pairs were recognised as /a/. [C-09-02-F03]



3–1 500-V-med 1951-syn [a]
R213848
R(i):1150-1350 F(i):1019-1494

3–2 w-V-med 1951-syn [a]
R213846
R(i):1150-1350 F(i):1133-1348

3–3 500-V-med 1951-syn [a]
R213898
R(i):1000-1500 F(i):1018-1494

3–4 w-V-med 1951-syn [a]
R213896
R(i):1000-1500 F(i):1015-1496

3–5 500-V-med 1951-syn [a]
R213853
R(i):1200-1300 F(i):1019-1494

3–6 w-V-med 1951-syn [a]
R213851
R(i):1200-1300 F(i):1186-1314

3–7 500-V-med 1951-syn [a]
R213903
R(i):1000-1500 F(i):1019-1494

3–8 w-V-med 1951-syn [a]
R213901
R(i):1000-1500 F(i):1015-1496

9.3 Conclusion

According to the prevailing acoustic theory, vowel-related spectral characteristics are understood within the perspective of speech production in terms of a source sound and its transformation by a vowel-related resonance pattern. Such an understanding of the vowel sound supposes that the vowel spectrum directly reflects the resonance pattern of production (of articulation). From this perspective, perception (recognition) plays a minor part. However, as shown, a phenomenological investigation of natural vowel sounds revealed a high complexity level and nonuniform character of vowel-specific spectral variation and a relation of the vowel spectrum to f_o. As discussed, both phenomenological findings challenge the prevailing source–filter concept. Further, and most importantly, the discovery that the relation of the vowel spectrum to f_o observed in natural vowel sounds is but an indication of a vowel–pitch relation (or its alternative) calls for a reconsideration of the role of perception and recognition for both sound production and acoustic characteristics. Thus, the question of whether the produced vowel sound directly reflects the resonance pattern of production or whether the reflection of resonances in the produced sound is mediated, involving a perceptual process – this process yet to be discovered – is posed.

In the context of the investigation, documentation and argument of this treatise, and in accordance with earlier claims made in the literature, the documentation of natural vowel sounds and the results of the synthesis experiment presented in this main chapter again raise doubts about a *general* detectability of resonances of vowel sound production based on an acoustic analysis of the produced sound. This calls for a thorough experimental revisitation of the actual processes of vowel sound production and perception (or vice versa), embracing an extended variation of sound production parameters. However, an investigation of these processes requires preceding progress in understanding general vowel-related acoustic characteristics beyond the indices put forth in this treatise.

Part III Commentary

The third part of this text formulates and discusses the primary and secondary indices derived from the documentation and experimentation presented and reflects on a future theory of the acoustics of the vowel. Also included is a third excursus.

10 Indices

10.1 A Text of Transition

As noted in the Introduction, this treatise is a transitional text. It resulted from the attempt and effort to formulate knowledge-based statements about the acoustic representation of vowel quality *in general* and, at the same time, provide corresponding evidence that can be apprehended and reproduced.

“Given the relatively long history of work on vowels, it may be tempting to regard the topic of vowel acoustics as basically settled and closed in contemporary science and practice.” (Kent and Vorperian, 2018) Indeed, this holds true if we only look at the “surface” of the literature on the matter and consult general textbooks. Most textbooks state that the formant pattern of a vowel sound, *in general,* represents its vowel quality acoustically: “The primary acoustic characteristic of vowels is the location of the formant frequencies, specifically, the first three formants (F1–F3). […] For a given speaker or a given speaker group of speakers with the same vocal tract length, each vowel is associated with a distinct acoustic formant frequency pattern.” (Reetz and Jongman, 2020, p. 207–208) (Note that although there is an explicit and extensive discussion in the specialist literature on the alternative option that either the formant pattern or the spectral shape of a vowel sound acoustically represents vowel quality – see e.g. Hillenbrand and Houde, 2003, and Swanepoel et al., 2012 –, this issue is rarely mentioned in introductory textbooks or the literature of other scientific disciplines.)

But from the very beginning of the spectrographic investigation and the related observations up to now, we have learned that a core characteristic appertains to the vowel sound that not only resists being grasped within a general theory of the formant pattern or the spectral shape representing vowel quality but also contradicts this general theory. (We will return to this matter in the Afterword.) We have written the Preliminaries and part of this treatise to describe, document, discuss and reference the main reasons for this conclusion. Yet, as we have to leave the prevailing theoretical ground, we do not have an alternative theoretical basis at our disposal. Therefore, we are constrained to refer to our phenomenological experience and knowledge of vowel sounds, their various types and modes of production, their calculated fundamental frequencies and spectra and the recognition of vowel qualities and pitch, and to attempt to formulate a few basic statements that apply to all recognisable vowel sounds *in general.* By doing so, we aim to contribute to the intellectual discourse, phenomenological knowledge and designs and methods of investigation with regard to a future theory building of the acoustics of the vowel.

In the present treatise, the basic statements are called primary indices of the actual acoustic characteristic representing vowel quality, this actual characteristic being undiscovered up to now. The secondary indices consist of the knowledge achieved within the prevailing framework of formant patterns or spectral shapes. Here, they are called secondary because they do not concern all recognisable vowel sounds *in general,* but they concern vowel sounds with limitations both in their parameters of production and in the methodological substantiation of their spectral analysis.

10.2 Primary and Secondary Indices

Firstly, the vowel sound relates to pitch (or to a comparable perceptual referencing to a sound pattern repetition over time). Secondly, the vowel sound is a kind of perceptual and acoustic foreground–background phenomenon. Thirdly, the spectral representation of vowel quality is nonuniform. These are the three main conclusions drawn from the documentation and the various experimental results presented.

The two characteristics that the vowel sound is a kind of perceptual and acoustic foreground–background phenomenon and that the spectral representation of vowel quality is nonuniform were already subjects of investigation and discussion in the Preliminaries. They are further elucidated here based on the new phenomenological database, the Zurich Corpus, and based on new experiments. The provided evidence of a vowel–pitch relation (or its alternative) is new.

These conclusions are assumed here to appertain to vowel sounds and their quality *in general* and to be the three primary indices for a future acoustic theory of the vowel. They are exposed in detail in the subsequent chapters, based on the present investigation of long Standard German vowels.

Furthermore, on the basis of the prevailing theory and related literature, we know that (i) spectral envelopes of natural voiced vowel sounds produced as monophthongs either by children or by women at lower f_o levels in their modal voice range (comparable to the range of relaxed speech) or by men at lower or middle f_o levels of their voice range exhibit a course with relative spectral energy maxima and minima, (ii) if such sounds are produced with medium vocal effort, for each of the three speaker groups and their f_o levels separately and with highly limited f_o variation, LPC analysis in most cases produces vowel-specific filter frequency patterns, although the estimation of these patterns cannot be completed without the intervention of the investigators (expertise in phonetics and sound analysis, parameter settings related to the vocal tract size of the speakers, visual spectrographic crosscheck, manual correction of LPC parameters and formant tracks), (iii) for these sounds produced with medium vocal effort, a source–filter resynthesis based on their estimated filter patterns and f_o levels in most cases produces replicas with maintained vowel qualities, that is, the natural sounds can be reproduced artificially using a (voiced-like) source and a filter. Also, according to the literature, phonation type and vocal effort variation have an additional impact on the vowel-related spectral characteristics mentioned, as can be true for the sound context and the dynamic character of natural sounds. As said, these conclusions drawn from the existing literature only apply to some of the vowel sounds. Therefore, they are understood here as representing but secondary indices for a future acoustic theory of the vowel.

10.3 The Relation of Recognised Vowel Quality to Pitch or a Comparable Perceptual Reference

Within the empirical limitations of vowel and pitch recognition experiments and the interpretation of results, with very few exceptions, the recognition of vowel quality for voiced and voiced-like sounds in our studies was indicated to relate to pitch directly. The experiments presented in Chapters 6.9 and 6.10 (and the corresponding details in Chapters M6.9 and M6.10) may serve as a paradigmatic demonstration of that relation because of the occurring parallel shifts of vowel qualities and pitch levels for sound manipulations for which spectral maxima were kept unchanged, these manipulations and associated recognition shifts including transitional sounds with double-vowel and/or double-pitch recognition. The many experiments and their results presented in the preceding chapters – the demonstration of the f_o range of recognisable vowel sounds and the relation of the sound spectrum of a vowel to f_o, the diverse demonstrations regarding formant pattern and spectral shape ambiguity, and the results of vowel and pitch recognition for whispered sounds, for sinewave replicas based on statistical *F*-patterns or estimated *F*-patterns of single natural sounds and for synthesised sounds with a variation of HCF – document the experimental and intellectual path leading to the conjecture and subsequent investigation of interrelated double-vowel and double-pitch recognition. Further, the presented experiments and their results also supported the notion of a perceptual vowel–pitch relation for natural whispered and synthesised whispered-like sounds. However, this relation is still to be investigated in more detail.

The question of the vowel–pitch relation (or its alternative) being uniform or nonuniform among vowel sounds is left open here. According to the results of the experiments conducted, pitch variation had no or only a minor impact below 200 Hz, and above that frequency level, it affected sounds of close and close-mid vowels more strongly and in a more systematic manner than open-mid or open vowels. However, this may be reformulated in future research. (On this matter, some indications are given in the following third excursus.) Moreover, the course of the spectral envelope and, with it, the foreground–background character of the vowel sound also has to be taken into account.

The question of the vowel–pitch relation (or its alternative) sometimes being dissociable is also left open here. However, according to the results of the experiments conducted, pitch level shifts without vowel quality shifts were often observed. Inversely, cases of vowel quality shifts without pitch level shifts were very rare.

To conclude, the indications resulting from the presented experiments that vowel quality recognition is related to pitch are very pronounced. However, because of the phenomenological perspective adopted here, it is left open whether vowel and pitch perception and recognition are indeed linked directly to each other or whether further differentiations are needed to describe the perceptual process in its referencing to sound characteristics and sound pattern repetition. In these terms, it is said here that recognised vowel quality relates to pitch or, as an alternative, to a comparable perceptual referencing to a sound pattern repetition over time.

10.4 Recognised Vowel Quality as a Kind of Perceptual and Acoustic Foreground–Background Phenomenon

If, in our study and related to the vowel qualities investigated, sounds of a close or close-mid front unrounded vowel were LP filtered with stepwise decreasing CFs, in most cases, the recognised vowel quality shifted to a front rounded vowel and then to a back vowel, maintaining vowel openness. In some cases, a direct front–back shift was also observed. Further, for the sounds of the close-mid front unrounded vowel, subsequent LP filtering in most cases resulted in an additional open–close shift. If sounds of a close or close-mid front rounded vowel were LP filtered, in most cases, the vowel quality shifted in a front–back and, possibly, in a subsequent open–close direction. If sounds of the front unrounded open-mid vowel were LP filtered, the vowel quality in most cases shifted first to the open vowel and, subsequently, in a front–back and open–close direction. If sounds of /a/ were LP filtered, the vowel quality shifted in an open–close direction. The same held true for some of the sounds of /o/. Within the limits of the general front unrounded–rounded, front–back and subsequent open–close shift directions, however, some variation of single vowel quality shifts for sounds of a given vowel was also observed.

Concerning both vowel quality recognition and vowel quality-related spectral characteristics of LP-filtered natural sounds, confirming earlier indications reported in the literature, the experiments and their results presented in this treatise provide evidence that sounds recognised as belonging to two vowel qualities can be filtered so as to result in sounds of only one recognised quality. Consequently, the lower vowel-related spectral energy of the sounds of the two vowels can be similar, while the subsequent higher spectral energy is then vowel-specific.

Reflecting on the findings of LP filtering and the resulting general shift directions of vowel quality, four additional aspects should be considered. Firstly, vowel quality shifts often occurred independently of the presence or absence of (expected) vowel-related spectral maxima of unfiltered sounds. Secondly, concerning vowel recognition, natural sounds of close-mid and close front unrounded vowels required a spectral representation surpassing c. 2.3 kHz, sounds of close-mid and close front rounded vowels required a spectral representation including the frequency range of c. 1.3–2.3 kHz, sounds in the range of /a–ɑ/ required a spectral representation including the frequency range of up to 1.5–2 kHz somewhat depending on middle or higher f_o levels, and sounds of back vowels required a spectral representation including the frequency range of up to 1 kHz or up to H_1 (/u/) or H_1–H_2 (/o/) for middle or higher f_o levels. Spectral ranges for sounds of /ɛ/ were hard to establish because of very pronounced spectral variations. However, as an approximation, sounds of this vowel required a spectral representation surpassing 2 kHz. Thirdly, as mentioned above, occurring variations of vowel quality shifts for sounds of a given vowel indicated that the specific individual distribution of spectral energy in the frequency range of a given sound has to be taken into account for shift prediction. Fourthly, after analysing the cases of unrounded–rounded, front–back and open–close shifts for LP-filtered sounds, a scheme of quality shifts can be derived that allows for a successful prediction of most filtered sounds: This scheme resembles the traditional vowel quadrilateral (ignoring the generally assumed relation of the vowel positions to $F1$–$F2$).

In this context, the results of the HP-filtering experiment of vowel sounds have also to be considered: If HP filtering is looked at from the perspective of stepwise increasing CFs from lower to higher frequency levels, as shown, natural sounds produced as close front vowels were indicated to first shift to close-mid or even open-mid vowels before they often reverted these shifts. Thus, close and close-mid (and sometimes open-mid) vowels can have a similar acoustic representation that is reflected in the higher part of the vowel spectrum, and they then differ only in relation to the preceding difference in lower frequencies.

In conclusion, LP filtering indicates that sounds of two vowels can have similar lower vowel-related spectral energy and that, then, they differ only in the subsequent higher spectral energy. Inversely, HP filtering indicates that sounds of two vowels can have similar vowel-related higher spectral energy and that, then, they differ only in the lower spectral energy. In these terms, it is said here that from a perceptual and acoustic perspective, the vowel sound is a kind of foreground–background phenomenon.

10.5 Nonuniform Spectral Representation of Vowel Quality

In our studies, vowel quality was shown to be recognisable for sounds with highly diverse spectral characteristics: With very different f_o and/or pitch levels, with or without spectral maxima, with many or with very few harmonics or even without a harmonic structure at all and so forth. Further, the conducted experiments often produced results that differed for different production parameters of the sounds – e.g. f_o levels and ranges of level variation, vocal effort, phonation type – and also for different vowel qualities investigated. Finally, the experiments and their results, including the corresponding documentation presented in this treatise, provide evidence that no average spectrum in terms of a spectral template can be identified as approximately representing all sounds of a given vowel, even if such a template were related to f_o. Also, evidence is provided that the findings and insights gained from an experiment based on only a limited set of vowels and/or a limited set of production parameters often do not allow for a generalisation applicable to other sounds of other vowels and/or other production parameters. In these terms, it is said here that the spectral representation of recognised vowel quality is nonuniform.

Excursus – Speculations

Introduction

In this treatise, we present a decided phenomenological and descriptive examination and discussion of the acoustic representation of vowel quality and the related basic perceptual aspects that have to be accounted for. In doing so, we relate to traditional methods of spectral analysis of natural vowel sounds and to vowel (re-)synthesis and sound manipulation experiments derived from or motivated by the descriptive findings of natural sounds and the questions arising thereof. But that is not to say that, during the many sound recordings made, the many acoustic analyses performed and scrutinised, the many experiments conducted and the many listening tests performed, and during the reflections made when reading the literature, we in the research team did not develop and discuss ideas exceeding the limitations of phenomenology, description, traditional spectral perspective and related experimental investigation.

Although these ideas – here referred to as speculations – underlie some of the documentation and experiments, for three main reasons, their exposure is kept separate from the main line of argument of the treatise: (i) They shall not blur the formulation of indices as statements of the acoustic representation of vowel quality *in general,* for which evidence is given, (ii) they were and are a matter of controversial debate within the research team, and (iii) there was and still is a lack of tools for a corresponding experimental investigation and verification. However, the speculations may be useful for future theory building and are therefore exposed in this excursus, with the main objective of sharpening the general question of the acoustics of the vowel posed in this treatise.

Sound-internal perceptual referencing to a repeating pattern and pattern-internal vibration form

Given a general vocal range of recognisable vowel sounds – how is it to be understood that, on the one hand, vowel quality is independent of pitch, in the sense that the same vowel quality can be produced and recognised on all pitch levels, but that, on the other hand, (at the same time) the perceptual recognition process relates to pitch (or its alternative)?

Speculating on this question, vowel quality may turn out to be a human discovery in that a sound quality and its possible quality-related contrast graduation and oppositions can be produced that are consciously recognised (quasi-)independently of variations in attributes commonly understood as attributes of the cardinal sound dimensions of loudness, pitch and timbre, and to some extent also independently of sound length. Since this quality conjectured here concerns sounds that can be produced in isolation and that, then, have somewhat (quasi-) monotonous acoustic and perceptual characteristics, it may turn out that the sounds always relate to a repeating pattern – even though the repetition may be very approximative and of limited duration, and the related sound spectra may not always manifest a clear harmonic structure or may not manifest such a structure at all, as is the case with many creaky and whispered sounds. (Diphthongs are considered here as a secondary phenomenon related to monophthongs.) Quality-related contrast graduation and oppositions may turn out to relate to the temporal extension of the pattern and, therefore, graduation and oppositions may turn out to be relational: Given a vocal range of recognisable vowel sounds, it may be possible to produce the same kind of vibration form subdivisions of the type of pattern in question for different temporal extensions of the pattern.

Thus, with respect to vowel sounds, we reflect on the idea of the possible production of a (repeating) sound pattern whose recognised vibration form relates to its temporal extension, this relation being at the core of the speculation put forward here. The relational characteristic in question would explain why

- the perceptual process for assessing vowel quality has to involve pitch (the recognition of the repetition of a vibration form as a pattern being the basis for the recognition of that pattern);
- consciously, a particular vowel quality is independent of pitch (the vowel quality-specific vibration form being in relation to its temporal extension and, therefore, that relation being transferable to other temporal extensions);
- in some cases of single sounds, two vowels and two pitch levels can be recognised (the sounds being ambiguous with respect to pitch level and, as a consequence, also with respect to vowel quality).

To reiterate in other terms, vowel quality may turn out to refer to a repeating vibration form as a repeating sound pattern of a particular kind, of which the temporal extension of the pattern directly relating to its vibration form is constitutive: Vowel quality may prove to refer to a sound dimension that allows (formal) articulation of contrasting vibration forms of a temporal extent. Notably, this is hardly an aspect of sound timbre comparable to other aspects commonly subsumed under the term timbre. Rather, sounds of this kind may raise the question of a separate system of vibration forms that are perceived and recognised as categorically distinct from all other aspects of sound characteristics, that is, a separate sound dimension (see below). This would explain why no instrument (comparable to a musical instrument) can play vowel sounds as humans can produce them.

According to this speculation, as stated, the repetition of the vibration form in question would relate to pitch (or its alternative) in the perceptual process of sound analysis, but the relational character of the vibration form (as a form in relation to a temporal extension) makes the conscious recognition of sound quality independent of pitch. From this perspective, the pitch quality (or its alternative) is "external" to the quality graduation of the sound pattern in question. Nevertheless, it is needed for the identification of the repeating pattern. Only vowel quality is "internal" to the pattern (in contrast to the prevailing theory, which claims that a sound-external form, a filter curve, is superimposed on an actual source sound). This would explain the seeming contradiction formulated above.

Following this idea, roughly speaking and formulated from a spectral perspective, it is a spectral energy distribution and a relation between lower and higher sound energy *within* an identified (repeating) vibration pattern that characterise the vowel sound (the temporal extension of the pattern being a constitutive part of it). To give a first simplified example: Sounds of /u/ have a vibration form that allows perception to refer to a sparse articulation of undulation within the repeating sound pattern, but sounds of /i/ have a vibration form that allows perception to refer to a very pronounced articulation of undulation within the pattern. From a spectral perspective, the perception and recognition of sounds of /u/ refer to a low vibrational frequency within the repeating pattern, while the perception and recognition of sounds of /i/ refer to a high vibrational frequency. Therefore, for all pitches of recognisable vowel sounds, the sounds of /i/ cannot be recognised if spectral energy of > c. 2 kHz is not present, and the sounds of /u/ cannot be recognised if spectral energy of < c. 2 kHz is not present. (Note that the recognised vowel quality of the sounds of /i/ was often maintained in HP filtering with a CF of c. 1 kHz; see Chapter 8.2.) Furthermore, the (repeating) patterns, spectral energy distribution and energy maxima of the sounds of /u/ at lower pitch levels may differ markedly from the sound patterns of this vowel at middle and higher pitch levels, and the same is true for the sounds of /i/. Yet, for sounds perceived at equal pitch levels, the vowel quality-related contrast of the vibration form for /u/ and /i/ within the patterns is always the same. (For illustration, see Figure 1.)

In a second example (see Figure 2), two sounds of /a/ are added to the sounds of the first example to address the question of the nonuniform character of vowel quality-related spectral characteristics. In contrast to the sounds of /u/ and /i/, the sounds of /a/ have a vibration form that allows perception to refer to an intermediate articulation of undulation within the repeating sound pattern, manifest in a spectral energy distribution in an intermediate frequency range relevant for vowel sounds. At the same time, in contrast to the natural sounds of /u/ and /i/, the range of prominent spectral energy and the spectral energy maxima (if manifest) of sounds of /a/ at lower and higher pitch levels are often comparable. On the one hand, this may be related to the fact that vowel sounds have two maximum opposites: /u/ versus /i/ (close back versus close front) and /u/ and /i/ versus /a/ (close versus open). On the other hand, and interrelated, this may also be understood in the context of the dichotomy of the vowel spectrum: "[…] while the dependence of vowel-specific spectral characteristics and formants on fundamental frequency for the lower frequency range ≤ 1.5 kHz is easily understandable and reproducible empirically, this is not the case for the higher frequency range" (Preliminaries, p. 67). In this sense, the perception and recognition of sounds of /a/ may be related to an intermediate (vowel quality-related) articulation of the vibration form of a pattern as a second contrast reference for a sparse articulation (/u/) and a very pronounced articulation (/i/), and because of this intermediate articulation, the associated range of spectral energy distribution and energy maxima may not be related to pitch variation in the same way as is the case for sounds of close vowels.

In a third example (see Figure 3), a sound of intended /u/ and a sound of intended /i/ produced with a very low vocal effort are shown to address the question of the foreground–background character of the vowel sound. According to the results of the standard listening test conducted when creating the corpus, the first sound of intended /u/ was recognised as /i/ by one of the five standard listeners of the Zurich Corpus, and a second listener recognised a "mixture of /u/ and /i/". The other three listeners labelled /u/. The sound of intended /i/ was recognised by one of the five standard listeners as /u/. The other four listeners labelled /i/. As far as the sound wave and the repeating period are concerned, visually, there is hardly any identifiable difference between the two vibration patterns. Looking at the sound spectrum, the difference is small, especially at the higher frequencies of > 2 kHz. Thus – although rarely observed for unmanipulated natural vowel sounds recognised by listeners with no reported hearing impairment – vowel perception and recognition can refer to two different (formal) articulations of the vibration form of a pattern (or, in spectral terms, to different frequency ranges and ratios of prominent spectral energy). Here, this is understood as the result of the vowel sound being a kind of foreground–background phenomenon. (Notably, several back–front or front–back confusions of individual listeners were observed in the resynthesis or synthesis experiments presented in this treatise; likewise, when LP filtering sounds of front vowels, CF levels causing recognised front–back shifts were found to be listener-specific.)

In a fourth example (see Figure 4), a natural reference sound of intended /i/ produced by a man with medium vocal effort in nonstyle mode at an intended f_o of 220 Hz, and four HP-filtered variants of that sound with CFs of 440–660–990–1320 Hz are shown to illustrate further the foreground–background character of the vowel sound. The natural reference sound was recognised by all five standard listeners of the Zurich Corpus as /i/. According to the labelling majority, the filtered variants were recognised by these listeners as follows (in the order of the above CFs): /e/ – /ɛ/ – /i/ – /i/. Thus, perception and recognition of vowel sounds refer to the configuration of higher spectral energy (here, the unfiltered frequencies > 1.5 kHz) dependent on the spectral energy configuration of the lower frequencies: In the present example, HP filtering of the natural reference sound of a close vowel with stepwise increasing CFs initially resulted in a vowel quality shift in a close–open direction, which was subsequently reversed. Therefore, the above formulation that the perception of sounds of /i/ refers to a high frequency of vibration needs further differentiation, and this differentiation concerns the vowel sound as such: The perception of a vowel sound refers to lower and higher frequencies of vibration as interrelated.

We refrain from giving further examples because we cannot propose a more precise formulation of the actual reference to which the perceptual process refers when analysing a vibration form of a repeating pattern of this (supposed) kind, that is, when analysing a particular vowel quality and its contrast with other vowel qualities. However, it is important to always have in mind that the reference to a vibration form is a perceptual reference: Thus, depending on the types of sounds and/or the individual conditions and analytic strategies of the listeners, different vowels and/or different pitch levels for the same sound may sometimes be recognised by different listeners, and some – but not all – listeners may sometimes recognise two vowels and/or two pitch levels for a single sound. (Note again that, in this view, vowel recognition of whispered and creaky sounds is assumed to also include a perceptual referencing operation to a sound pattern repetition over time.)

Obviously, what is missing in this account is an explanation of the actual perceptual and acoustic sound dimension and the actual system of sound quality differentiation to which perception refers when assessing (assumed) vibration form contrasts of sound patterns. What is missing is a new phonetic theory of the vowel.

An aspect among others of sound timbre or a sound dimension?

The definitions of cardinal perceptual sound dimensions and their acoustic correlates are a matter of debate. On the one hand, it is common to understand a sound as being perceived by its length and by the three cardinal dimensions of loudness, pitch and timbre. (Occurring interactions of these sound characteristics are ignored here.) On the other hand, especially concerning sound timbre, there is no clear and scientifically satisfying definition (see e.g. Bregman, 1990, pp. 92–93), and the phenomena discussed and subsumed under the term timbre is of high complexity (see e.g. McAdams and Giordano, 2009; Siedenburg and McAdams, 2017). However, for the below argument, we will adopt an undifferentiated "naive" perspective by referring to the definitions of loudness, pitch and timbre as given by the American National Standards Institute (ANSI, 2013). For the present discussion of the vowel sound and as an argumentation strategy – with the intention of bringing forward a possible categorical difference between the recognised vowel quality and perceptual aspects commonly discussed as aspects of sound timbre – it may prove to be sufficient to do so.

As stated, it is common to understand a sound as being perceptually characterised by its length and by three cardinal dimensions of loudness, pitch and timbre. In parallel, it is common to understand the sound length and the three cardinal perceptual dimensions as being related to acoustic characteristics such as sound duration over time, sound pressure level, fundamental frequency and (static or dynamic) spectral specificity of a sound. Accordingly, sound length is commonly understood as the perceptual attribute of auditory sensation based on which sounds can be ordered from short to long, and according to the American National Standards Institute, loudness is defined as the perceptual attribute of auditory sensation based on which sounds can be ordered from quiet to loud, pitch is defined as the perceptual attribute based on which sounds can be ordered from low to high, and timbre is defined as the perceptual attribute based on which two sounds with the same length, loudness and pitch can be differentiated.

In the literature on vowel sounds, the terms vowel timbre and vowel quality are commonly used as synonyms, vowel quality being understood as belonging to the dimension of sound timbre. This understanding has a long tradition back to “Die Lehre von den Tonempfindungen” by von Helmholtz (1863), who explicitly subsumed the vowel sound as a phenomenon of musical timbres (“musikalische Klangfarben”) and discussed it in the context of timbres of different musical instruments (see von Helmholtz, 1863, pp. 113–181). (For a more recent example of the terms vowel quality and vowel timbre being used as synonyms, see Hillenbrand and Houde, 2003; for an exception, see the below citation of Titze, 2000.)

However, as indicated above, there are reasons for the claim to categorically separate vowel quality from aspects of sound timbre and to understand vowel quality as a phenomenon of a category of sound attributes on its own: Above all for sounds of long vowels, it is in the very “nature” of vowel quality to be of the abstract kind, that is, to be recognisable for shorter and longer and for soft and loud sounds, for sounds at different pitch levels, for sounds produced by different speakers with different phonation types, different production styles and speaking modes and so forth, and even for sounds that can only be produced in sound synthesis. It is in the very “nature” of vowel quality to be recognisable independent of all of these sound aspects, that is, independent of length, loudness, pitch and timbre, independent of the actual sound production. (An exception is the difference between sounds of long and short vowels, which we consider here as a secondary phenomenon. Therefore, in our argumentation, we refer to long vowels in the first instance: Long vowels are recognisable independent of a highly limited sound length, and consequently, they can be sung with pronounced length variation.) Accordingly, Titze (2000, p. 281), in introducing a discussion of vocal register, states: “The perception of voiced sounds is sometimes reduced to four dimensions: pitch, loudness, vowel (or voiced consonant), and quality. The last of these dimensions, quality, is a poorly defined term that includes all the leftover perceptions after pitch, loudness, and phonetic category have been identified.” (Note the terminological usage of Titze, which is not in accordance with our own use of the terms; note also in this context the differentiation of timbre and identity of vowel sounds made by de Cheveigné and Kawahara, 1999; see the corresponding citation in the introduction to Chapter M9.2).

From this perspective, sound timbre may be a reference for speaker identification but not for vowel differentiation: Speakers have individual sound timbres, but not individual vowel qualities. If they had different vowel qualities, language could not have emerged. Similarly, sound timbre may often be a direct reference for sound production, but not a direct reference for vowel differentiation: Different production parameters such as phonation types, vocal effort, pitch, register, nasalisation, production styles, speaking modes and so forth relate to different sound timbres, but not to different vowel qualities. Individual speech sounds have different timbres in many respects, but they do not have individual vowel qualities. Once again, if vowel quality belonged to these aspects of timbre, language could not have emerged.

However, the following counterargument to the above argumentation may be brought forward: As McAdams and Giordano (2009, p. 72) state, "[timbre] is one of the primary perceptual vehicles for the recognition, identification, and tracking over time of a sound source (singer's voice, clarinet, set of carillon bells)". Indeed, vowel perception and recognition and vowel production must be fused, and pronunciation must be learned by blending phonation and articulation of sounds and heard vowels of a language, and vice versa. Based on the perception and recognition of vowel qualities of natural vowel sounds, the qualities can be reproduced by phonation and articulation. In this sense, the source of the vowel quality of a vowel sound can be identified (in terms of its reproducibility) based on vowel recognition. Yet, the matter is complex. Above all, there is no separation of production and perception for vowel sounds: When they are produced, perception and production are one. Thus, perceptual sound source identification (related to vowel quality) does not follow production. It is inherent in production. Vowel sounds do not in simple terms reflect the physical and physiological conditions of sound production which would then be "traced" by perception (they do not reflect the resonances of the vocal tract in a direct and unmediated way; in this context, the physiological difference between children and adults with respect to phonation and articulation is also worth noting, since language learning is not affected by this difference; see also the reference to Ohala, 1996, given in the introduction to Chapter M9.1). It is in view of the foregoing that we speculate that sound timbre may be a reference for the identification of sound sources from the outer world and sounds from various instruments used to produce them, but the vowel sound is not comparable to the sounds of instruments: As stated, there is no instrument (other than digital devices) that can produce vowel sounds in a manner comparable to the human voice. In this sense, the voice is not comparable to a musical instrument.

In conclusion, vowel quality is recognised (quasi-)independently of aspects commonly understood as aspects of sound timbre, in a comparable and similarly “abstract” manner to the way in which pitch is recognised. If pitch perception and recognition are understood as being categorically distinct from the perception and recognition of other sound characteristics, why should we not understand vowel quality perception and recognition accordingly? If sound characteristics of pitch constitute a dimension of sound, why should we not consider the sound characteristics of vowel quality as a dimension of sound, too? If pitch is defined as the perceptual attribute based on which sounds can be ordered from low to high, why should we not say that vowel quality is defined as the perceptual attribute based on which sounds can be ordered within a vowel system (its quality contrasts, relations, and oppositions), e.g. from perceptual attributes that replace the production-related attributes close to open, back to front and unrounded to rounded? Finally, as will be discussed below, since pitch level differentiation can be represented within notation systems, why should we not say that the same is true (in a comparable manner) for vowel quality differentiation?

In these terms, we conjecture that the prerequisite for recognising vowel quality is that it be categorically distinct from other perceptual aspects of timbre in order to achieve its status as an element of speech and language. Our conjecture that vowel quality is a perceptual phenomenon of a repeating sound pattern of a particular kind, for which the temporal extension of the pattern and its vibration form directly related to this extension are constitutive of quality recognition, has to be considered from this perspective.

A note on the distinction between “speaker quality” and “phonetic quality” made by Ladefoged

Ladefoged (1967, pp. 56–62) raised the question of whether vowel quality is a sound quality on its own, and he discussed this question at length. For two reasons, we will elaborate on his argumentation: Firstly, reference has to be made to his claim that vowel sounds are of a different kind than most sounds of musical instruments, thus expressing a somewhat different view of the vowel sound than many other scholars in the tradition of von Helmholtz. Secondly, however, having posed this most central question of the vowel sound, his assertion that the recognised vowel quality is a peculiar phenomenon of perception only but not of the acoustic characteristics of the sounds draws his reflection on the matter to a close without providing sufficient evidence.

Below, longer excerpts from his text are quoted, as his book may not be available to the reader.

He begins his consideration of the particular status of recognised vowel quality with a differentiation of the terms phonetic quality and personal quality: "[…] we must note that speech sounds can be equated in a way peculiarly their own. Many other sounds can be classified according to three separate factors: their pitch, their loudness, and their quality. Thus we can disregard loudness and quality and compare the pitch of two sounds, considering only whether they are on the same note or not; similarly we can compare their loudness irrespective of the other two factors; and we can also compare their quality irrespective of everything else. It might appear that in comparing the vowel sounds of a soprano and a bass we are assessing their quality. But it is not as simple as this; in the case of speech sounds we can consider pitch, loudness and two sorts of quality, which we may label phonetic quality and personal quality. Thus when we say that two vowels are different we usually do not mean to imply anything about their pitch and loudness, nor about their personal quality, but only that they differ with regard to that one aspect of their quality which we term phonetic quality. […] Thus when listening to speech sounds observers can assess, with a high degree of convergence, four variables: loudness, pitch and two aspects of quality.

"Most people cannot consider any other sounds in terms of four variables. It is only in the case of speech sounds that nearly everybody can say that they may be similar as regards one aspect of quality, but different in another.

"This point has often been overlooked in discussions of vowel quality. There is a tendency to assume that variations in the personal quality of the vowels of two speakers are of the same kind as differences in phonetic quality." (pp. 56–57)

"The distinction between phonetic quality (the attributes of the auditory sensation that enable a phonetician to consider a speech sound as part of a sociolinguistic system) and personal quality (some other features of the auditory sensation) is one of the basic assumptions of phonetics. […] there is the dichotomy between personal quality and phonetic quality. As we have defined it, this opposition operates on the sociolinguistic level of analysis." (p. 61)

Ladefoged points out that it is possible to perceive two sounds as similar with respect to a particular aspect commonly attributed to sound timbre – namely, vowel quality – but at the same time perceive the two sounds as different with respect to another aspect also commonly attributed to sound timbre – namely, the speaker-specific characteristic. In his example of the vowel sounds of a soprano and a bass, however, it is unclear whether he is referrring to an individual voice characteristic of a woman and a man or to a (singing) style-specific characteristic of voice production that is actually not a personal characteristic but a characteristic of a production style. Reflecting on his example, moreover, vowel quality is recognised not only for different speakers or different production styles, but also for different phonation types, vocal efforts, registers, for nasalisation and so forth, aspects that can themselves be recognised independently of each other and of vowel quality (see also below).

To return to Ladefoged's statements: He claims that vowel quality is an auditory sensation related to a sociolinguistic system, unlike the other sound characteristics that are not related to that system. Thus, vowel quality is a sound characteristic that is coded as part of a linguistic system. Accordingly, the code is the only perceptual specificity of this sound characteristic compared to other characteristics.

Ladefoged continues, however, by insisting – seemingly – on a general difference between vowel sounds and other sounds: "When we assess the quality of a speech sound we need a frame of reference which cannot easily be applied to the assessment of the quality of any other kind of sound. We cannot, for instance, consider differences in the quality of two violins playing the same note as being in any way comparable with the differences between two people saying the same vowel. An expert violin maker can easily tell violins apart by what might be called the 'personal' quality of each violin. But this does not mean that there are two dimensions of quality for violins. […] By no means is it possible to represent two different dimensions of perceived quality of musical instruments." (pp. 57–58) (However, see the below discussion of sounds of organs, for which Ladefoged claims two different dimensions of quality.)

This consideration can be interpreted as a statement that sound characteristics generally relate to three perceptual variables (three sound dimensions), namely loudness, pitch and sound quality (sound timbre), but the characteristics of vowel sounds (and other speech sounds) in particular relate to four variables (four sound dimensions), namely loudness, pitch, sound quality and phonetic quality. As indicated above and discussed again in the below note on sound systems and their notation, the sound dimensions of sound length, loudness, pitch and timbre are usually differentiated not only because of categorically distinct perceptual attributes but also because of categorically distinct acoustic correlates, that is, sound duration versus sound intensity versus sound pattern repetition over time versus (static or dynamic) spectral differences of sounds. Sound dimensions are understood as being in parallel perceptually and acoustically different in this categorical sense.

Questioning whether the perceptual difference in the quality of speech sounds compared to other types of sounds is related to a corresponding specific acoustic sound characteristic, Ladefoged states: “It should be noted that we are not saying that there is anything peculiar about vowel sounds considered as physical entities. It is only in the way that they are normally perceived that they differ from other sounds. We could no doubt learn to assess other sounds in terms of two kinds of quality. Indeed, it is possible that this is actually done by some observers. A musician considering organs could possibly tell organ 1 from organ 2, irrespective of the stops; in such a case it would be possible to represent the perceived quality in two dimensions, the quality of being organ 1, or organ 2, being shown in one dimension, and the quality of the different stops (diapason, violin, trombone, etc.) being represented in the other. However, in practice musicians are not often concerned with this kind of assessment; whereas phoneticians are continually specifying speech sounds in terms of their phonetic quality. Nearly all phonetic theory relies on the tacit assumption that it is possible to recognize two kinds of quality. Nevertheless, this fundamental point is rarely considered and the different kinds of quality are seldom explicitly distinguished.

“The peculiar position of speech sounds is due to their being habitually assessed as part of a means of communication. Every speaker has learnt to separate personal quality from phonetic quality as a result of his constant experience of the sociolinguistic system. Phoneticians, who are trained observers of sociolinguistic systems, have become highly skilled in their assessments of the different kinds of quality. We may conclude that when a phonetician equates vowels spoken by different voices he does so because in the appropriate sociolinguistic system there could be no difference of codified information conveyed by the differences in the sounds.” (pp. 59–60)

By answering this way, he closes the question: Vowel quality is not a sound dimension comparable to loudness and pitch, but merely one aspect among others of sound timbre with the specificity of being a learned encoded aspect of timbre within a sociolinguistic system. There is nothing special about vowel sounds per se with respect to their sound characteristics. Through learning, other types of sounds can also be perceptually differentiated and comparably coded.

However, this assertion also needs to be thoroughly reconsidered. According to the example of Ladefoged for a musical instrument, an actual sound may be recognisable as the sound of a specific type of instrument (an organ), a specific single instrument (organ 1 or 2) and a specific stop applied during sound production. In the example, therefore, the sound of one organ may be recognised as different from that of another organ regardless of the stops selected, and vice versa. That is, two sources (this term used in a broad sense, not in the sense of source and filter) of sound production of the same kind can be perceptually held apart independent of the different modes of sound production applied, and vice versa.

If this is considered in a broader sense, what other examples can be brought into this context? For machines, the sounds of two washing machines of the same brand and type may be recognised by a technician as different regardless of a selected wash program (and conversely, the wash program may be recognised independently of the two machines). In human voices, the sounds of two children may be recognised as different by their parents regardless of whether they are nasalised (and conversely, nasalisation may be recognised independently of the two individuals). In a further abstraction, concerning human voices, the sounds of two children may be recognised as different by their parents regardless of the emotional expression of joy or anger or sadness (and conversely, the emotional expressions may be recognised independent of the individuals).

But these examples only indicate that some aspects of the sound timbre of two actual sounds of the same kind can be distinguished independently of other actual aspects of timbre, and that this may relate to a specific ability and learning process of a listener. Yet, if so, this is hardly a question of separate sound dimensions comparable to loudness and pitch. The fact that vowel quality is recognised not only for different speakers or different production styles, but also for different phonation types, vocal efforts, registers and so forth, and that these aspects of sounds can in turn be recognised independently of each other and of vowel quality, has to be considered in this context of single identifiable aspects of sound timbre.

Returning to Ladefoged's answer to the question about the specific sound character of the vowel sound as the result of a sociolinguistic code and its acquisition, this view corresponds to a structuralist perspective: “[…] les sons offriraient-ils par eux-mêmes des entités circonscrites d'avance? Pas davantage. La substance phonique n'est pas plus fixe ni plus rigide; ce n'est pas un moule dont la pensée doive nécessairement épouser les formes, mais une matière plastique qui se divise à son tour en parties distinctes pour fournir les signifiants dont la pensée a besoin. Nous pouvons donc représenter le fait linguistique dans son ensemble, c'est-à-dire la langue, comme une série de subdivisions contiguës dessinées à la fois sur le plan indéfini des idées confuses […] et sur celui non moins indéterminé des sons […]. Le rôle caractéristique de la langue vis-à-vis de la pensée n'est pas de créer un moyen phonique matériel pour l'expression des idées, mais de servir d'intermédiaire entre la pensée et le son, dans des conditions telles que leur union aboutit nécessairement à des délimitations réciproques d'unités. La pensée, chaotique de sa nature, est forcée de se préciser en se décomposant. Il n'y a donc ni matérialisation des pensées, ni spiritualisation des sons, mais il s'agit de ce fait en quelque sorte mystérieux, que la ‘pensée-son’ implique des divisions et que la langue élabore ses unités en se constituant entre deux masses amorphes.” (de Saussure, 1916/1995, pp. 155–156)

(For the English translation, see de Saussure, 1916/1959, p. 112: “[…] would sounds by themselves yield predelimited entities? No more so than ideas. Phonic substance is neither more fixed nor more rigid than thought; it is not a mold into which thought must of necessity fit but a plastic substance divided in turn into distinct parts to furnish the signifiers needed by thought. The linguistic fact can therefore be pictured in its totality – i.e. language – as a series of contiguous subdivisions marked off on both the indefinite plane of jumbled ideas […] and the equally vague plane of sounds […]. The characteristic role of language with respect to thought is not to create a material phonic means for expressing ideas but to serve as a link between thought and sound, under conditions that of necessity bring about the reciprocal delimitations of units. Thought, chaotic by nature, has to become ordered in the process of its decomposition. Neither are thoughts given material form nor are sounds transformed into mental entities; the somewhat mysterious fact is rather that “thought-sound” implies division, and that language works out its units while taking shape between two shapeless masses.”)

From this perspective, the sound characteristics produced by the human vocal organ are considered amorphous (“matière plastique”, “plan indéterminé des sons”, “masse amorphe”), not themselves offering any entities circumscribed in advance (“des entités circonscrites d'avance”), at least as far as vowel quality is concerned. But is this evident in all aspects mentioned?

Before further approaching this question, and to avoid misunderstandings: It is evident that a specific vowel quality of a given language as an entity of that language is not circumscribed in advance. If it were so, there would be little reason for the multitude of languages and vowel systems. A particular vowel quality of a particular language is determined within a language-specific system of qualities and their contrasts and opposites as a coded position within that system. So it is a result of the development and state of that language.

But to say that an actual vowel quality is determined within a language-specific system of qualities and their contrasts does not imply that the contrast building itself (as a structural aspect) is language-specific nor that the perceptual and acoustic reference is undetermined: Different languages have different vowel systems, but (probably) all languages have vowel contrasts that are structurally inherent in all vowel systems, that is, not the results but the basis of a code. Yet, vowel contrast building needs a frame of reference, both perceptually and acoustically. This being the case, the specific assumption of the structuralist perspective is addressed here, that the sound characteristics produced by the human vocal organ are amorphous, not themselves offering any contrast building in advance. Is this evident?

For many scholars, such doubt may not arise as long as their assumption can be confirmed that "The formants of a [vowel] sound […] are directly dependent on the shape of the vocal tract and are largely responsible for the characteristic quality. […] when there is a vibration in the rate at which pulses are produced by the vocal cords, there will be a change in the pitch of the sound (although there will be no change in the formants, and hence no change in the characteristic vowel quality)." (Ladefoged, 1996, pp. 94–95, and 99) According to this assumption, the acoustic and perceptual reference of the vowel sound is the physical specificity of sound production (that is, the physical specificity of resonances), comparable to any other type of physical sound production, and the field or range of sounds of all possible variations and combinations of formants (peaks in the sound spectrum) is indeed not structured, but amorphous (although the peaks as an assumed reference structure is a predefinition). Yet, the assumption is empirically contradicted: As extensively shown and discussed here, there is no acoustic representation of vowel quality by formants *in general,* as there is no acoustic representation of vowel quality independent of pitch. On the contrary, the perceptual and acoustic characteristics of vowel quality are in principle pitch-related (or related to an alternative sound characteristic comparable to pitch), and this represents not only a perceptual and acoustic "circumscription in advance" but also a very special sound condition of human utterances, for which it has not yet been clarified whether there is any other sound of the same kind at all, or whether any other type of sound allows for a system of contrast formation comparable to the vowel systems, including the notation systems associated with it.

In these terms, it is not evident to say that “there is nothing peculiar about vowel sounds considered as physical entities”. And if it is not evident, the question of whether vowel quality is a sound dimension in its own right remains open.

Pitch and vowel quality as comparable sound dimensions? A note on sound systems and their notation

As stated, pitch is commonly considered a sound dimension both in perceptual and acoustic terms.

With respect to perception, pitch enables (i) the general differentiation of a sound as lower or higher than another sound, (ii) the assessment of the extent of the pitch level difference in terms of pitch intervals, and (iii) the assessment of octave equivalence. Based on these perceptual aspects, pitch levels and intervals can be organised into a scale and coded in terms of systems of levels and their interrelations. While an actual system of pitch levels and intervals is a result of a code and depends on the tradition and actual context and state of the system, the perceptual basis itself – the structural basis of level differences, level scaling and level equivalences – is not coded, but it is a condition for the code. There is a (structural) perceptual “circumscription in advance”.

With respect to acoustics, there is no simple theory for correlating pitch levels, pitch intervals and octave equivalence with acoustic sound characteristics. However, for most types of vowel sounds, there is some kind of describable sound pattern repetition over time that can be identified in the sound wave, even if that repetition may be far from regular (as is the case e.g. for creaky vowel sounds), and pitch levels, intervals and octave equivalence can in most cases be estimated with regard to a repetition rate over time, to whatever extent the approximation of this estimation may be. Thus, there is also an acoustic “circumscription in advance” for pitch recognition of vowel sounds of this type, that is, an acoustic characteristic of the sound wave pitch refers in a (quasi-)systematic way and in a (quasi-)categorical difference to other acoustic sound characteristics that are not perceived as pitch. (The question of the pitch of whispered vowel sounds and also of synthesised vowel sounds that lack harmonicity in the sound spectrum, e.g. sinewave vowel sounds related to statistical *F*-patterns, is left open here.)

This parallelism of perceived sound quality and acoustic sound characteristics enables not only pitch notation systems, but also corresponding systems of acoustic characteristics. (Therefore, musical instruments emerged with which the acoustic characteristics related to pitch could be produced accordingly.) In this sense, the pitch notation systems that have emerged throughout history testify to pitch as a perceptual and acoustic sound dimension all of its own.

Is there reason to consider perceived vowel quality as comparable to pitch, and therefore to consider it as a perceptual and acoustic sound dimension in its own right?

With respect to perception, vowel quality enables (i) the general differentiation of a sound as being in a quality contrast to another sound of its kind, independent of loudness and pitch and other perceptual sound attributes, (ii) the assessment of the extent of this difference (vowel “intervals”), and (iii) the assessment of maximal quality contrasts (opposites). Based on these perceptual aspects, vowel qualities and quality differences can be organised into a schema and coded in terms of systems of vowels and their interrelations. While an actual system of vowels and vowel "intervals" and oppositions is a result of a code of a language and depends on the tradition and actual context and state of the system, the perceptual basis itself – the structural basis of vowel quality differences, vowel “intervals” and opposites – is not coded, but it is a condition for the code. There is again a (structural) perceptual “circumscription in advance”, and comparable to pitch, vowel sounds also allow for notation systems, which in their turn have an uncoded (structural) basis.

With respect to acoustics, we have only some indications of the correlation of differentiation, “intervals” and oppositions of vowel qualities with acoustic sound characteristics. These indications are described in this treatise. But because of the lack of an acoustic theory of the vowel, the question of whether there is an acoustic characteristic of the vowel sound that is categorically distinct from all other possible acoustic characteristics of sounds cannot yet be answered. However, the question is not closed either: The fact that notation systems for vowel quality differentiation may bear witness to vowel quality as a perceptual and acoustic sound dimension all of its own, and our conjecture that vowel quality is a phenomenon of a repeating sound pattern of a special kind, of which the temporal extension of the pattern and its vibration form directly related to this extension are constitutive, offers an option to reflect on such a particular category of sound characteristics.

With regard to the conjectured structural basis for the codes of the two notation systems of pitch levels and vowel qualities, note that some aspects of pitch notation systems are manifest among many different individual systems, such as e.g. pitch level scaling and octave equivalence: "[…] the use of discrete pitch relationships, as well as the concept of octave equivalence seem, while not universal in early and prehistoric music (Nettl 1956; Sachs 1962), rather common to current musical systems (Burns, 1999)." (Honingh and Bod, 2011) Likewise, some aspects of vowel systems and their symbolised representation also have an almost universal character, since they are generally manifest in different individual systems, such as e.g. quality "intervals" and opposites: "Probably every language uses at least three distinct vowels. […] Languages that have only three vowels usually have sounds that can be symbolized i, a, o or i, a, u." (Ladefoged, 2005, p. 175)

A phenomenon of produced form of expression, an "intervention" of perception and recognition

In the present context, probably the most difficult question to pose and to discuss is the question of a produced form of expression that cannot be derived in all its very constitutive characteristics from a physical model or a physiological condition.

Referring to the considerations of the first excursus: Within the framework of the source–filter model, the only attempt to explain why a specific timbre or quality of a sound may not be recognised by the resonance characteristic of its production is the decreasing resolution ("undersampling") of the resonance curve with increasing f_o and the resulting detection problem, including f_o levels that exceed the statistical frequency levels of the assumed first resonance of the sounds of a vowel in relaxed speech. However, the vowel spectrum does not deviate from the assumed vowel-specific resonance curves or formant patterns or spectral shapes because of poor spectral resolution or f_o exceeding lower F_1, but it deviates in principle from these assumed vowel-specific characteristics. This is the core phenomenon that has to be faced and questioned when reflecting on the vowel sound: How can the deviation from the prevailing prediction be understood? Why does the attempt to derive the *general* acoustic representation of vowel quality from a physical and physiological model of source and filter lead to predictions that can be empirically falsified?

There is a missing link, and this link concerns perception and recognition.

It is neither a vowel-specific resonance characteristic nor a vowel-specific spectral peak structure (in the proper sense of these terms) of the radiated sound to which the recognition process directly refers. Several authors of previous studies have taken that stand, and the Preliminaries and the present treatise once again provide empirical evidence for this rejection. Thus, the perception and recognition of vowel quality are not simply reflecting the conditions and predefined acoustic characteristics of the sound-producing organ or apparatus in that they only serve to differentiate these sound characteristics. Perception and recognition of vowel quality are indicated to act as an agent, that is, they are actively involved in the production of these sound characteristics *as such,* bringing production into service for a kind of vibration form and its differentiation that cannot be derived from a physical and physiological model alone. Thus, vowel perception and recognition are indicated as not being subordinated to sound production. They either intervene in or are superordinated to vowel sound production.

Obviously, if this holds true, the following interrelated questions arise: How do we address the vowel sound as a phenomenon of a produced form that cannot be derived from a physical model or physiological condition in all its very constitutive characteristics, and how do we link vowel perception and vowel production?

In the Preliminaries, we discussed these two questions as follows: "Prevailing theory is characterised by its explanation and description of vowel sounds within a physical model unspecific to speech: all kinds of sounds and noises are transformed by filters in the same way, irrespective of whether or not they concern utterances (speech events).

"One possible way to respond to the difficulties of understanding prevailing theory in terms of its intellectual re-enactment and to the fact that empirical findings can contradict its predictions might be to supplement the existing source–filter model or to replace that model with another physical model external to language and speech.

"Another approach might be to assume that the production and formation of vocal sounds is speech specific and, based on such a premise, to develop a method for describing vowel sounds in form-related terms. This second approach assumes that the vowel sound and its manifestations elude description within a purely physical model. […]

"Either resonances as such, and thus the corresponding pharyngeal, oral and nasal resonance patterns of the vocal tract, fail to represent in full the physical quantity to which language and speech directly refer, but another physical quantity can be found instead; if this is the case, then it is simply a matter of replacing the existing (physical) model with another (physical) model rather than adopting a fundamentally different perspective; […] or, aside of the human voice, no construction, no instrument and no process can be found to exist in physics that would explain and allow for the production of vowel sounds including basic variations of sound characteristics, for example, fundamental frequency and phonation type; then, the physical representation of human voice cannot be related to a simple voice-independent physical quantity, but instead, the voice would produce a 'substance' or 'quantity'." (pp. 83–84)

Conclusion

In conclusion, the very likely failure of the attempt to derive the *general* acoustic representation of vowel quality primarily from a physical and physiological model of source and filter, and possibly from any primarily physically and physiologically based model, and the finding that vowel-specific spectral characteristics relate to pitch (or to its alternative), provide grounds for arguing that vowel quality is a sound dimension in its own right, a sound quality in contrast to other aspects of sound commonly subsumed under the term timbre, a produced form of vocal expression with a unique aptitude for a linguistic function, an achievement of the human voice itself (Preliminaries, p. 6). In these terms, the very likely failure of the above attempt and the observed vowel–pitch relation (or its alternative) are understood here as indicating an "intervention" of perception and recognition in the production of vowel sounds.

These formulations represent but a rough and indicative sketch of reflections and speculations that have arisen during our investigation of the vowel sound. We are aware, of course, that it is uncommon to reflect on vowel sounds in these terms. However, it is not our priority to defend a thesis formulated in speculative terms only. As stated, even within the research team, the ideas exposed here were and still are controversially discussed. In the first instance, it is the aim of the present exposure to draw attention to our general interpretation of the various empirical findings presented in this treatise: These findings do not concern a minor problem of occurring spectral variation of vowel sounds that could be solved by additional differentiations of the prevailing acoustic theory. Any future acoustic theory of the vowel has to face the questions of why there is a vowel–pitch relation (or its alternative), why the vowel sound appears to be a kind of foreground–background phenomenon, and why the spectral representation of vowel quality is nonuniform. Any future theory has to provide a new theoretical basis for these general indications. In these terms, the empirical findings presented call for a paradigmatic change in understanding the vowel sound as a core phenomenon of the human voice.

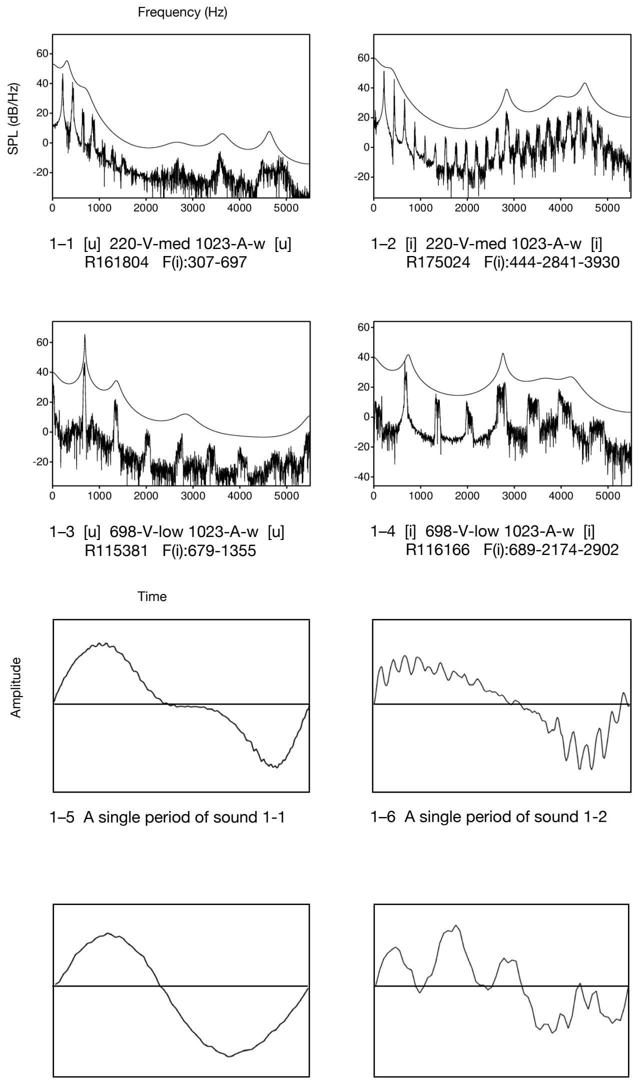
Figure 1. Sound-internal perceptual referencing to a repeating pattern and pattern-internal vibration form: General illustration for sounds of /u/ and /i/. Four sounds produced by a woman in V context at intended f_o of 220 Hz (sounds 1 and 2) and 698 Hz (sounds 3 and 4). Average spectra of the middle 0.3 sec. of the sounds are shown at the top of the figure, and single periods extracted from the middle of the sounds are shown at the bottom. (Note that a single period as shown here corresponds to what is discussed in this excursus as a repeating sound pattern.) If the two sounds of /u/ are compared with each other (sounds 1 and 3), due to the difference in the f_o (and pitch) levels, the spectral energy in the lower frequency range < 1.5 kHz and the vibration form of the period differ. The same is true for the comparison of the two sounds of /i/ (sounds 2 and 4). However, if the two sounds of /u/ and /i/ produced at the lower intended f_o (and pitch) level of 220 Hz are compared with each other, the same contrast of the vibration form of the periods is manifest as for the two sounds of /u/ and /i/ produced at the higher f_o (and pitch) level of c. 698 Hz: Sparse articulation of the undulation for sounds of /u/, very pronounced articulation of the undulation for sounds of /i/. This illustrates that the undulation contrast is independent of f_o (and pitch) if sounds of different vowels with similar f_o (and pitch) levels are compared with each other.
[C-E3-F01] ↗

Figure 2. Sound-internal perceptual referencing to a repeating pattern and pattern-internal vibration form: General illustration for sounds of /u/, /a/ and /i/. Six sounds produced by women in V context at intended f_o of 220 Hz and 698 Hz. For the sounds of /u/ and /i/, see Figure 1. Two sounds of /a/ are added to the sounds of the close vowels. Average spectra of the middle 0.3 sec. of the sounds and single periods extracted from the middle of the sounds are again shown. If the two sounds of /a/ are compared with each other (sounds 2 and 4), there is no pronounced difference concerning spectral maxima < 1.5 kHz and the range of prominent spectral energy that would relate to the difference in f_o (and pitch) levels, in contrast to the sounds of /u/ and /i/. This illustrates the non-uniform character of the spectral representation of vowel quality. (However, there is a pronounced difference for the periods of the two sounds of /a/ due to the difference in the number of harmonics in the spectrum.) If, for both f_o (and pitch) levels separately, the three sounds of /u/, /a/ and /i/ are compared with each other, comparable contrasts of the vibration form of the periods are manifest: Sparse articulation of the undulation for sounds of /u/, intermediate articulation of the undulation for sounds of /a/, very pronounced articulation of the undulation for sounds of /i/. Again, the undulation contrast is independent of f_o (and pitch), if sounds of different vowels are compared with each other, given similar f_o (and pitch) levels of the sounds.
[C-E3-F02] ↗

Figure 3. Sound-internal perceptual referencing to a repeating pattern and pattern-internal vibration form: Sounds of /u/ and /i/ produced with very low vocal effort illustrating the impact of the foreground–background character of the vowel sound on perceptual referencing. A sound produced by a man at an intended f_{o} of c. 349 Hz and intended as /u/, and a sound produced by a woman at an intended f_{o} of 440 Hz and intended as /i/. Average spectra of the middle 0.3 sec. of the sounds and single periods extracted from the middle of the sounds are shown. Visually, there is no difference between the two periods to identify. Obviously, this is, in the first instance, only a result of the type of graphic display. However, spectrally, the difference in the higher frequencies > 2 kHz is very small. Perceptually, the sound of /u/ may be perceived by some listeners as /i/ but not by others and, vice versa, the same holds true for the sound of /i/ possibly being recognised by some listeners as /u/ (see text). This illustrates the impact of the foreground–background character of the vowel sound on perceptual referencing.
[C-E3-F03] ↗

Figure 4. Sound-internal perceptual referencing to a repeating pattern and pattern-internal vibration form: A natural sound of /i/ and HP-filtered variants thereof again illustrating the impact of the foreground–background character of the vowel sound on perceptual referencing. Extract of Chapter M8.2, Table 1 (see sound 4 of /i/ in this table). A natural reference sound of /i/ produced by a man at an intended f_{o} of 220 Hz and four HP-filtered variants thereof with CFs of 440–660–990–1320 Hz are shown. HP filtering this natural sound of a close vowel with stepwise increasing CFs resulted in an initial vowel quality shift in a close–open direction, which subsequently reversed: The filtered variants were recognised as /e e e e ei/ – /ɛ ɛ ɛ ɛ i/ – /e ei i i i/ – /e e i i i/ (labelling details of the five standard listeners, given in a phonetic order). This again illustrates the impact of the foreground–background character of the vowel sound on perceptual referencing.
[C-E3-F04] ↗

Figure 1. Sound-internal perceptual referencing to a repeating pattern and pattern-internal vibration form: General illustration for sounds of /u/ and /i/. [C-E3-F01]



1–1 [u] 220-V-med 1023-A-w [u]
R161804 F(i):307-697

1–2 [i] 220-V-med 1023-A-w [i]
R175024 F(i):444-2841-3930

1–3 [u] 698-V-low 1023-A-w [u]
R115381 F(i):679-1355

1–4 [i] 698-V-low 1023-A-w [i]
R116166 F(i):689-2174-2902

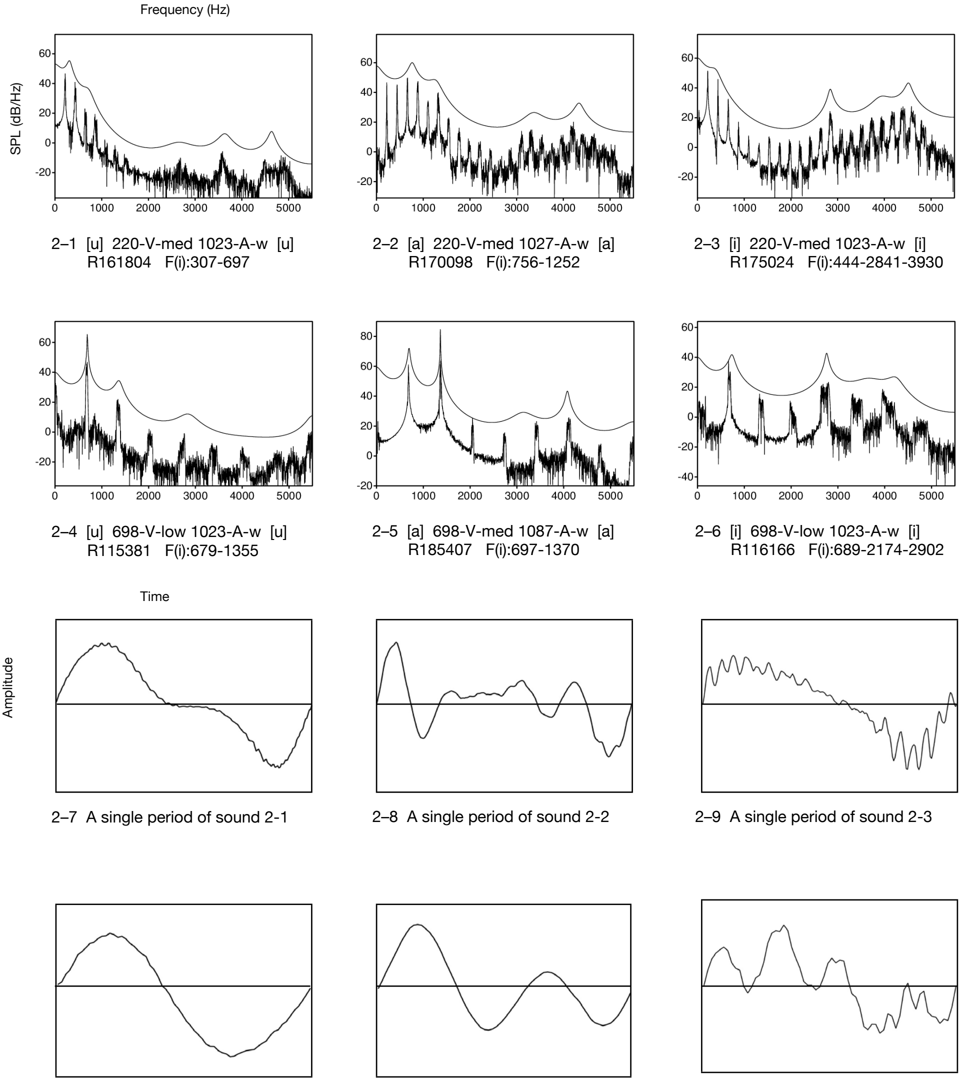
1–5 A single period of sound 1-1

1–6 A single period of sound 1-2

1–7 A single period of sound 1-3

1–8 A single period of sound 1-4

Figure 2. Sound-internal perceptual referencing to a repeating pattern and pattern-internal vibration form: General illustration for sounds of /u/, /a/ and /i/. [C-E3-F02]



2–1 [u] 220-V-med 1023-A-w [u]
R161804 F(i):307-697

2–2 [a] 220-V-med 1027-A-w [a]
R170098 F(i):756-1252

2–3 [i] 220-V-med 1023-A-w [i]
R175024 F(i):444-2841-3930

2–4 [u] 698-V-low 1023-A-w [u]
R115381 F(i):679-1355

2–5 [a] 698-V-med 1087-A-w [a]
R185407 F(i):697-1370

2–6 [i] 698-V-low 1023-A-w [i]
R116166 F(i):689-2174-2902

2–7 A single period of sound 2-1

2–8 A single period of sound 2-2

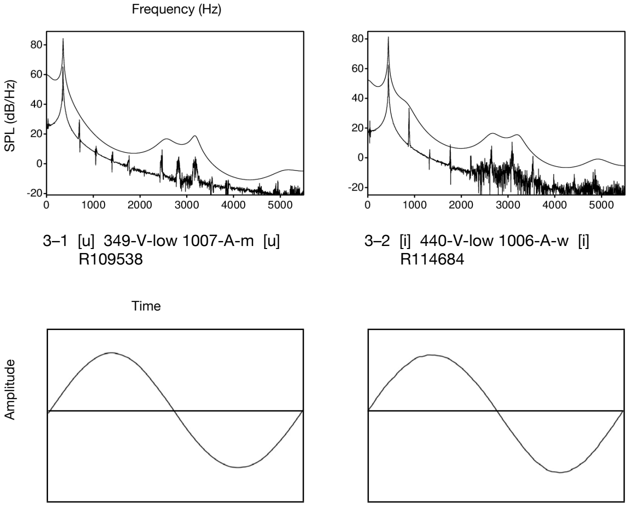
2–9 A single period of sound 2-3

2–10 A single period of sound 2-4

2–11 A single period of sound 2-5

2–12 A single period of sound 2-6

Figure 3. Sound-internal perceptual referencing to a repeating pattern and pattern-internal vibration form: Sounds of /u/ and /i/ produced with very low vocal effort illustrating the impact of the foreground–background character of the vowel sound on perceptual referencing. [C-E3-F03]



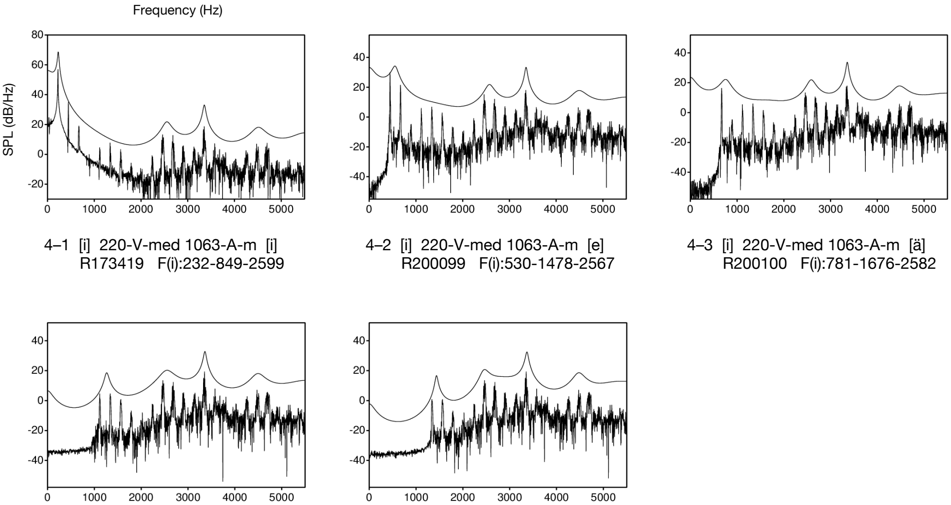
3–1 [u] 349-V-low 1007-A-m [u]
R109538

3–2 [i] 440-V-low 1006-A-w [i]
R114684

3–3 A single period of sound 3-1

3–4 A single period of sound 3-2

Figure 4. Sound-internal perceptual referencing to a repeating pattern and pattern-internal vibration form: A natural sound of /i/ and HP-filtered variants thereof again illustrating the impact of the foreground–background character of the vowel sound on perceptual referencing. [C-E3-F04]



4–1 [i] 220-V-med 1063-A-m [i]
R173419 F(i):232-849-2599

4–2 [i] 220-V-med 1063-A-m [e]
R200099 F(i):530-1478-2567

4–3 [i] 220-V-med 1063-A-m [ä]
R200100 F(i):781-1676-2582

4–4 [i] 220-V-med 1063-A-m [i]
R200101 F(i):1290-2368-2616

4–5 [i] 220-V-med 1063-A-m [i]
R200102 F(i):1439-2503-2875

11 Towards a Future Acoustic Theory of the Vowel

11.1 Prerequisites – Perspective, Empirical References and Phenomenological Indices

An acoustic theory of the vowel addresses sounds that are produced and perceived as belonging to one vowel quality of a given language in contrast to sounds of other vowel qualities of that language in the sense of quasi-ideal parallelism of vowel intention and vowel recognition for speakers and listeners.

On the basis of the documentation and experimentation presented in this treatise, we conclude that our current knowledge allows for the formulation of three general prerequisites for the creation of a future acoustic theory of the vowel: Adopting a specific initial perspective of investigation, setting empirical references and facing phenomenological indications.

As stated in the Introduction, we propose to initially adopt a perspective of parallelism of vowel recognition and acoustic sound characteristics akin to a psychophysical perspective, with the specificity that recognition may precede acoustics in the investigation process. (This proposition may seem trivial in that the perspective mentioned is understood as already adopted with the prevailing theory. However, this is not the case. According to our understanding, the prevailing theory does not face phenomenological indications and then derive vowel-related acoustic characteristics in parallel, but refers primarily to the source–filter model of speech production.)

We also propose that, for the building of a future theory of the acoustic representation of vowel quality *in general,* two conditions should initially be adhered to: Firstly, to differentate and hierarchise diffferent types of vowel sounds – e.g. sounds with quasi-static spectral characteristics versus sounds with a marked variation of spectral characteristics, sounds produced in isolation versus sounds produced in a consonantal context, sounds in nonsense context versus sounds in a specific semantic context and so forth – and, secondly, to start with an investigation of monophthongs produced in isolation (V context) with quasi-static spectral characteristics (or comparable sound fragments extracted from a sound context). For an initial investigation, further restrictions regarding the sounds examined may be added to strengthen the condition of quasi-static sounds (that is, excluding on- and offsets of the sounds), avoid effects of sound timbre by nasalisation (that is, excluding sounds with nasalisation all together) and avoid a direct impact of sound duration on vowel production and recognition (that is, investigating only sounds of the long vowels of a language). A new theory should first and foremost address a successful prediction of the acoustic representation of vowel quality for sounds with these restricted characteristics. Other sounds should subsequently be investigated. In the course of theorising, the appropriateness of such an initial hierarchisation and specific focus of investigation may become a matter of debate.

As indicated, the creation of a new acoustic theory of the vowel is expected here to have to superordinate sound perception and recognition over sound production initially. In the first instance, the theory should aim to predict the vowel-specific acoustic characteristic of any sound produced as a monophthong of the type mentioned in the above paragraph, independently of the production parameters of the sound, but given an unambiguous vowel quality recognition. Only in the second instance may production-specific, additional characteristics be assessed. Again, in the course of theorising, the appropriateness of such an initial superordination of sound perception and recognition over sound production may become a matter of debate.

Obviously, with regard to this proposition, assumptions are involved. Above all, it is assumed that isolated vowel sounds with quasi-static spectral characteristics are principally intelligible and that this fact is central to human voice and speech: From a structural point of view, vowel sounds must be intelligible as such, per se, quasi-independently of their context, because their character as elements of an autonomous system – manifest in the arbitrary relation between signifying elements and the signified and in the aptitude of speech for a phonetic system of writing – is at the core of speech and language (see also the Preliminaries, p. 6). And because vowel sounds are transmitted acoustically, vowel qualities are represented by a specific acoustic characteristic, invariant among different sounds of a vowel – or, more precisely, the differences between the vowel qualities of any given language are represented by differences (contrasts) of a specific acoustic characteristic.

We are aware that many scholars are fundamentally critical of basing an acoustic theory of the vowel on isolated steady-state vowel sounds and question the recognisability and linguistic function of such sounds. This critical stance generally refers to a linguistic definition of the vowel as syllabic and not existing, per se, independently of a semantic function: “Vowels are defined by the physiological characteristic of their having no obstruction in the vocal tract, and by their function within a phonologically defined syllable." (Ladefoged and Maddieson, 1996, p. 282) Moreover, according to prevailing theory and many related empirical studies, no single vowel quality-related acoustic characteristic as an invariant characteristic for all sounds of a vowel is expected if sounds produced with different production parameters (above all, phonation type, age and gender of the speakers, vocal effort, duration) are included in the investigation. Furthermore, some scholars assume that vowel-inherent spectral change should be considered a principal aspect for the assessment of vowel-specific acoustic characteristics. Finally, in discussions during our research, some scholars rejected the idea that vocal expression and speech can be traced structurally to perceptual and acoustic sound elements, such as vowels, quasi-independently of the production and context of the sounds. In their view, there is no structural basis for vowel sounds being related to a particular vowel-specific acoustic characteristic. Rather, the brain is able to develop multiple strategies for recognising vowels, depending on production, context and the particular acoustic characteristics associated with them.

However, as detailed in the Preliminaries, the present line of argument does not concur with the notion of a fundamental opposition between isolated versus context-bound sounds, static versus dynamic spectral processes and "functionless" utterances versus those with a linguistic function. Much could be said in response to such oppositions and the critical take on the attempt to assess an invariant characteristic as representing recognised vowel quality for isolated, monotonous vowel sounds (for details, see the Preliminaries, pp. 90–91). Here, however, this debate is left to a future discussion in the process of future theory building, with the exception of the following remark: Whatever the manifold, sometimes surprising, unexpected and extraordinary phenomena of the perception and recognition of vowel sounds as a result of the general capacity of human perception, of the development of learned strategies in ontogeny and of specific voice and listening training may be, there is no proof that these phenomena are opposed to a basic structural reference of vowel quality recognition and a related acoustic characteristic. Indeed, a structural reference may prove to be indispensable for the very complex processes of the perception and recognition of vocal expressions.

Concerning empirical sound samples as references, in relation to the perspective proposed, large-scale, language-specific databases of sounds of long vowels are needed, and they should be created and published in an open-access form to define empirical references based on which any thesis of the acoustic representation of vowel quality is verifiable or falsifiable. These databases should include an extensive variation of basic production parameters and a (useful) redundancy of investigated sound characteristics. Most importantly, the f_o range should cover the entire range of recognisable natural vowel sounds. (With this perspective and intent, the Zurich Corpus was created, focusing mainly on natural sounds of the long Standard German vowels and additionally presenting numerous speech extracts and manipulated and resynthesised and synthesised sounds.) In parallel, large-scale databases of everyday speech extracts and utterances from the performing arts field are also needed in order to demonstrate the significance of production parameter variation for vowel sounds, above all in the context of speech, but also for singing. Concerning vowel sound manipulation and vowel (re-)synthesis, large-scale sound compilations as sound references with a systematic structure will have to be defined in the process of theory building.

Concerning the phenomenological indications, according to the main conclusion of the Preliminaries and this treatise, a future theory has to address the three primary indices of the vowel–pitch relation (or its alternative) for both vowel recognition and the acoustic representation of vowel quality, the perceptual and acoustic foreground–background character of the vowel sound and the nonuniform spectral representation of vowel quality, including the associated aspects detailed above. At the same time, theory building has to evaluate the significance of the secondary indications summarised above.

11.2 Theory Building – Method of Acoustic Analysis, Thesis, Verification–Falsification Criterion

Theory building addressing the question of the *general* vowel quality-related acoustic characteristics faces challenges, especially with regard to the method of acoustic analysis and the formulation of theses.

A uniform, fully objective and “reversible” method of analysing the acoustic characteristic of vowel quality has to be developed: Uniform in the sense that the analysis procedure applies to all recognisable vowel sounds independent of the parameters of their production; fully objective in the sense that, in the course of analysis, there is no intervention needed on the part of the investigator; “reversible” in the sense that, based on the results of acoustic analysis, the analysed sound can be reproduced (resynthesised) with only minor and predictable changes in the sound timbre and with no change in the recognised vowel quality. (Note in this context that creating new software tools for acoustic analysis and high-quality resynthesis and synthesis of vowel sounds is pivotal for the theoretical endeavours undertaken here, including open access to these tools.)

Based on such a method of acoustic analysis and a new phenomenological investigation of vowel sounds, the main focus lies in determining a particular kind of acoustic sound characteristic that represents vowel quality *in general.* In the course of the investigation, intermediate theses and predictions for sounds with limited production parameters and limited sound variations may be useful (see e.g. the second excursus on vowel quality and the harmonic spectrum).

Most importantly, for any hypothesis, a criterion for either verification or falsification has to be formulated. (Note that no verification or falsification criterion is given in the literature as a reference with regard to the two theses of either formant patterns or spectral shapes being vowel-specific.)

Once such a framework for investigating vowel acoustics is established, theory building must proceed in the search for a general understanding of the voice and the perception and production of sounds with vowel qualities and attempt to explain the acoustic characteristic that represents vowel quality.

Afterword

A thread to seize and trace

To say that the formant pattern is the primary acoustic characteristic of vowel quality *in general* cannot be comprehended despite the many repetitions of this statement. The same holds true for the thesis that, if it is not the formant pattern, it is the spectral shape, the course of the spectral envelope.

The first resistance emerges from mere observation. Given the ability to estimate the levels and variation range of pitch as an aspect of intonation, any attentive listening to speech in various contexts of everyday life leads to the assessment that the upper limit of the pitch range of recognisable speech is 800 Hz at a minimum: As a rough estimate, many women can speak in an intelligible way up to approximately 800 Hz with register changes that are difficult to allocate to frequency ranges in general. Many men can speak up to approximately 400 Hz in the mixed register. In the falsetto register, their pitch often rises to 700 Hz and, in some cases, even above that level, comparable to women's speech. Children's speech may be associated with the middle and higher range of pitch levels of women. Exceptions of speech at higher levels may also occur.

When it comes to pitch variation in speech, expressions observed within the field of the performing arts deserve the greatest attention: They bear witness to the wide pitch range of recognisable speech in a way and with compelling evidence that only artistic virtuosity, stylistic variation and expressiveness can provide. And what holds true for speech also holds true for singing.

However, if speech is recognisable up to these pitch levels, vowels as isolated sounds have to be expected to be distinguishable, too. Futher, what is heard and recognised as the pitch of natural speech generally corresponds to the fundamental frequency as an acoustic correlate. Thus, the upper limit of f_o of recognisable voiced speech sounds is 800 Hz at a minimum.

This observational assessment calls for verification, which is given in the first and second chapters of this treatise in terms of a broad phenomenological basis of vowel sounds provided in the Zurich Corpus and, as extracts of the corpus, vocalisations of single speakers with extended f_o variation, recognisable vowel sounds produced at high f_o levels, minimal pairs produced with extended f_o variation and speech extracts of everyday life and the field of the performing arts, in their turn including middle and high f_o levels as an aspect of intonation. References to earlier studies on the matter are also given in the Materials.

The observation and demonstration that not only speech but also single vowel sounds as monophthongs are recognisable up to f_o of 800 Hz at a minimum challenge both theses of the formant pattern or the spectral shape as acoustically representing vowel quality *in general:* Firstly, the f_o of recognisable vowel sounds can far surpass the upper f_o limit of formant measurement procedures and their methodological substantiation. How, then, to verify the thesis of the formant pattern or the spectral shape as acoustically representing vowel quality *in general?* Secondly, the f_o of recognisable vowel sounds can far surpass statistical F_1 commonly given for sounds of all close vowels, and it can even surpass statistical F_1 for sounds of close-mid vowels and approximate or equal F_1 of sounds of open-mid or open vowels. In these terms, the range of f_o of recognisable vowel sounds covers almost the entire range of statistical F_1 of sounds of all vowels produced by children, women and men. Likewise, the f_o of recognisable sounds of a vowel can in part or entirely surpass prominent spectral energy that is manifest for sounds of that vowel produced at low f_o levels. How, then, can it be claimed that F_1 *as such* (or prominent lower spectral energy *as such*) and, with it, the vowel-related F-pattern *as such* is vowel quality-specific?

In these terms, the first resistance emerging from mere observation turns into the thesis that any phenomenological investigation of the acoustic characteristics of vowel sounds and vowel quality will reveal systematic deviations of these characteristics from predictions of the formant or the spectral shape theses. The line of subsequent investigation is then partly predetermined: The deviating manifestations have to be ascertained and documented, and an attempt has to be made to understand their causes. Figuratively speaking, the systematic deviation is a thread to seize and trace, conceptually and experimentally.

As our contribution, core aspects of these deviations were investigated and are documented and discussed in this treatise, that is, aspects that are a consequence of or parallel the appraisal of the actual frequency range of f_o and of the lower vowel spectrum being related to f_o to be observed for recognisable natural vowel sounds. For natural vowel sounds and also for their resynthesis, these aspects are (i) formant pattern and spectral shape ambiguity, (ii) possible decrease, disappearance and inversion of supposed age- and gender-related lower spectral characteristics of vowel sounds, (iii) possible disappearance of supposed phonation-related spectral characteristics of vowel sounds, (iv) variation of supposed vowel-related spectral characteristics due to variation of vocal effort, (v) different spectral peak numbers for sounds of a given vowel, (vi) inverted spectral minima and maxima for sounds of a given vowel, and (vii) sounds with flat vowel-related spectra or spectral parts. For manipulated natural or synthesised vowel sounds, in addition, recognisable sounds with only incomplete harmonic series (including sounds with equal amplitudes of the harmonics), with very few harmonics as well as with very few partials lacking a harmonic relation were also investigated and are documented and discussed, not allowing for spectral shape estimation and also often not allowing for *F*-pattern estimation.

In the course of the phenomenological investigation of these core aspects – and as a direct consequence of the comparison of voiced and whispered vowel sounds and their spectral characteristics – it became necessary to keep pitch apart from fundamental frequency for the investigation of the vowel sound. Following this imperative and conducting experiments in which pitch and fundamental frequency were contrasted, it could be demonstrated that it is not fundamental frequency but pitch (or a comparable perceptual reference) to which vowel recognition and the corresponding spectral sound characteristics relate.

Furthermore, in the course of the phenomenological investigation, the observed character of the relation between recognised vowel quality and spectral sound characteristics strongly supported earlier indications that this relation is nonuniform. In order to provide corresponding evidence, it is demonstrated here that spectral characteristics of vowel sounds depend on vowel qualities, levels and ranges of pitch and pitch variation (and of f_o, if quasi-equal to pitch), and also the individual spectral energy distribution of the sounds, and while the lower part of the vowel spectrum is strongly affected by pitch variation, this does not hold true for the higher part of the spectrum. We have attributed this last phenomenon to the dichotomy of the vowel spectrum (see the Preliminaries, p. 67 and p. 238).

In consequence, the vowel–pitch relation (or its alternative) and the nonuniform spectral representation of vowel quality, including the dichotomy of the vowel spectrum, turn out to be invariant in the many diverse spectral manifestations of sounds of a vowel: They are predictable.

In the further course of the investigation, referring to earlier studies reported in the literature and questioning the relation between lower and higher parts of a given sound spectrum and their representation of vowel quality, LP and HP sound filtering became a matter of attention. Indeed, as earlier studies have indicated, LP or HP filtering of vowel sounds does not generally impair or corrupt vowel quality recognition but causes vowel quality shifts, which, in their turn, are predictable, supporting the foreground–background thesis, as is shown in this treatise.

Thus, in the course of exploring and describing the systematic deviations of spectral characteristics from predictions of the formant or the spectral shape theses, three statements have emerged that predict the parallelism between vowel quality recognition and related spectral characteristics of the vowel sound: Vowel quality recognition (and with it, spectral sound characteristics) relates to pitch (or a comparable perceptual reference), the vowel sound is a kind of perceptual and acoustic foreground–background phenomenon, and spectral representation of vowel quality is nonuniform. (We prefer this order of the statements.) These three findings apply to the vowel sound *in general.* They are what can be said about all recognisable vowel sounds and their acoustic representation of vowel quality. Because of this, they represent a turning point in the quest for the acoustic representation of vowel quality: They explain the causes for the above deviations from predictions of the prevailing theory, and they provide three primary indices for a future theory. In these terms, they transform criticising the prevailing acoustic theory of the vowel into providing indices for a new theory.

Finally, a supplementary study addressed the detectability of resonances of sound production based on the analysis of the radiated sound. As could be expected against the background of the general methodological limitation of formant and spectral shape estimation, earlier studies on the parallelism of articulator positions and spectral characteristics of vowel sounds and the many arguments and experiments conducted in the context of this treatise, it is again demonstrated that resonances of sound production cannot always be detected in the acoustic analysis of a vowel sound.

After having observed the actual pitch range of recognisable speech and realising that, for itself, this observation contradicts the formant pattern and spectral shape theses, this far we could trace the thread of argument, investigation and understanding on our part. Obviously, there will be a need for replication and extension of the experiments presented here, including larger sound samples and/or larger numbers of speakers and/or extended parameters of sound production and/or other software tools and/or larger numbers of listeners and so forth. Also, additional aspects of the matter not presented in this treatise may be demonstrated. However, we conclude that the present description of the systematic deviations of spectral characteristics from predictions of the formant pattern and the spectral shape theses, and the present report and elaboration of three primary indices concerning the acoustic representation of vowel quality *in general* – including, above all, the discovery of the vowel–pitch relation (or its alternative) – suffice to approach the question of vowel acoustics in a new way, beyond the theoretical framework of formant patterns or spectral shapes. The description and the indices provided here suffice to argue for a new attempt at a theory of vowel acoustics. At the same time, as concluded in the last main chapter, the next necessary steps are predefined: The development of a uniform, objective and "reversible" method of analysing the acoustic characteristics of vowel sounds combined with the development of new software tools for high-quality sound resynthesis or synthesis, phenomenological investigation of vowel-related acoustic characteristics, and formulation of hypotheses that attempt to predict these characteristics *in general.* We take the stance that substantial progress in understanding vowel acoustics must be based on such a new approach.

How does it come about?

How is it that, up to now, almost all textbooks introducing phonetics still assert that formants are the *primary* acoustic cue of vowel quality recognition? How is it that many specialist articles still claim that either formant patterns or spectral shapes acoustically represent vowel quality *in general* (see citations in Chapter 10.1)?

Obviously, as a result of this, the present treatise – or any treatise of the same kind – will meet with scepticism. Furthermore, the plausibility of voice production as a process of phonation and articulation (of source and filter), the character of the prevailing theory to interrelate production, acoustics and perception of vowel sounds, the results of statistical formant pattern measurements and the ability to produce vowel sounds using a formant-based synthesiser seem to give strong reasons for such scepticism. But in our view, these seemingly convincing arguments obscure the facts that (i) there has never been a methodological substantiation of acoustic analysis that would allow for verification of hypotheses and related predictions of the acoustic characteristics of vowel sounds *in general,* (ii) there has never been empirical proof that the prevailing theory of vowel-specific formants or spectral shapes can be applied to vowel sounds *in general,* even for sounds for which methodological substantiation of acoustic analysis is not an issue, (iii) it is not possible to translate the many diverse experiments and results reported in the specialist literature into any stable prediction of the acoustic cue(s) of vowel quality, and (iv) ever since early statistical studies of analysing spectra and spectrograms of vowel sounds, scepticism against the formant theory (and subsequently also against the spectral shape theory) was continuously expressed by specialists in the field of phonetics. In these terms, the prevailing theory should, in its turn and already at an earlier point, have been met with much stronger scepticism, and this scepticism should have been communicated in an explicit manner to scholars of other disciplines and students of phonetics.

However, here, we dispense with further details on the matter. The question of why the formant pattern and spectral shape theses have persisted as the prevailing theses of vowel acoustics to this day may become the subject of a historical examination in the future. Only one remark shall be repeated: There has been a serious lack of large-scale references for voices and vowel sounds with extensive variation of production parameters that were compiled and edited for scientific use. If we had had comprehensive vowel sound corpora at our disposal at an earlier stage, including sounds related to various speaking styles and habits in everyday life and also, most importantly, to various speaking and singing styles in the field of the performing arts, then, simplified concepts of “normal/average” versus “high” fundamental frequencies of speech, speaking versus singing, age- and gender-related sound characteristics, characteristics of “normal” speech in contrast to “emotional” expressions and so forth would barely have dominated the understanding of vowel acoustics.

Some shortcomings

This treatise has some shortcomings regarding its form. Firstly, it was written over a long period and, in consequence, textual consistency is not always fully achieved. Secondly, the text, figures and tables are very extensive and, therefore, the entire presentation is subdivided into two parallel parts: A main body that is reduced to the general line of investigation and argument and also to exemplary illustrations of the general findings only, and a Materials section in which the same content is presented but with extended background information and main references, details of the experiments conducted and their results, extended discussions, full documentation of experimental results and links to the entire samples of the sounds examined. This parallelism of presentation entails numerous repetitions of text parts given in both the main body and the Materials, and if readers read both parts, they are confronted with corresponding text redundancy.

However, these aspects should be considered with regard to (i) the phenomenological perspective adopted, (ii) the related effort of creating and investigating the comprehensive and systematic Zurich Corpus, the main references of this treatise being the sounds of this corpus, (iii) the results of exploring the deviations of vowel-related acoustic characteristics from predictions of the formant pattern or the spectral shape theses and, above all, (iv) the elaboration of three primary indices for a future acoustic theory of the vowel. If the reported findings and the elaboration of the primary indices are presented in a comprehensible manner and in a form that allows for an experimental replication with limited effort – the sound samples investigated published open access and linked to software tools for analysis and (re-)synthesis purposes in order to provide full traceability – then the shortcomings are not of major importance.

A limitation

This treatise also has a limitation regarding the generalisation of the reported vowel and pitch recognition results. As discussed in the Introduction (see p. 22–23), for most of the experimental studies, five expert listeners that participated in the standard vowel recognition test conducted when creating the Zurich Corpus (expert listener panel) also performed the vowel and pitch recognition tests. Thus, the question is posed as to whether the recognition results obtained on this basis allow for generalisation, and future research is needed to address this question. However, this limitation has to be appraised with respect to the considerations set out in the Introduction. In short and in summary, we would not have been able to create the Zurich Corpus, explore the perceptual and acoustic characteristics of vowel sounds concerning questions for which stable references in the literature are lacking, develop and evaluate experimental designs and software tools, then conduct the many experimental studies and, in their course, discover the vowel–pitch relation (or its alternative) – and demonstrate the consistency of the individual experimental results explained by that discovery – without having involved and interacted with expert listeners and having limited their number. Furthermore, almost all sounds investigated are made accessible in combination with four software tools. This allows direct evaluation by researchers, not only to support verification and replication and to limit the corresponding effort, but also to support the conceptual work of future experiments.

A note on the adopted terms for vowel quality categorisation

As indicated in the Introduction, vowel qualities are specified in this treatise according to close–open, front–back and unrounded–rounded differences. This categorisation and terminology, based on articulatory phonetics, is common in the literature. For readability, it was therefore adopted in this treatise.

Obviously, in a demonstration of the extensive, pitch-related and non-uniform variation of spectral characteristics that can be observed for sounds of a given vowel and that explain the occurring ambiguity of formant patterns and spectral shapes, the use of the above terms without corresponding reflection leaves room for criticism: Above all, the traditional, direct association of the terms with vowel quality-specific resonance patterns as a result of vowel quality-specific positions of the articulators poses a basic problem for the present context. However, we were unable to approach this terminological matter because of the sheer workload and effort involved in providing evidence for the three primary indices put forward here and embedding them into the general context of vowel acoustics, including the extensive documentation of sound samples compiled for verification.

Production, acoustics, perception

As said, the prevailing acoustic theory assumes that it provides a consistent explanation for vowel-specific resonances due to sound production, for formant patterns or spectral shapes that reflect these resonances in the acoustic characteristics of the radiated sounds and for these patterns or shapes being the primary cue for vowel quality recognition. Any theory of the vowel sound has to strive for such a consistent interrelation of sound production, acoustic characteristics and sound quality recognition.

In this treatise, however, no such attempts were made to interrelate production, acoustics and perception in general (apart from particular aspects such as varying sound production parameters in order to create variations in the sound spectrum, comparing resonances of sound synthesis with formant estimation of the radiated sounds, testing vowel quality recognition, and relating pitch as a perceptual characteristic to the vowel spectrum as a characteristic of acoustic measurement). It was indeed not the aim to address their *general* mutual relations. These relations have to be clarified within the framework of a future theory.

Perceptual referencing in the process of vowel production and recognition

The demonstration that there is no average spectral shape that represents vowel quality acoustically, even for sounds of single speakers, leads us to the conclusion that the perceptual referencing does not relate directly (in an unmediated way) to a resonance characteristic external to an actual sound, and the demonstration of the vowel–pitch relation (or its alternative) is ground for the thesis that the perceptual referencing is an operation that is related to “sound-internal” acoustic characteristics. As indicated in the third excursus, one way to understand a perceptual referencing to “sound-internal” acoustic characteristics is to consider the “nature” of the vowel sound as a perceptual relation to a sound pattern repetition and form differentiation of the repeating pattern, with this form differentiation possibly being incomparable to any other kind of sound and its production. Obviously, then, the question arises about the interrelation of the two processes of production and perception (or vice versa) of vowel sounds.

Vowel spectrum

In the context of the perceptual referencing mentioned, the fact that the commonly taken acoustic perspective on vowel sounds is a spectral perspective becomes worth considering. Thereby, in our view and repeating earlier considerations, the main argument for questioning the appropriateness of a spectral characterisation of vowel quality *in general* concerns the indication that any attempt to establish vowel quality-related spectral templates will fail due to both methodological conditions and limitations of spectral analysis on the one hand (see the excursus on vowel quality and the harmonic spectrum) and, on the other hand, the experimental finding of the nonuniform spectral variation of sounds produced with varying production parameters.

Concerning spectral analysis, firstly, the harmonic spectra of two natural voiced vowel sounds produced at different f_o levels are not directly comparable because of the different numbers and frequencies of the harmonics. Secondly, the harmonic envelopes of two natural voiced vowel sounds produced at very different f_o levels are, in their turn, often incomparable because the methodological substantiation for spectral envelope estimation for sounds produced at middle and high f_o levels is lacking. The same holds true for spectral peak patterns and *F*-patterns. Thirdly, as a consequence, the methodological substantiation for a spectral comparison of natural voiced vowel sounds with extensive f_o variation and natural whispered vowel sounds is lacking, too. These conditions and limitations regarding spectral analysis lead to the conclusion that, unrelated to the f_o of the sounds, no vowel-specific templates of harmonic spectra or spectral envelopes or spectral peak patterns or *F*-patterns can be developed due to methodological reasons.

It may be tempting to relate spectral characteristics to f_o. However, as said with regard to the estimation of spectral envelopes, spectral peak patterns and *F*-patterns, the lack of methodological substantiation for the estimation still stands against a *general* attempt of this kind. Concerning the harmonic spectrum, as we have discussed in the second excursus, the attempt to create f_o- and vowel-related templates also fails because of the foreground–background character and the extent of spectral variation observable for sounds of a vowel: Above all, for sounds of two adjacent vowels produced at equal f_o, it is not possible to create two templates of harmonic spectra in reference to which all sounds of the first vowel would always result in a smaller spectral distance to their respective template when compared with the spectral distance to the template of the sounds of the second vowel, and vice versa. Thus, even if related to the f_o of the sounds, no vowel-specific spectral templates can be created because of methodological and empirical reasons.

As far as the experimental results are concerned, in addition to the documented nonuniform manifestations of vowel-related spectral characteristics and their variation for natural vowel sounds, the various presented findings concerning sound manipulation or resynthesis and the presented findings concerning synthesised sounds with spectral characteristics alien to natural sounds in their turn run counter to the attempt to create vowel-related spectral templates.

Finally, the vowel–pitch relation (or its alternative) – including the finding that, most importantly, the relation between vowel quality recognition and pitch is not always substitutable by a relation of vowel quality recognition and measured f_o – underpins these counterarguments: Pitch is not an acoustic but a perceptual characteristic, and it is not always reflected in a measurable characteristic as a result of acoustic analysis.

Thus, further theorising must question the appropriateness of a spectral characterisation of vowel quality *in general,* although an alternative perspective is yet lacking. However, as indicated in the third excursus, we speculate upon a system of vibration form differentiation that is related to the temporal extension of a perceptually referenced repeating period (perceived and possibly recognised as pitch) and is, as a differentiation system, transferable to another temporal extension of a perceptually referenced repeating period. If this holds true, a corresponding method of acoustic analysis and description has to be created that, most importantly, allows for comparing sounds with different pitch (and f_o) levels.

Significance

In conclusion, the ending of the main text of the Preliminaries (p. 93) shall be resumed and taken further here: Our vocal cords – when modulating air expelled from the lungs – produce sound. The resonances of the pharyngeal, oral and nasal cavities could form the initial characteristics of the source sound into a formant pattern or a spectral shape that always and uniquely represents a vowel physically and thus allows us to perceive it accordingly. Empirical investigation reveals, however, that the spectral characteristics of vowel sounds systematically deviate from this.

This observation and experience call attention to the vowel sound as being insufficiently explained by a general physical model of sound production and sound transformation via vowel-specific resonances or by a human-specific physiological model of airflow, firstly modified by the vibration of the vocal folds or by a partial closure of them, and secondly modified by vowel-specific resonances of the vocal tract. In short and in these terms, the observation and experience call attention to the vowel sound as insufficiently explained by physics or physiology.

All this allows for the thesis that humans do not speak by simply creating sound contrasts according to given physical (acoustic) characteristics that are, per se, external to perception and recognition (with perception and recognition playing only a minor role in producing and identifying language-specific contrasts). Humans do not produce speech sounds in the same way they produce other sounds using objects or instruments. If this holds true, the observation and experience of the spectral characteristics of vowel sounds systematically deviating from the predictions of the prevailing source–filter and phonation–articulation models are of primary anthropological significance. This should be communicated accordingly.

Deepening the knowledge of the deviating spectral characteristics from predictions of the prevailing theory, interrelating the findings of individual studies on the matter with each other, further investigating vowel and pitch recognition and including sounds for which fundamental frequency and pitch are not interchangeable, a structure appears in the spectral variation for vowel sounds: Vowel quality recognition – and with it, spectral sound characteristics – relate to pitch (or to a comparable perceptual reference to a sound pattern repetition over time), the vowel sound is a kind of perceptual and acoustic foreground–background phenomenon, and spectral representation of vowel quality is nonuniform. The Preliminaries have shown how spectral characteristics commonly assumed to represent vowel quality deviate from theoretical prediction. In this treatise, the three primary indices presented now explain said deviations, at least for a substantial part. At the same time, they provide knowledge-based statements predicting vowel sound characteristics. Thus, this second treatise transcends the critical reflection on and the investigation of the formant and spectral shape theses into indices for a future acoustic theory of the vowel.

In this sense, we have traversed the preliminary stages of understanding the acoustic representation of the vowel. We did have knowledge of spectral manifestations of vowel sounds produced under certain conditions, but now, we have some understanding of the spectral manifestation of the vowel sound *in general.* This knowledge, indicating the actual acoustic characteristic representing vowel quality, gives reason for a change in paradigm, a new theoretical and empirical quest to understand the vowel sound and, with it, the human voice.

Materials

The Materials section presents background information and details of the method, results and discussions for each documentation and experimentation presented in Part II of the main text, including main references, detailed tables and sound series.

The chapters in this section correspond to the chapters of Part II, Chapters 2 to 9, and are numbered accordingly, starting with the letter “M”.

No further details are given for Chapter 1 in the Materials. They are presented in the Handbook of the Zurich Corpus.

M2 Natural Vowel Sounds, Vowel Spectrum and f_o

M2.1 Vocalises

Introduction

In this first chapter on natural vowel sounds, vowel spectrum and f_o, we address the question of the spectral characteristics of recognisable vowel sounds produced by single speakers with extensive variation of f_o (vocalises). The two subsequent chapters are concerned with investigating the upper-frequency limit of vowel recognition, and the fourth chapter addresses the question of f_o contours of intelligible speech, demonstrating the actual frequency extension of f_o contours observed in everyday speech and in the performing arts.

In the literature, vowel recognition and spectral characteristics of natural vowel sounds produced in isolation or syllable context with extensive f_o variation are discussed mainly in the context of singing (for an overview and discussion of studies related to f_o variation in singing, see Sundberg, 2012; Maurer et al., 2014; Friedrichs et al., 2017). Apart from this, f_o variation for these kinds of vowel sounds is also investigated in the context of vowel synthesis (for an overview, see Maurer and Landis, 1995; Maurer et al., 2000), albeit in most cases again for a limited frequency range (most studies investigated f_o levels below 350 Hz). Variation of f_o in speech – albeit in most cases for a more limited frequency range – is discussed in the context of various specific aspects of speech, such as highly emotional, loud or infant-directed speech or with regard to speech within the field of the performing arts (for details, see Chapter M2.4). However, the question of whether or not vowel-specific spectral characteristics relate to f_o levels and the question of the actual f_o range for which vowel sounds are recognisable lie at the core of the understanding of vowel acoustics, independent of vowel context and independent of differentiations such as speech versus singing, relaxed versus loud speech, “normal” versus “emotional” utterances, indoor versus outdoor utterances and so forth.

Until now, no consensus could be established regarding an appraisal of the effect of f_o on vowel recognition, as is explained in more detail in the next chapter. The same holds true for the relation of vowel-related spectral characteristics to f_o: On the one hand, many scholars report a marginal or very limited effect of f_o on the vowel-related spectral characteristics of sounds produced by speakers equal in age (or size) and gender (see the conclusions of Cheveigné and Kawahara, 1999; Barreda and Nearey, 2012). However, most of these studies were based on sounds with values for a limited f_o variation of up to 350 Hz. On the other hand, studies including higher f_o levels indicated a substantial effect of f_o on the vowel spectrum (see the conclusion of Maurer and Landis, 1995; Maurer et al., 2000). Findings of a relation of vowel-related spectral characteristics to f_o were either interpreted as calling for some kind of intrinsic normalisation of f_o and formants, possibly also related to the paralinguistic variation of vocal effort, or as an indication of f_o (and pitch) related spectral representation of vowel quality, a perspective adopted here (see also Maurer, 2018).

In one of our early studies on vowel acoustics (Maurer and Landis, 1995), we investigated the spectral characteristics of natural vocalisations of Swiss German vowel sounds produced by untrained speakers as monophthongs at various f_o levels. As an observational result, the vowel spectrum indeed appeared to be related to f_o: Above all, for close and close-mid vowel sounds, spectral peaks and estimated formants below 1.5 kHz were found to shift upwards with substantially increasing f_o levels. For each of the eight vowels /i, y, e, ø, ɛ, a, o, u/, this observation was documented in terms of sound pairs produced by single speakers at two very different levels of f_o. We named this relation the f_o-dependence of the vowel spectrum. (However, as will be shown in the course of the argument here, this notion has to be further clarified.)

In the Preliminaries (see pp. 158–169), we have extended this type of documentation regarding the relation of the vowel spectrum to f_o by including sounds of one given vowel produced at intermediate f_o levels. However, because the sounds that these examples and illustrations were based on were recorded under varying conditions and with varying sound qualities, and the rights for online playback could not be retrospectively obtained from all speakers, we replicated, further extended and systematised the documentation based on the Zurich Corpus. This replication was conducted in the context of an Interspeech Show and Tell conference presentation (Maurer et al., 2019, online presentation Chapter 5). It is transferred to the present treatise, including additional new functionalities of the online presentation as created for the second version of the Zurich Corpus.

Experiment

Selection of speakers and sounds: Vocalisations of single speakers of the Zurich Corpus were investigated for which systematic series of recognisable vowel sounds are given in the corpus for all Standard German vowels and covering large ranges of f_o (extracts of Parts 1 and 5 of the corpus). The examination of the vocalises documented in the corpus was limited to sounds produced in nonstyle mode with medium vocal effort in V context and with successive f_o variation according to the C-major musical scale.

Based on such an examination, sounds of three single speakers, one child (untrained speaker), one woman (CS singer) and one man (CS singer) were selected for exemplary documentation in this treatise because of the large range of f_o variation covered and the consistency of successful vowel recognition according to vowel intention: For each of the three speakers and each of the eight long Standard German vowels, series of sounds produced within a range of intended f_o of 22 semitones for the child speaker and of 31 and 34 semitones for the adult speakers were compiled. The f_o ranges were 220–784 Hz for the 112 selected sounds of the child, 131–784 Hz for the 152 sounds of the woman and 110–784 Hz for the 168 sounds of the man.

Almost all sounds were recognised in the standard listening test of the Zurich Corpus according to vowel intention, with an 80–100% recognition rate. Exceptions were five sounds that scored a rate of 60% only (/e/ and /ɛ/ produced by the child at intended f_o of 698 and 440 Hz, respectively; /ɛ/ and /o/ produced by the woman at intended f_o of 587 and 659 Hz, respectively; /u/ produced by the man at an intended f_o of 110 Hz).

Acoustic analysis and examination of the relation of the vowel spectrum to f_o: Spectral analysis and calculation of f_o accorded with the standard procedure of the Zurich Corpus. On this basis, the occurrence of spectral variation in relation to f_o was examined for the sounds produced by each speaker.

A note on the sounds produced by the child speaker: For sounds of /o/ and the two intended f_o levels of 659 and 698 Hz, no sound produced with a medium vocal effort that was recognised as /o/ was available. Therefore, for these two f_o levels, recognisable sounds of /o/ produced with low vocal effort were included.

A note on the sounds produced by the adult speakers: The adults were professionally trained singers and actresses/actors and, at the time of the recordings, were actively performing on stage within the musical theatre genre. However, only vocalisations produced in nonstyle mode were selected: According to the sound production and recording procedure of the Zurich Corpus, nonstyle utterances were made by the speakers in terms of favouring the intelligibility of vowel quality over sound timbre. Consequently, and most importantly, the professionals had to attempt to partially or fully abandon their style-specific vocal training (see Chapter 1.1). Therefore, the resulting nonstyle sounds are not to be considered sung in opposition to spoken – in our view, they represent vowel sounds produced at different f_o levels with a focus on clear vowel quality pronunciation without further qualification of the utterances – and they are comparable to sounds of untrained speakers (for verification, see corresponding sounds of untrained speakers in the Zurich Corpus). As said, the two adult speakers were selected for this experiment because of their large vocal ranges in general and the level of successful vowel recognition in the listening test (regarding the sounds they produced within this extensive vocal range) in particular.

Additions: Additional vocalises produced by other speakers manifesting vowel-specific spectral variations will be given in the Additions section online.

Results

Table 1 in the chapter appendix lists the selected sound series and provides sound links for each of the three speakers and each of the eight vowels investigated. (Note that, for each sound, the vowel recognition details are displayed in the Zurich Corpus.) Below, the examination results of the relation of the vowel spectrum to f_o are given first for the vocalises produced by the man (most extensive range of f_o) and subsequently for the vocalises produced by the woman and the child.

Spectral variation of < 1.5–2 kHz due to f_o variation for sounds produced by the man: For the sounds of the close front vowels /i, y/ produced by the man, with increasing f_o above c. 200 Hz, the first harmonic of the spectrum became dominant, and the lowest spectral peak increased in parallel to f_o. The same held true for sounds of the back vowel /u/ for f_o levels above c. 300 Hz, with a decreasing spectral slope starting from f_o levels of ≥ 250 Hz.

For the sounds of the close-mid front vowels /e, ø/ produced by the man, with increasing f_o levels from c. 200 Hz to c. 260 Hz, an upward shift of the lowest spectral peak was indicated (dominant *H*2). The same held true for f_o levels up to c. 330 Hz for sounds of /o/. For higher f_o levels, the sound spectra for these three vowels were difficult to interpret with regard to peak estimation because of the large frequency distance of the harmonics. However, for the sounds of all three vowels, the f_o level of the high-pitched sounds markedly surpassed the frequencies of the first spectral peaks of the low-pitched sounds.

The relation between f_o and F_1 for the sounds of close and close-mid vowels was difficult to assess because of the general methodological problem of formant estimation (see Introduction), and a Klatt resynthesis of some of the calculated F-patterns did not replicate the vowel quality of the natural sounds (author's estimate; for examples, see Table 2 in the chapter appendix). However, when F_1 was interpreted as being near the lowest dominant harmonic for sounds of this kind, a parallel increase of f_o and F_1 was indicated for all sounds of the close and close-mid vowels documented here.

Only limited and inconsistent indications of a variation for vowel-specific lower spectral peaks with increasing f_o were manifest for the sounds of the open-mid vowel /ɛ/, and the sounds of the open vowel /a/ showed no clear indications of such variation.

Spectral variation of < 1.5–2 kHz due to f_o variation for the sounds produced by the woman and the child: Highly comparable spectral characteristics as described for the man's utterances were found for all vocalises of the woman and the child.

Spectral variation of > 1.5 kHz due to f_o variation for sounds of front vowels: No consistent spectral peak shifts in the upper-frequency range above 1.5 kHz were found for rising f_o levels of the sounds of a vowel. However, spectral peak estimation was often problematic because of the increase in the frequency distance of the harmonics and parts of flat energy distribution in the spectra. (Notably, for the sounds of /i/ of the man, a spectral peak in the frequency region of 2400–2800 Hz was manifest up to f_o levels of c. 300 Hz, but it was not manifest for higher f_o levels. This peak "disappearance" may be associated with a register change.)

Discussion

Firstly, the above phenomenological findings and documentation provide a first reference for vocal ranges within which the qualities of long vowels intended by a speaker (with excellent vocal abilities) during vocal production can be successfully recognised. Secondly, the vocalises presented here reconfirm and again document the relation of the lower part of the vowel spectrum to f_o (and pitch) for natural vowel sounds. This relation is not relativised by the methodological problem of formant estimation because, for the sounds of close and close-mid vowels, middle and higher f_o levels of the vocalises of all three speakers surpassed the first spectral peak and also the first estimated F_1 of the sounds at lower f_o levels.

Most importantly, the spectral variation to be observed for sounds of close or close-mid vowels due to f_o variation is not a phenomenon of a specific production style. Above all, it is not a phenomenon to observe only in singing: The speakers were asked to produce the vowel sounds in nonstyle mode, and f_o variation per se appertains to both speech and singing. The documented spectral variation is also not a phenomenon of age or gender differences among the speakers. In fact, in the first place, it is not a phenomenon of speaker differences at all. It is a general phenomenon of the vowel sound *as such* and how the (lower) vowel spectrum is related to f_o (and pitch). This statement does not counter the finding and confirmation of earlier claims that the relation of the lower vowel spectrum to f_o is to be considered nonuniform, above all because it is dependent on vowel qualities and frequency ranges and frequency shifts of f_o. Indeed, the indication of a nonuniform characteristic of the relation in question constitutes the third main aspect of the present documentation of vocalises, which the acoustic theory of the vowel must take into account.

None of the above statements run counter to additional and secondary spectral variations that may arise due to speaker differences or differences in phonation types, other production modes, or vocal effort. These aspects will be addressed in more detail below (above all, see the main Chapters M5 and M7).

To conclude, there are three notions here that are considered important references for vowel acoustics in general and taken as a basis for the following arguments in this treatise in particular: (i) Natural vowel sounds are recognisable within an extensive range of f_o, (ii) vowel-specific spectral characteristics < 1.5–2 kHz are related to f_o (and pitch) and (iii) this relation is nonuniform. The sound series presented further document the lack of a methodological substantiation to estimate F-patterns for all recognisable vowel sounds independent of their f_o level. Notably, the methodological problem already occurred for sounds at f_o below 250 Hz (for exemplary illustration, see Table 2 in the appendix to this chapter.)

Some additional aspects are also worth noting. For the vocalises of /u/, (i) the f_o levels of the high-pitched sounds produced by the adult speakers surpassed the second spectral peak (or the corresponding peak indications) and the calculated F_2 of the low-pitched sounds, (ii) some sounds produced by the adult speakers at middle or higher f_o manifested a dominant first harmonic in a frequency range in which the sounds at low f_o manifested a relative spectral minimum in between two spectral peaks, an aspect referred to here as inversion (for details of this matter, see Chapter M7.2), (iii) for the sounds produced by all speakers at middle or higher f_o, there was only a single lower spectral energy maximum manifest near or at the frequency of the first harmonic, and (iv) a Klatt resynthesis related to the calculated LPC filter curves of sounds produced by all speakers at high f_o, this f_o level also applied to the resynthesis, produces in many cases sounds with /u/–like vowel qualities (author's estimate; reproducible using the Klatt tool in the user interface of the Zurich Corpus; thus, although there is no methodological substantiation given for the LPC filters to be understood as representing formants, the resynthesis based on these filters does not substantially affect the recognised vowel quality; however, if the f_o level is stepwise lowered, the vowel quality changes to /o/, /ɔ/ and eventually even to /a/).

For the sounds of /a/ produced by the speakers at an f_o level above 350 Hz, the methodological problem of identifying two lower spectral peaks and estimating two lower formant frequencies < 1.5–2 kHz is well exemplified in the present documentation.

For the sounds of /ø/, the often weak spectral peak structure < 2 kHz for sounds above 400 Hz contrasted with the clear peak structure of low-pitched sounds, above all for the vocalises of the adults.

Chapter appendix

Table 1. Vocalises of the eight long Standard German vowels produced by a child, a woman and a man. The sounds shown were produced in nonstyle mode with medium vocal effort in V context and with successive f_o variation according to the C-major musical scale. Column 1 = vowel intended and recognised (V). Columns 2 and 3, 4 and 5, and 6 and 7 = sound series and sound links (S/L) and number of sounds (N) for the three speakers.
[M-02-01-T01]

Table 2. Examples of natural vowel sounds for which the intended and recognised vowel quality is not maintained in Klatt resynthesis based on estimated *F*-patterns. Column 1 = vowel intended and recognised (V). Columns 2 and 3, and 4 and 5 = sound series and sound links (S/L) and number of sounds (N) for the two adult speakers.
[M-02-01-T02]

Table 1. Vocalises of the eight long Standard German vowels produced by a child, a woman and a man. [M-02-01-T01]

V	**Child**		**Woman**		**Man**	
	fo = 220–784 Hz		**fo = 131–784 Hz**		**fo = 110–784 Hz**	
	(22 semitones)		**(31 semitones)**		**(34 semitones)**	
	S/L	**N**	**S/L**	**N**	**S/L**	**N**
i	1 ↗	14	9 ↗	19	17 ↗	21
y	2 ↗	14	10 ↗	19	18 ↗	21
u	3 ↗	14	11 ↗	19	19 ↗	21
e	4 ↗	14	12 ↗	19	20 ↗	21
ø	5 ↗	14	13 ↗	19	21 ↗	21
o	6 ↗	14	14 ↗	19	22 ↗	21
ɛ	7 ↗	14	15 ↗	19	23 ↗	21
a	8 ↗	14	16 ↗	19	24 ↗	21

Table 2. Examples of natural vowel sounds for which the intended and recognised vowel quality is not maintained in Klatt resynthesis based on estimated *F* -patterns. [M-02-01-T02]

V	**Woman**		**Man**	
	fo = 169–181 Hz		**fo = 179–222 Hz**	
	S/L	**N**	**S/L**	**N**
i	1 ↗	2	3 ↗	2
u	2 ↗	2	4 ↗	2

M2.2 Isolated Vowel Sounds Produced at High Levels of f_o

Introduction

In the literature, the upper-frequency limit of f_o for vowel recognition is a matter of debate. In short, three positions can be identified (see Maurer et al., 2014): According to the first, vowel sounds lose their intelligibility if f_o surpasses the statistical F_1 assumed to acoustically represent vowel quality (see e.g. Joliveau et al., 2004b). According to the second position, vowel recognition of all vowels can be maintained up to f_o levels of approximately 500 Hz; for f_o levels higher than 500 Hz, successful vowel differentiation drops substantially, and the sounds tend to be recognised as /a/–like (see e.g. Sundberg, 2012). According to the third position, vowel intelligibility can be maintained for f_o levels of up to 660–1050 Hz, depending on vowel quality, conditions of vowel production and related listening tests (see e.g. Smith and Scott, 1980; Maurer and Landis, 1996; Smith et al., 2005; Maurer et al., 2014; Friedrichs, Maurer and Dellwo, 2015; Friedrichs, Maurer, Suter and Dellwo, 2015; see also the Preliminaries, Chapter M8.2).

In the previous chapter, examples of recognisable vowel sounds up to calculated f_o in the frequency range of 700–800 Hz are documented, supporting the third position mentioned above. Further examining the entire sample of the Zurich Corpus with regard to all eight long Standard German vowels, we found numerous sounds in the f_o range of 700–800 Hz, produced in V context by ten or more different speakers, that reached a recognition rate of 100% (5/5 listeners) according to vowel intention. However, the number of recognised sounds in this f_o range markedly depended on vowel qualities: Numerous fully recognised sounds of the vowels /i, y, a, u/ were documented for this range of f_o, but fewer recognised sounds for the vowels /e/ and /ɛ/ and only a few recognised sounds for the vowels /ø/ and /o/. Similarly, numerous sounds produced in V context in the range of 950–1100 Hz by ten or more different speakers and with a recognition rate of 100% are documented in the corpus for the vowels /i, y, a, u/, with the fewest examples for /a/, a higher number of sounds for /y, u/, and most sounds for /i/. Note, in this context, that the vowels /i, a, u/ represent corner positions of the periphery of a vowel triangle or a vowel quadrilateral (no differentiation is made here between /a/ and /ɑ/), marking the extreme oppositions close–open and front–back, and that /y/ represents a middle position in between the close front and close back corner positions.

Thus, the vowel sounds of the Zurich Corpus discussed here confirm the assumptions made in the above third position (and the related studies) regarding successful possible vowel recognition up to 660–1050 Hz. However, the listening test conducted during the creation of the Zurich Corpus was based on entire sounds (including on- and offsets), and the sounds of each speaker were tested separately (speaker-blocked test condition), further separating nonstyle and style productions. (Sounds produced in V and sVsV conditions were tested together.) The sound series of the previous chapter have to be considered in this context.

In view of the foregoing, in two experiments, we addressed the question of whether or not successful vowel recognition of entire sounds produced at high f_o, as found for numerous sounds of the Zurich Corpus, could be confirmed if only the respective sound nuclei were investigated (excluding on- and offsets) and if sounds of different speakers were mixed in the listening tests conducted.

Experiment 1

Sound selection: Based on the vowel sounds documented in the Zurich Corpus, which were unanimously recognised according to vowel intention by the five standard listeners involved in the standard listening test when building up the corpus, for each of the eight long Standard German vowels, 20 sounds of ten or more different speakers, produced in V context at calculated f_o in the frequency range of 700–800 Hz, were selected by the author (best sound and vowel quality according to the author's estimate). Production style and vocal effort of the sounds were ignored in this experiment.

Sound editing: On- and offsets of the selected sounds were extracted. If the sound duration after the extraction was > 1 sec., the middle 1 sec. sound nucleus was used for subsequent investigation; otherwise, the sound fragment after deleting the on- and offset was used. Finally, a fade in/fade out of 0.05 sec. was applied, and the sounds were normalised in amplitude. As a result, a sample of 160 sound nuclei with calculated f_o in the frequency range of 700–800 Hz and with a sound duration in the range of 0.5–1 sec. was compiled.

Acoustic analysis: For all natural sounds and their sound nuclei, acoustic analysis was conducted according to the standard procedure of the Zurich Corpus.

Listening test: The standard listeners of the Zurich Corpus performed the vowel recognition test. The entire sample (in random order of the sounds) was divided into three test subsets of 55, 55 and 50 sounds, which were separately tested in order not to overstrain the listeners. For each subset, the listeners first listened twice to all sounds without assigning vowel qualities in order to become familiar with the high-pitched sound nuclei. Subsequently, the actual test was run: For a single sound, the listeners were asked to assign one of the eight long Standard German vowels, or ɔ, or schwa, or “no vowel recognised” (forced choice, no vowel boundaries). All other parameters of the listening test accorded with the standard procedure of the Zurich Corpus.

Results of Experiment 1

Table 1 in the chapter appendix summarises the recognition rates for the sounds of the vowels. As the table shows, for the close vowels /i, y, u/, the vowel recognition rate for the sound nuclei proved to be 100% in most cases, equal to the recognition rate of the original sounds, including on- and offsets. The same holds true for the open vowel /a/. For the open-mid vowel /ɛ/, in 18 of 20 cases, the vowel recognition could either be maintained at 100% or dropped slightly to 80%. Conversely, for the sounds of the close-mid vowels /e, ø, o/, vowel recognition proved to be substantially impaired or even confused for the sound nuclei compared with the original sounds, above all for sounds of /ø, o/: Only three sounds of /o/ were recognised with a rate of 80%, and only seven sounds of /ø/ were recognised with a rate of 80–100%.

However, despite this impairment, the results indicate that speakers with excellent vocal abilities can produce sounds of all long Standard German vowels at *f*ₒ levels in the range of 700–800 Hz in a way that listeners with experience in vowel recognition tests can differentiate and recognise them on the basis of isolated sound nuclei. In order to document this phenomenon, “best” examples in terms of three sounds for each vowel with the highest recognition rates were selected by the author. These are listed in Table 2, including links to the original sounds and respective sound nuclei.

As said, production styles were not taken into account in the experiment. However, the number of sounds in the tested sample associated with stylistic vowel production was very small: Only 22 sounds of a total of 160 sounds were produced in ST style (eight sounds) or CS style (14 sounds), and only three of these sounds (one sound in ST style and two sounds in CS style) were selected as “best” examples, as documented in Table 2.

As could be expected, if recognised vowel quality differences were found for sounds produced at high *f*ₒ levels, corresponding spectral differences could also be observed. However, an exact formulation of these differences is not a simple matter. This will be addressed in detail in the context of a general discussion of the relation between vowel recognition and related characteristics of harmonic spectra (see the below excursus on vowel quality and harmonic spectrum).

Note also that *H*1 is either not very pronounced or does not manifest a relative spectral energy maximum for the sounds of /i/ and /y/, and that the sounds of /ɛ/ show flat spectral energy distribution.

Experiment 2

Sound selection, sound editing and acoustic analysis: The same procedures of sound selection, sound editing and acoustic analysis as described for experiment 1 were applied for sounds of the vowels /i, y, a, u/ produced at calculated f_o in the frequency range of 950–1100 Hz. As a result, a sample of 80 sound nuclei with a sound duration in the range of 0.5–1 sec. was compiled.

Listening test: This sample (in random order of sounds) was divided into three test subsets of 30, 30 and 20 sounds, respectively. Vowel recognition was tested for each subset separately, with the listening test procedure according with experiment 1. The listeners were informed that vowel quality distribution within a subset was not uniform.

Results of Experiment 2

Table 3 in the chapter appendix summarises the recognition rates for the sounds of the four vowels investigated. As the table shows, for the majority of the sounds of the close vowels, the vowel recognition rate for the sound nuclei proved to be 100% or 80%, again equal to or near to the recognition rate of the original sounds, including on- and offsets. In contrast, for the open vowel /a/, the recognition rate for the sound nuclei dropped substantially: A 100% rate was obtained for only one sound, an 80% rate was obtained for four sounds and a 60% rate for another four sounds. The other sounds were not recognised by a majority of the listeners according to vowel intention and according to the results of the standard listening test of the corpus (including on- and offsets of the sounds and conducted with speaker- and style-blocked conditions). However, investigating the confusion matrix for the sounds of /a/ as shown in Table 4, most of the sounds were assigned to open-mid or open vowels, and with one exception, no confusion with a close vowel occurred.

In these terms, the results indicated that speakers with very good vocal abilities can produce sounds of the long Standard German corner vowels and the intermediate vowel /y/ in between /i/ and /u/ at f_o levels of c. 1 kHz in a way that listeners with experience in vowel recognition tests can differentiate and recognise them on the basis of sound nuclei only, above all in terms of differentiation of corner positions close versus open and front versus back, including rounded versus unrounded. In order to document this phenomenon, “best” examples in terms of three sounds for each vowel with the highest recognition rates were again selected by the author. These are listed in Table 5, including links to the original sounds and respective sound nuclei. (Please use state-of-the-art headphones to listen to the sounds.)

As was the case in experiment 1, production styles were not taken into account, but the number of sounds related to style-specific production was again very small: Only seven of a total of 80 sounds were produced in ST style (three sounds), CS style (three sounds) or EC style (one sound), and none of these sounds is shown in Table 5 presenting “best” examples.

Discussion

As a main finding, the two experiments and their results confirmed that vowel recognition of entire sounds produced in V context (including on- and offsets), as documented in the Zurich Corpus, can also be demonstrated for a substantial part of the respective sound nuclei (excluding on- and offsets) for all long Standard German vowels in the f_o range of 700–800 Hz and for the vowels /i, y, a, u/ at an f_o of approximately 1 kHz. This general result is in line with the findings of Maurer and Landis (1996) and Maurer et al. (2014), with some even higher f_o values in the current study. The same holds true with regard to the findings of Friedrichs et al. (2017) concerning recognisable sounds of the corner vowels /i, a, u/ up to f_o of c. 1 kHz (see also Zhang et al., 2022, for the differentiation of /i, y, e/ versus /a/ versus /o, u/). However, in the present study, a higher recognition rate was found for the sounds of /y/ at f_o of c. 1 kHz, and higher recognition rates were found for the sounds of /e, ø, ε, o/ at f_o in the range of 700–800 Hz than in Friedrichs et al. (2017) and Zhang et al. (2022). Further, the rates of the present study are also in line with the findings of Friedrichs, Maurer and Dellwo (2015) and Friedrichs, Maurer, Suter and Dellwo (2015), who reported maintained 80% vowel recognition rates for minimal pairs up to an f_o range of 740–880 Hz (entire sounds; in this context, note also the correspondence of the results with the high-pitched sounds of the vocalises documented in the previous chapter.) In contrast, the results did not confirm the very differentiated recognition of /i, I, ε, æ/ reported by Smith and Scott (1980) at f_o of c. 1 kHz.

Thus, evidence is again provided that the f_o range of successful vowel recognition covers the entire range of F_1 for all vowels, as given in almost all formant statistics. (Concerning statistical F-patterns for sounds of Standard German produced by women and men, the frequency range up to 700–800 Hz represents the F_1 range related to sounds of all vowels, including open vowels; see e.g. Maurer et al., 1992; Pätzold and Simpson, 1997.) Notably, if sounds of all long vowels at f_o of 700–800 Hz can be recognised, it is evident that no formant concept can account for vowel differentiation when f_o surpasses the level of statistical F_1 for almost all vowels. Moreover, if sounds of the close corner vowels can be recognised even at f_o levels up to 1 kHz, it is evident that "undersampling" of the assumed resonances of the vocal tract is not directly related to an impairment of vowel recognition, especially considering that the reported statistical F_1 of close vowels is the lowest of all vowels.

The results further indicated that style-specific productions are in many cases less recognisable than nonstyle productions (when building up the two samples of experiments 1 and 2, most of the recognised vowel sounds found in the Zurich Corpus were produced in nonstyle mode), a conclusion which is important to consider for existing studies of singing, above all for singing in EC style. In our understanding, the tendency observed in the literature to generally understand vowel sounds of the middle and higher f_o levels as sung vowels is misleading. As we argue – and as we demonstrate below (see Chapter M2.4) – there is no principal difference between recognisable spoken and sung vowels that would directly relate to f_o levels of sounds.

The sounds produced at high f_o levels investigated and documented here represent vowel sounds at the upper limit of the vocal ranges of the speakers. Therefore, many sounds manifest phonation stress. However, this does not limit the recognition results.

Chapter appendix

Table 1. Recognisable sound nuclei of natural vowel sounds at calculated f_o in the frequency range of c. 700–800 Hz: Vowel recognition results of experiment 1. Columns 1–3 = sounds (VO = vowel openness; c = close vowels, c-m = close-mid vowels, o-m = open-mid vowel, o = open vowel), vowel quality intended by the speakers (V) and range of calculated average f_o (in Hz) for the mid 0.3 sec. of the sound nuclei. Columns 4–7 = number of sounds per vowel recognition rate according to vowel intention (in %).
[M-02-02-T01]

Table 2. Recognisable sound nuclei of natural vowel sounds at calculated f_o in the frequency range of c. 700–800 Hz: "Best" examples of experiment 1. Three sounds per vowel produced at f_o in the frequency range of c. 700–800 Hz are shown. Columns 1–7 = sounds (VO = vowel openness; V/L = vowel quality intended by the speaker and sound link to the original reference sound and its sound nucleus; SP = ID number of the speaker in the Zurich Corpus; SG = speaker group, where c = children, w = women, m = men; R org = record number of the original reference sound in the Zurich Corpus; R nuc = record number of the extracted sound nucleus in the Zurich Corpus; fo = calculated average f_o for the mid 0.3 sec. of the sound nucleus, in Hz). Columns 8 and 9 = vowel recognition (V = the five vowel qualities labelled by the five listeners; % = vowel recognition rate according to vowel intention, in %).
[M-02-02-T02]

Table 3. Recognisable sound nuclei of natural vowel sounds at calculated f_o in the frequency range of c. 950–1100 Hz: Vowel recognition results of experiment 2. Columns 1 and 2 = sounds (V = vowel quality intended by the speakers; fo = range of calculated average f_o for the mid 0.3 sec. of the sound nuclei, in Hz). Columns 3–6 = number of sounds per vowel recognition rate according to vowel intention.
[M-02-02-T03]

Table 4. Recognisable sound nuclei of natural vowel sounds at calculated f_o in the frequency range of c. 950–1100 Hz: Confusion matrix for the sounds of /a/ of experiment 2. Columns 1 and 2 = sounds (V = vowel quality intended by the speakers; fo = calculated average f_o for the mid 0.3 sec. of the sound nuclei, in Hz). Column 3–10 = confusion matrix (ns = not specified, no vowel quality assigned).
[M-02-02-T04]

Table 5. Recognisable sound nuclei of natural vowel sounds at calculated f_o in the frequency range of c. 950–1100 Hz: "Best" examples of experiment 2. Three sounds per vowel produced at f_o in the frequency range of c. 950–1100 Hz are shown. Columns, see Table 2.
[M-02-02-T05]

Table 1. Recognisable sound nuclei of natural vowel sounds at calculated fo in the frequency range of c. 700–800 Hz: Vowel recognition results of experiment 1. [M-02-02-T01]

Sounds			Sounds per vowel recognition rate			
VO	**V**	**fo (Hz)**	**100%**	**80%**	**60%**	**<50%**
c	i	700–798	19	–	1	–
	y	703–799	16	4	–	–
	u	700–799	15	5	–	–
c-m	e	704–801	6	9	5	–
	ø	702–789	1	6	9	3
	o	702–801	0	3	11	6
o-m o	ɛ	703–798	9	9	1	1
	a	710–798	19	–	1	–

Table 2. Recognisable sound nuclei of natural vowel sounds at calculated fo in the frequency range of c. 700–800 Hz: “Best” examples of experiment 1. [M-02-02-T02]

Sounds							Vowel recognition	
VO	**V/L**	**SP**	**SG**	**R org**	**R nuc**	**fo (Hz)**	**V**	**%**
c	i ↗	1102	w	191326	205045	798	i i i i i	100
	i ↗	1023	w	115689	204889	797	i i i i i	100
	i ↗	1032	w	138213	204915	790	i i i i i	100
	y ↗	1069	m	165671	204980	796	y y y y y	100
	y ↗	1023	w	115717	204891	794	y y y y y	100
	y ↗	1063	m	192200	205048	763	y y y y y	100
	u ↗	1036	w	138968	204922	799	u u u u u	100
	u ↗	1068	w	163543	204972	792	u u u u u	100
	u ↗	1059	w	176140	205017	792	u u u u u	100
c-m	e ↗	1023	w	175152	205015	801	e e e e e	100
	e ↗	1036	w	170979	204995	778	e e e e e	100
	e ↗	1001	w	189182	205040	765	e e e e e	100
	ø ↗	1034	c	183318	205032	766	ø ø ø ø ø	100
	ø ↗	1056	c	143466	204935	715	ø ø ø ø ø	100
	ø ↗	1037	c	121865	204905	753	e ø ø ø ø	80
	o ↗	1070	m	179232	205022	792	ɔ o o o o	80
	o ↗	1046	w	160056	204953	759	ɔ o o o o	80
	o ↗	1069	m	174602	205004	715	o o o o u	80
o-m/o	ɛ ↗	1034	c	120995	204902	781	ɛ ɛ ɛ ɛ ɛ	100
	ɛ ↗	1069	m	174632	205008	780	ɛ ɛ ɛ ɛ ɛ	100
	ɛ ↗	1052	w	160687	204957	777	ɛ ɛ ɛ ɛ ɛ	100
	a ↗	1023	w	116444	204896	796	a a a a a	100
	a ↗	1050	m	136671	204913	796	a a a a a	100
	a ↗	1052	w	142053	204926	791	a a a a a	100

Table 3. Recognisable sound nuclei of natural vowel sounds at calculated fo in the frequency range of c. 950–1100 Hz: Vowel recognition results of experiment 2. [M-02-02-T03]

Sounds		Sounds per vowel recognition rate			
V	**fo (Hz)**	**100%**	**80%**	**60%**	**<50%**
i	982–1084	16	3	–	1
y	960–1067	14	3	3	–
u	962–1058	11	6	1	2
a	950–1042	1	4	4	11

Table 4. Recognisable sound nuclei of natural vowel sounds at calculated fo in the frequency range of c. 950–1100 Hz: Confusion matrix for the sounds of /a/ of experiment 2. [M-02-02-T04]

Sounds		Vowel recognition (confusion matrix)							
V	**fo (Hz)**	**y**	**e**	**ø**	**ɛ**	**a**	**ɔ**	**o**	**ns**
a	1035					5			
	1032					4	1		
	1004				1	4			
	985				1	4			
	982					4	1		
	1009				1	3	1		
	994				2	3			
	989				2	3			
	988			1	1	3			
	1042				2	2	1		
	1022				2	2	1		
	990				1	2	1		1
	984				1	2	1	1	
	982				2	2		1	
	979			1	2	2			
	984			2	2	1			
	968			2	1	1	1		
	950				1	1	1	2	
	1015		1	1	3				
	992	1			4				

Table 5. Recognisable sound nuclei of natural vowel sounds at calculated fo in the frequency range of c. 950–1100 Hz: “Best” examples of experiment 2. [M-02-02-T05]

Sounds							Vowel recognition	
VO	**V/L**	**SP**	**SG**	**R org**	**R nuc**	**fo (Hz)**	**V**	**%**
c	i ↗	1046	w	147231	205078	1084	i i i i i	100
	i ↗	1023	w	174782	205125	1058	i i i i i	100
	i ↗	1068	w	163875	205103	1049	i i i i i	100
	y ↗	1052	w	142095	205072	1067	ü ü ü ü ü	100
	y ↗	1023	w	174784	205126	1043	ü ü ü ü ü	100
	y ↗	1059	w	176343	205132	1023	ü ü ü ü ü	100
	u ↗	1018	w	164390	205106	1058	u u u u u	100
	u ↗	1039	w	171760	205122	1056	u u u u u	100
	u ↗	1023	w	161500	205091	995	u u u u u	100
o	a ↗	1023	w	174793	205127	1035	a a a a a	100
	a ↗	1046	w	163321	205100	1004	ɛ a a a a	80
	a ↗	1023	w	161504	205093	985	ɛ a a a a	80

M2.3 Minimal Pairs Produced at High Levels of f_o

Introduction

As indicated in Chapter 1.1, during the creation of the Zurich Corpus (see Part 3 of the corpus), minimal pairs produced at various levels of f_o by selected speakers with excellent vocal abilities were recorded, and vowel recognition was tested. Two studies analysing sets of minimal pairs produced by a woman at intended f_o of up to 880 Hz were published earlier (see Friedrichs, Maurer and Dellwo, 2015; Friedrichs, Maurer, Suter and Dellwo, 2015), and additional sets produced by two women at various levels of intended f_o up to 1047 Hz were investigated in the present treatise. These three experiments are described and discussed in this chapter, with a focus on the recognition results for utterances produced at intended f_o of 784, 880 and 1047 Hz, and exemplary sound series are presented.

Experiment 1

Speaker, utterances and f_o levels: Friedrichs, Maurer and Dellwo (2015) investigated sounds of all eight long Standard German vowels in a minimal pair context, each minimal pair produced as a word pair in one utterance (one recording) by a woman. The speaker was asked to focus on the intelligibility of speech and, if necessary, to ignore the aesthetic qualities of her singing and acting style. Table 1 in the chapter appendix shows the minimal pairs and vowel contrasts.

The minimal pairs were produced and recorded in two runs in AB and BA order at nine f_o levels of 220–440–587–659–698–740–784–831–880 Hz. The lowest f_o level corresponded to the statistical average f_o for utterances in citation-form words (see e.g. Hillenbrand et al., 1995), and the entire frequency range of the investigated f_o covered the range of statistical average F_1 for Standard German vowels produced by women (see e.g. Pätzold and Simpson, 1997).

Sound selection, sound editing: The word pair (AB or BA) that, per the estimate of the second author of Friedrichs, Maurer and Dellwo (2015), had a perceptually more salient vowel contrast was chosen for further investigation. Subsequently, two sound sets were prepared for the vowel recognition tests, the first consisting of single words extracted from the recordings of the selected word pairs, the second consisting of steady-state vowel nuclei (middle 250 ms of the vowel sound in question) extracted from these single words, resulting in two experimental conditions, words and isolated vowels.

Listening tests: Two recognition tests involving 40 native German listeners (students of the University of Zurich) were performed. Listeners were randomly divided into two groups (20 per group; one group for the word condition and one for the isolated vowel nucleus condition; gender balanced across groups). In the first test, single words in random order were presented to the listeners via headphones, and the minimal pair that the presented word was extracted from was presented on a computer screen. Listeners were asked to assign the sound signal to one of the two displayed words. The same procedure was applied to the vowel nuclei in the second test.

Analysis of the recognition results: In Friedrichs, Maurer and Dellwo (2015), the listeners' identification performance was given as calculated with the bias-free non-parametric sensitivity measure A' (Signal Detection Theory, see Stanislaw and Todorov, 1999) with Praat scripts written by the third author according to formulas in Pallier (2002). For the present treatise, in addition, the corresponding values in % were also calculated.

Results 1

Table 1 in the chapter appendix shows the results of the recognition tests of experiment 1 for the two levels of f_o of 784 and 880 Hz. Included in the table are links to the minimal pairs. As the results indicate, for the word condition, vowel contrast recognition for the minimal pairs according to the intention of the speaker was generally maintained up to intended f_o of 784 and 880 Hz with a rate of ≥ 90%, except for the contrasts of /y/–/ø/, /e/–/ø/ and /e/–/ɛ/, which were recognised in the range of 82–88%. However, note that the contrast /y/–/ø/ was recognised with a rate of 97% for the f_o level of 880 Hz, despite the drop in the rate for the f_o level of 784 Hz to 85%. This result may have been a consequence of within-speaker differences in pronunciation.

For the sound nucleus condition, the vowel contrast recognition rate dropped for some of the minimal pairs; however, for the sounds at f_o of 784 Hz, recognition was maintained at ≥ 80%, except for /y/–/ø/ (78%) and /e/–/ø/ (72%), and the same held true for the sounds at f_o of 880 Hz, except for /e/–/ɛ/ (72%), /ø/–/ɛ/ (78%) and /ɛ/–/a/ (63%).

For further details of results and interpretation of the recognition performance measured with A', see Friedrichs, Maurer and Dellwo (2015). Note that, with only one exception, calculated f_o levels of all vowel sound nuclei of this experiment did not exceed ± 4% of the intended levels. The same held true for the sounds of the subsequent experiments.

Experiment 2

Speaker, utterances and f_o levels: Friedrichs, Maurer, Suter and Dellwo (2015) investigated “lVgen” utterances related to the seven Standard German vowels /i, y, e, ø, ɛ, a, o/ (all long vowels except /u/), the words produced by a woman (same speaker as in experiment 1) as minimal pairs for all possible vowel contrasts of all front vowels and /a/ and for the contrast of /a/ and the back vowel /o/. “lVgen” minimal pairs concern vowel contrasts of long Standard German vowels in the words “liegen–lügen–legen–lögen–lägen–lagen–logen”. The “lVgen” variants were paired into sets of minimal pairs (one recording, two words), first in AB order, then in a second run in BA order. f_o was varied for all minimal pairs in AB and BA order, with f_o levels at 220–440–587–659–698–740–784–831–880 Hz. The speaker was again asked to focus on the intelligibility of speech and, if necessary, ignore the aesthetic qualities of her singing and acting style. Because one of the two listening tests was performed by inexperienced listeners (see below), the vowel /u/ was excluded from the investigation in this experiment because the word “lugen” (“ausschauen, spähen”) is outdated and therefore uncommon in the German language spoken today.

Sound selection, sound editing: Listening to the utterances (first and second author of the above-mentioned study), for each vowel and each f_o level investigated, one “lVgen” token that appeared to manifest the optimal correspondence between intended and perceived vowel quality was chosen for further investigation and was extracted from the corresponding recording as a single word.

Listening test A: A recognition test involving 28 native German listeners (students of the University of Zurich) was performed. In the test, single “lVgen” utterances in random order were presented to the listeners via headphones, and buttons labelled with the seven “lVgen” options were presented on a computer screen, arranged in random order in a circle. The listeners had to assign the word they heard to one of the seven “lVgen” options presented on the screen (seven-alternative forced-choice word identification task; for further details of the method, see Friedrichs et al., 2015b).

Listening test B: Vowel recognition for the “lVgen” utterances (same sample as used for test A) was also tested according to the standard procedure of the Zurich Corpus and involving the five standard listeners. This listening test was not a part of the earlier study, but its results are integrated into the discussion of vowel recognition at high levels of f_o.

Results 2

Table 2 in the chapter appendix shows the results of the recognition tests of experiment 2 for the two intended f_o levels of 784 and 880 Hz. The table includes sound links.

Based on listening test A involving 28 inexperienced listeners, vowel recognition rates of the sounds at high levels of f_o were found as:

- > 90% for /i, y, ø, a, o/ and > 50% for /e, ɛ/ at an intended f_o of 784 Hz
- > 95% for /i, y, a, o/, 86% for /ɛ/, 64% for /ø/ and 50% for /e/ at an intended f_o of 880 Hz.

However, note that the word with the vowel /ɛ/ was recognised with a rate of 86% for the f_o level of 880 Hz, despite the drop in the rate to 57% for the f_o level of 784 Hz. This result may again have been a consequence of within-speaker differences in pronunciation.

Based on listening test B involving the five experienced listeners of the Zurich Corpus, the vowel recognition rates for the sounds at high levels of f_o were as follows:

- 100% for /i, y, ø, ɛ, o, a/ and 80% for /e/ at an intended f_o of 784 Hz
- 100% for /i, y, e, ɛ, o, a/ and 80% for /ø/ at an intended f_o of 880 Hz

In general, the study indicates that sounds of long vowels produced in minimal pair context can be recognised up to an f_o of 880 Hz, with a higher recognition rate for experienced listeners compared with inexperienced listeners and, by tendency, with a more robust recognition of the close and open vowels compared with close-mid and open-mid vowels.

When considering the high recognition rate for /o/, note that /u/ was not given as an option in the recognition task.

For further details, see the confusion matrices in Friedrichs et al. (2015b). For the documentation of the investigated sounds produced at high f_o levels and the related spectra of the vowel nuclei, see the sound links in Table 2.

Experiment 3

Speakers, utterances and f_o levels: In addition to the previous studies and for the present discussion of recognisable vowel sounds at high f_o levels, utterances of two selected speakers of the Zurich Corpus (women, CS singers) who produced utterances of “lVgen”, “bVden” and “schVf” as single words at f_o levels according to the C-major scale up to 1047 Hz (extract of Part 3 of the corpus) were investigated.

The “lVgen” minimal pairs involved vowel contrasts of all long Standard German vowels. Above all, the enlarged set of sounds of the Zurich Corpus was used in addition to experiment 2 in order to investigate all long vowels (including /u/), the upper-frequency range of f_o of 880–1047 Hz of vowel sound production and recognition, and vowel recognition of very experienced listeners.

“bVden” minimal pairs involved the vowel contrasts /a–o–u/ (“baden–Boden–Buden”). This set was used to re-examine vowel differentiation of back vowels and /a/ at very high pitches.

“schVf” minimal pairs involved the corner vowel contrasts /i–a–u/ (“schief–Schaf–schuf”). This set was used to re-examine vowel differentiation of corner vowels at very high pitches.

Listening test: A recognition test (entire words) according to the standard procedure of the Zurich Corpus was performed. (Note that, according to the standard procedure, the test was performed in speaker-blocked condition.)

Selection of “best” cases and related vowel recognition rates: For each of the three sets of minimal pairs, each vowel and each of the two f_o levels of 880 Hz and 1047 Hz, the utterance with the highest recognition rate (vowel recognition according to vowel intention) in terms of “best” cases was selected for presentation and discussion in this treatise.

Results 3

Table 3 in the chapter appendix shows the results of the recognition test of experiment 3 for the selected utterances as “best” cases. The table includes sound links. For all vowels of all three sets of minimal pairs, vowel recognition of the selected utterances was maintained up to an intended f_o of 880 Hz with a rate of 100%. For the selected “lVgen” utterances produced at an intended f_o of 1047 Hz, the recognition rate was maintained at 100% for all vowels except /e/ and /a/: At this highest range of f_o investigated, the recognition rate for /e/ was 80%, with an additional labelling of vowel quality in the /e–i/ boundary. The recognition rate for /a/ was 60%, with two additional labellings of the back vowel /o/. However, both utterances of /a/ in the second and third sets of minimal pairs produced at an intended f_o of 1047 Hz were fully recognised, as was true for the utterance of /u/ and the utterance of /i/ in the second set. At that f_o level, only the recognition rate for /o/ in “bVden” context slightly dropped to 80%, with an additional labelling of vowel quality in the /o–u/ boundary.

Discussion

The three experiments on vowel recognition for minimal pairs at high levels of f_o differed in their method (words, sound production, entire sounds versus extraction of sound nuclei, listening tests), and they also differed somewhat in their results, above all concerning the vowel quality-related decrease of the recognition rate at high f_o. Generally, however, the results showed that minimal pairs with long vowels produced by speakers with exceptional vocal abilities in the range of intended f_o of 784–880 Hz could be successfully differentiated, both in production and perception. Furthermore, for the five experienced listeners, vowel differentiation was maintained up to f_o of 1047 Hz, not only for the corner vowels but also for other vowels or vowel contrasts. Thus, referring to Friedrichs, Maurer and Dellwo (2015) and integrating their conclusion into the present context, the examples presented here indicate that the phonological function of vowels in word context can be maintained at fundamental frequencies up to c. 1 kHz.

Depending on vowel qualities, the recognition rate for sound nuclei dropped somewhat in experiment 1 compared to the recognition rate obtained for the respective entire sounds. However, the results indicated that isolated sound nuclei of long vowels produced in the context of minimal pairs could be recognised above chance level at an f_o of 880 Hz.

These findings of vowel recognition for sounds produced in minimal pair context are in line with the corresponding findings for isolated, entire vowel sounds (including on- and offsets) as well as the finding that the impairment of vowel recognition at high f_o for extracted sound nuclei was limited in its extent (see the previous two chapters). These findings are also in line with the findings of a study of single syllables and isolated vowel sounds produced by a female Cantonese opera singer at various levels of f_o in a live performance, which demonstrated recognisable vowel sounds at levels of f_o of up to 700–860 Hz (see Maurer et al., 2014).

Chapter appendix

Table 1. Recognisable German minimal pairs produced at intended f_o of 784 and 880 Hz: Vowel recognition results of experiment 1. Columns 1 and 2 = minimal pairs with words and vowel contrasts (VC) tested and sound links (L; for a single vowel contrast and the two intended f_o levels of 784 and 880 Hz, the link relates to the word pairs investigated). Columns 3–6 = recognition results for entire utterances, given as A' (see text) and recognition rates in % for intended f_o of 784 Hz and 880 Hz. Columns 7–10 = recognition results for the extracted vowel sound nuclei. Colour code: Dark blue = recognition rate ≥ 90%; light blue = recognition rate = 80–89%.
[M-02-03-T01]

Table 2. Recognisable German "lVgen" minimal pairs produced at intended f_o of 784 and 880 Hz: Vowel recognition results of experiment 2. Columns 1 and 2 = sounds (MP = minimal pair context; V = vowel qualities tested). Columns 3 and 4 = vowel recognition results of test A (in %; 28 inexperienced listeners) for intended f_o of 784 Hz and 880 Hz of the utterances. Columns 5 and 6 = vowel recognition results of test B (in %; five experienced standard listeners of the Zurich Corpus). Last row = sound links for the two samples at f_o of 784 Hz and 880 Hz. Colour code: Dark blue = recognition rate ≥ 90%; light blue = recognition rate = 80–89%.
[M-02-03-T02]

Table 3. Recognisable German "lVgen", "bVden" and "schVf" minimal pairs produced at intended f_o of 880 and 1047 Hz: Vowel recognition results of experiment 3. Columns 1 and 2 = sounds (MP = minimal pair context; V = vowel qualities tested). Columns 3–6 = vowel recognition results (identifications of the five standard listeners of the Zurich Corpus and labelling majority in %) for intended f_o of 880 Hz and 1047 Hz of the utterances. Last row of a section = sound links for the two samples at f_o of 880 Hz and 1047 Hz. Colour code: Dark blue = recognition rate = 100% (5 of 5 identifications); light blue = recognition rate = 80% (4 of 5 identifications).
[M-02-03-T03]

Table 1. Recognisable German minimal pairs produced at intended fo of 784 and 880 Hz: Vowel recognition results of experiment 1. [M-02-03-T01]

Minimal pairs		Vowel contrast (VC) recognition							
Words	VC/L	Word condition				Sound nucleus condition			
		fo=784Hz		fo=880Hz		fo=784Hz		fo=880Hz	
		A'	%	A'	%	A'	%	A'	%
Biene–Bühne	/i/–/y/ ↗	0.98	97	0.99	100	0.95	90	0.95	90
siegen–Segen	/i/–/e/ ↗	0.95	90	0.99	100	0.93	88	0.98	97
biegen–Bögen	/i/–/ø/ ↗	0.99	100	0.99	100	0.98	97	0.95	90
Schielen–schälen	/i/–/ɛ/ ↗	0.99	100	0.99	100	0.97	95	0.89	80
siegen–sagen	/i/–/a/ ↗	0.99	100	0.99	100	0.99	100	0.98	97
lügen–legen	/y/–/e/ ↗	0.99	100	0.99	100	0.9	82	0.95	90
rühren–Röhren	/y/–/ø/ ↗	0.92	85	0.98	97	0.87	78	0.93	88
schürfen–schärfen	/y/–/ɛ/ ↗	0.97	95	0.99	100	0.96	93	0.91	85
Sühne–Sahne	/y/–/a/ ↗	0.99	100	0.98	97	0.99	100	0.93	88
Lehne–Löhne	/e/–/ø/ ↗	0.98	97	0.93	88	0.84	72	0.9	82
legen–lägen	/e/–/ɛ/ ↗	0.99	100	0.91	82	0.95	90	0.83	72
segen–sagen	/e/–/a/ ↗	0.99	100	0.99	100	0.99	100	0.96	93
töte–täte	/ø/–/ɛ/ ↗	0.99	100	0.97	95	0.97	95	0.86	78
Söhne–Sahne	/ø/–/a/ ↗	0.97	95	0.98	97	0.96	93	0.9	82
schälen–Schalen	/ɛ/–/a/ ↗	0.98	97	0.99	100	0.96	93	0.75	63
Baden–Boden	/a/–/o/ ↗	0.98	97	0.99	100	0.98	97	0.96	93
Baden–Buden	/a/–/u/ ↗	0.99	100	0.99	100	0.98	97	0.98	97
Boden–Buden	/o/–/u/ ↗	0.98	97	0.98	97	0.92	85	0.89	80

Table 2. Recognisable German “lVgen” minimal pairs produced at intended fo of 784 and 880 Hz: Vowel recognition results of experiment 2. [M-02-03-T02]

Sounds		Vowel recognition (in %)			
MP	Vowel	Inexperienced listeners (test A)		Experienced listeners (test B)	
		fo=784Hz	fo=880Hz	fo=784Hz	fo=880Hz
l-V-gen	/i/	100	100	100	100
	/y/	93	100	100	100
	/e/	54	50	80	100
	/ø/	93	64	100	80
	/ɛ/	57	86	100	100
	/a/	93	96	100	100
	/o/	100	100	100	100
	SL			↗	↗

Table 3. Recognisable German “lVgen”, “bVden” and “schVf” minimal pairs produced at intended fo of 880 and 1047 Hz: Vowel recognition results of experiment 3. [M-02-03-T03]

Sounds		Vowel recognition (matrix and in %)			
MP	V	fo=880Hz		fo=1047Hz	
l-V-gen	/i/	i i i i i	100	i i i i i	100
	/y/	y y y y y	100	y y y y y	100
	/e/	e e e e e	100	e e e e ei	80
	/ø/	ø ø ø ø ø	100	ø ø ø ø ø	100
	/ɛ/	ɛ ɛ ɛ ɛ ɛ	100	ɛ ɛ ɛ ɛ ɛ	100
	/a/	a a a a a	100	a a a o o	60
	/o/	o o o o o	100	o o o o o	100
	/u/	u u u u u	100	u u u u u	100
	SL	↗		↗	
b-V-den	/a/	a a a a a	100	a a a a a	100
	/o/	o o o o o	100	o o o o uo	80
	/u/	u u u u u	100	u u u u u	100
	SL	↗		↗	
sch-V-f	/i/	i i i i i	100	i i i i i	100
	/a/	a a a a a	100	a a a a a	100
	/u/	u u u u u	100	u u u u u	100
	SL	↗		↗	

M2.4 Extensive Ranges of f_o Contours of Speech

Introduction

In the literature, no frame of reference is given as a standard range regarding f_o variation for recognisable speech that has to be adhered to for speech acoustics in general (see also the Introduction). However, most frequency ranges reported concern measurements of citation-form words (or comparable utterances, above all investigated in studies on formant statistics), read text, or so-called normal (sic!) or conversational speech. Thus, in most cases, the related frequency ranges are given for relaxed speech and the lower part of the actual vocal range of the speakers only. To give an example, in a comprehensive book on voice production and perception, Kreiman et al. (2011, pp. 58–60) indicate an average f_o level of c. 220 Hz for the speech of women and of c. 115 Hz for the speech of men, and typical f_o ranges of speech are given as 164–262 Hz for women and 82–164 Hz for men. Levels of f_o exceeding these ranges are then attributed to singing.

However, as Andreeva et al. (2015) state: “Level and span of fundamental frequency are key ingredients of pitch profiles that have been shown to be characteristic for specific linguistic communities ([…], for different dialects, […] for different languages, […] for bilingual speakers).” Accordingly, when investigating pitch level and pitch span differences for English, German, Bulgarian and Polish, they found an f_o range of c. 180–370 Hz for the speech of women and c. 90–200 Hz for the speech of men, with some language-specific differences within these ranges. (See also Traunmüller and Erikkson, 1995, who report that different mean values and standard deviations of f_o reported in the literature depend on language, type of text, type of discourse, and emotional state of the speaker. For a discussion of different f_o levels and ranges of language-related speech, see also Keating and Kuo, 2012. Finally, further note in this context gender-affirming voice adaptation; see e.g. Quinn et al., 2022.)

More extended levels and ranges of f_o are reported for specific types of speech, such as for highly emotional, loud or infant-directed speech or speech within the field of the performing arts. Concerning emotional speech, to give two examples for Standard German, Paeschke and Sendlmeier (2000) reported f_o ranges of approximately ten semitones for neutral speech (terminology used by these authors) and 20 semitones for emotional speech (boredom, sadness, happiness, anger, fear). However, Probst and Braun (2019) reported a much more reduced f_o range of approximately 11 semitones for emotional speech (fear, disgust, joy, sadness, hot anger, cold anger). Note in this context that, discussing a study of emotional expressions, Kreiman et al. (2011, pp. 320–324) refer to an f_o range of up to 400 Hz for a single female speaker, a more extended range than that given for normal speech. Comparably, Signorello et al. (2020) reported an f_o variation of up to 21 semitones in charismatic speech.

Concerning f_o variation as a function of distance and vocal effort, Traunmüller and Eriksson (2000) investigated the f_o ranges for a single spoken sentence produced by speakers addressing a person positioned at five different distances ranging from 0.3 m to 187.5 m. According to their results, f_o varied from 213 (± 24) Hz to 423 (± 235, probably an erroneous indication) Hz for women (mean values and standard deviations) and from 110 (± 17) Hz to 274 (± 28) Hz for men. For children, f_o levels of up to 532 (± 48) Hz were also found. (For similar upper f_o frequencies for utterances of women, see also Meyer et al., 2018. Note also in this context increased f_o for Lombard speech, that is speech modification due to ambient noise; for an overview of the phenomenon, see Brumm and Zollinger, 2011.)

Concerning f_o variation as an aspect of infant-directed speech, Fernald et al. (1989) investigated the prosodic features of parental speech in different languages (French, Italian, German, Japanese, British English and American English). According to their results, women's infant-directed speech can manifest f_o levels of up to c. 450 Hz. Grieser and Kuhl (1988) investigated Mandarin-speaking mothers and found corresponding values for f_o ranges, except for one woman for whom an f_o range of c. 150–780 Hz was found.

Finally, to give a last example, Melton et al. (2020) investigated actresses and actors performing classical material without electronic amplification in outdoor spaces. Independent of gender, they found extended frequency ranges of c. 75–530 Hz.

These examples from the literature illustrate that most scholars consider vowel sounds produced at f_o levels in the lower part of the vocal range of speakers as corresponding to sounds and levels of so-called normal or conversational speech, vowel sounds produced at middle f_o levels up to c. 500 Hz as an aspect related to very specific speech characteristics, and vowel sounds produced at higher f_o levels as an aspect of singing. However, according to our estimate, there is no difference between the f_o ranges of recognisable speech and recognisable singing – above all not concerning the related upper f_o limit –, the distinction of normal or conversational versus highly specific speech is problematic, and the f_o ranges of speech as such, given in the literature, are in most cases too limited and too low. We had already come to this conclusion at the very beginning of our studies on the acoustic characteristics of vowel sounds because of both intellectual reasoning and observations of speech in different contexts of everyday life and, most importantly, of speech in the field of the performing arts and film. With regard to intellectual reasoning, many adult speakers have a vocal range of two octaves, and some speakers with great vocal abilities have a vocal range of up to three octaves or even more. To give an example, investigating voice range profiles of women and men, Sanchez et al. (2014) reported minimum and maximum f_o levels (mean and SD) of 118 (± 21) Hz to 1275 (± 269) Hz for vocally untrained women and 75 (± 11) Hz to 773 (± 186) Hz for vocally untrained men. To give a further example for an age-specific voice range profile, Andersen et al. (2021) reported minimum and maximum f_o levels (mean and SD) of 144 (± 22) Hz to 1064 (± 160) Hz for vocally untrained women aged 18 to 28. (Note that even wider ranges are given by Pabon and Ternström, 2020, including profiles for female and male singing students.) Why, then, should intelligible speech not cover most of these ranges? With regard to observations in different contexts of everyday life, many speakers have an f_o range exceeding one octave which is related neither to specific emotions (in the way in which emotions are differentiated in the literature) nor to loud speech. Also, a one-octave difference in the average f_o of speech can sometimes be observed for different female speakers, as is true for different male speakers. Further, the habitual speech of many speakers includes register changes, and it manifests correspondingly large f_o ranges. (Note in this context the indications for registers and f_o ranges given by Wolfe et al., 2020; Lee et al., 2021). Finally, with regard to observations in the field of the performing arts and film, extended f_o ranges of speech – often also including register changes – are a general phenomenon.

Because of the considerable effort of creating the vowel sound database of the Zurich Corpus, we had no resources to create a large sample of speech of different speakers, different contexts and different production styles and modes, the sample also having a systematic structure. However, in order to provide some evidence for our understanding and estimate, we have presented various examples of large f_o contours of intelligible speech in the Preliminaries, which point to the actual frequency extension of the contours that can be observed for everyday speech and for the speech of performing artists (see Preliminaries, Chapter M8.2). Yet, open-access sound playback was not provided because of potential legal issues. Therefore, for the present treatise, the documentation was revised and further extended, including new examples of speech extracts with legal permits or consent for sound playback for all utterances. (Note that a part of this renewed documentation was conducted in the context of an Interspeech Show and Tell conference presentation, Maurer et al., 2019, which is transferred to and integrated into the present treatise.)

Experiment

Creation of a sample of speech extracts: On the basis of the sounds and speakers documented in Part 2 of the Zurich Corpus (see Chapter 1.1), one or several speech extracts per speaker were selected (different languages), focussing on speech contours with upper f_0 levels of 500 Hz and higher for women and 350 Hz and higher for men, respectively. (Note that if these two frequency limits are surpassed by f_0, the statistical F_1 for sounds of close vowels and subsequently also for close-mid vowels is also surpassed and F-pattern estimation loses methodological substantiation, although speech intelligibility is maintained). However, some compilations of speech extracts of single speakers also include utterances produced at lower f_0 levels so as to demonstrate both the upper f_0 levels and the vocal range of the speaker in question. Extracts of nonstyle speech were separated from extracts of speech produced in an artistic style. However, the main focus of the investigation concerned artists performing on stage, acting in film or doing voice-over work. As explained in the Preliminaries (see pp. 5–6 and 91), we consider performing artists' speech to be the most direct approach to understanding the possible variation of basic speech characteristics. The number and duration of the speech extracts per speaker varied. Also, for some speakers, a long extract was retained, and short sequences of this extract were created in addition to the presentation. Some extracts were retained as single compilations of several cuts to exclude interacting voices or focus on parts of speech with extensive f_0 variation. Approximate pitch ranges of the extracts of a speaker were estimated by the author. As a result, a sample of c. 1300 single speech extracts produced by c. 100 speakers was created.

Selection of speech extracts for presentation: On the basis of the above sample, for this treatise, 517 single speech extracts of 80 adult speakers (48 women and 32 men) illustrating f_0 ranges for nonstyle speech and the speech of professional performing artists were selected. Further examples will continuously be added in the Additions section of the Zurich Corpus.

The text intelligibility of the extracts presented was not tested explicitly. However, as this treatise is based on an open-access sound database with a playback feature, the intelligibility should prove to be self-evident.

Results

Table 1 in the chapter appendix lists the selected extracts of nonstyle speech and their respective (approximative) f_o ranges, including sound links. Tables 2 and 3 list the extracts and f_o ranges for the speech of performing artists. For all three tables, the production context is given in Column 8. For nonstyle speech, the entire f_o range documented covers a frequency range of c. 125–1000 Hz for women and 100–600 Hz for men. For the speech of performing artists, the entire f_o range documented covers a frequency range of c. 110–1000 Hz for women and c. 90–850 Hz for men.

Discussion

Although lacking a systematic structure, the documentation nevertheless provides evidence that the f_o range of intelligible speech corresponds to the f_o range of recognisable vowel sounds, as discussed in the previous chapters. The documentation exemplifies anew that an acoustic theory of speech sounds has to account for an f_o range of intelligible speech of up to 800 Hz at a minimum. The investigation of isolated vowel sounds has to be evaluated from this perspective: Extensive f_o (or pitch) variation is not a side but a core phenomenon of speech and, thus, of the vowel sound. That is why this treatise introduces its course of argument with the documentation and discussion of vocalises, of recognisable vowel sounds at high f_o levels above statistically given levels of F_1 for almost all vowels, and of speech examples with f_o contours up to these high f_o levels, all these topics forming an ensemble of phenomena to which an acoustic theory of the vowel has to relate as a general reference and premise.

With regard to future research on speech acoustics, there is a need for large-scale sound corpora that include speech produced with extensive variation of basic production parameters documenting both everyday speech and the speech of performing artists. Such corpora are necessary in order to create a frame of reference for the investigation of the acoustic characteristics of speech in general and of vowel sounds in particular.

Chapter appendix

Regarding the speech extracts presented in the tables below, please note: The graphic illustration of the f_o contour does not always correspond to the actual f_o contour of the speech extract because of interfering noise or measurement problems. Please refer to the indications of perceived pitch ranges (approximations; author's estimate) given in the tables. References (origins of the speech extracts) are given online in the comment field of the sounds presented.

Table 1. Extracts of intelligible nonstyle speech demonstrating extensive ranges of f_o variation. Columns 1–4 = speech extracts (SG = speaker group; S/L = sound series and sound links; SubG = speaker subgroup (for details, see the Introduction); SP = speaker ID in the Zurich Corpus. Columns 5 and 6 = range of f_o documented (speech = frequency range for intelligible speech, in Hz; sounds = additional upper f_o frequency limit for high-pitched exclamations, in Hz). Column 7 = production context.
[M-02-04-T01]

Table 2. Extracts of intelligible speech produced by female performing artists demonstrating extensive ranges of f_o variation. Columns, see Table 1. Note CO = Chinese Opera actresses and singers.
[M-02-04-T02]

Table 3. Extracts of intelligible speech produced by male performing artists demonstrating extensive ranges of f_o variation. Columns, see Table 1.
[M-02-04-T03]

Table 1. Extracts of intelligible nonstyle speech demonstrating extensive ranges of fo variation. [M-02-04-T01]

Speakers and speech extracts				fo range (Hz)		Production context
SG	S/L	SubG	SP	Speech	Sounds	
woman	1 ↗	N	2172	220–800	1000	Market
	2 ↗	N	2220	200–500		TV coverage
	3 ↗	N	2430	200–500		TV coverage
	4 ↗	N	2458	200–500	650	TV coverage
	5 ↗	N	2477	200–550		TV coverage
	6 ↗	N	2479	220–600		TV coverage
	7 ↗	N	2480	200–600		TV coverage
	8 ↗	N	2506	200–650		TV coverage
	9 ↗	N	2412	200–800		Political speech
	10 ↗	N	2429	300–600		Political speech
	11 ↗	ST	2336	300–650	1000	TV show
	12 ↗	N	2348	200-700		TV show
	13 ↗	N	2300	200–650	750	TV show
	14 ↗	ST	2421	200–800		TV show
	15 ↗	N	2505	125–530		TV show
	16 ↗	N	2379	200–1000		Infant-directed speech
men	17 ↗	N	2420	170–600	700	Preaching
	18 ↗	N	2428	200–450		Political speech
	19 ↗	N	2433	130–440		Political speech
	20 ↗	N	2495	100–600		Political speeches
	21 ↗	N	2238	150–400		TV show
	22 ↗	N	2498	130–600		TV show
	23 ↗	N	2439	110–440		Sports reporting
	24 ↗	N	2456	150–400		Sports reporting
	25 ↗	N	2380	150–600		Infant-directed speech

Table 2. Extracts of intelligible speech produced by female performing artists demonstrating extensive ranges of fo variation. [M-02-04-T02]

Speakers and speech extracts				fo range (Hz)		Production context
SG	S/L	SubG	SP	Speech	Sounds	
women	1 ↗	CS	1001	170–650		Stage Performance
	2 ↗	EC	1005	250–800		Stage Performance
	3 ↗	ST	1052	350–1000		Stage Performance
	4 ↗	CO	1300	300–850		Stage Performance
	5 ↗	ST	2177	130–780		Stage Performance
	6 ↗	ST	2178	200–850		Stage Performance
	7 ↗	ST	2204	200–800	1100	Stage Performance
	8 ↗	ST	2212	300–660		Stage Performance
	9 ↗	ST	2234	200–850		Stage Performance
	10 ↗	ST	2251	300–800		Stage Performance
	11 ↗	ST	2275	150–600	850	Stage Performance
	12 ↗	ST	2276	300–800		Stage Performance
	13 ↗	CO	2284	200–850		Stage Performance
	14 ↗	ST	2305	400–800		Stage Performance
	15 ↗	ST	2350	200–900	1050	Stage Performance
	16 ↗	ST	2410	200–700	850	Stage Performance
	17 ↗	ST	2470	180–800	1050	Stage Performance
	18 ↗	ST	2277	200–700		Film
	19 ↗	ST	2304	170–1000		Film
	20 ↗	ST	2413	150–600		Film
	21 ↗	ST	2435	300–700		Film
	22 ↗	ST	2175	260–800	1000	Film
	23 ↗	ST	2196	300–880		Film (voice over)
	24 ↗	ST	2223	200–800		Film (voice over)
	25 ↗	ST	2258	200–700		Film (voice over)
	26 ↗	ST	2285	250–950		Film (voice over)
	27 ↗	ST	2298	300–1000		Film (voice over)
	28 ↗	ST	2419	200–750		Film (voice over)
	29 ↗	ST	2446	200–600	800	Film (voice over)
	30 ↗	ST	2447	270–600	800	Film (voice over)
	31 ↗	ST	2216	110–880		Radio broadcast

Table 3. Extracts of intelligible speech produced by male performing artists demonstrating extensive ranges of fo variation. [M-02-04-T03]

Speakers and speech extracts				fo range (Hz)		Production context
SG	S/L	SubG	SP	Speech	Sounds	
men	1 ↗	ST	1003	250–550		Stage Performance
	2 ↗	EC	1007	250–550	600	Stage Performance
	3 ↗	ST	2163	100–580		Stage Performance
	4 ↗	ST	2194	130–700	800	Stage Performance
	5 ↗	ST	2214	250–700		Stage Performance
	6 ↗	ST	2225	200-780		Stage Performance
	7 ↗	CS	2274	110–580		Studio recording
	8 ↗	ST	2297	150–600		Stage Performance
	9 ↗	ST	2351	150–550	800	Stage Performance
	10 ↗	ST	2388	400–750		Stage Performance
	11 ↗	ST	2411	200–800		Stage Performance
	12 ↗	ST	2432	150–500		Stage Performance
	13 ↗	ST	2465	100–600		Stage Performance
	14 ↗	ST	2468	150–700		Stage Performance
	15 ↗	ST	2494	90–550		Stage Performance
	16 ↗	ST	2282	110–580		Film
	17 ↗	ST	2453	130–650		Film
	18 ↗	ST	2169	100–650		Film (voice over)
	19 ↗	ST	2294	200–800	950	Film (voice over)
	20 ↗	ST	2394	300–600		Film (voice over)
	21 ↗	ST	2422	150–850	1000	Film (voice over)
	22 ↗	ST	2501	400–780		Film (voice over)
	23 ↗	ST	2295	160–500		TV show (performance)
	24 ↗	ST	2484	110–650	800	TV show (performance)

M3 Ambiguity of Spectral Peaks, Estimated Formant Patterns and Spectral Shapes

M3.1 Source–Filter Synthesis Based on Statistical Formant Frequencies, Including Variation of f_o

Introduction

One way to interpret estimated statistical *F*-patterns of natural vowel sounds and evaluate their role in vowel quality recognition is to use a source–filter synthesiser to reproduce sounds based on these *F*-patterns and then test the recognition thereof.

Hillenbrand and Gayvert (1993) investigated vowel recognition for synthesised vowel sounds (Klatt synthesis) based on *F*-patterns reported by Peterson and Barney (1952) for natural hVd utterances, both *F*-patterns and related f_o applied being speaker- and speaker group-specific. The study showed a substantial decrease in vowel recognition for the synthesised sounds compared to the natural sounds: The average vowel recognition rate of 94.4% for natural sounds dropped down to 74.8–72.7% for the synthesised sounds (depending on the use of f_o with a falling or a flat contour in synthesis). However, pronounced recognition rate differences that related to vowel quality were also found, with a maximum rate of 95.4–96.2% for sounds of /i/ (f_o contour and flat f_o, respectively) and a minimum rate of 51–55% for sounds of /ɑ/ (f_o contour and flat f_o, respectively).

Hillenbrand et al. (1995) conducted a formant measurement study comparable to the study of Peterson and Barney (1952) but on a new methodological basis and taking into account duration and spectral change. (Note that in the Peterson and Barney study, formant frequencies and amplitudes were measured by “estimating a weighted average of the frequencies of the principal components” of a vowel spectrum; see Potter and Steinberg, 1950, for details of this procedure. In the study of Hillenbrand et al., 1995, formant frequencies were measured using LPC analysis and manual editing of the resulting formant tracks.) Subsequently, based on a 300-utterance subset of this new sound database used for the statistical analysis of *F*-patterns, Hillenbrand and Nearey (1999) investigated vowel recognition by comparing the natural /hVd/ sounds with two different types of synthesised /hVd/ versions (Klatt synthesis), one type of synthesis using the originally measured formant contours, the other type of synthesis related to static (averaged) *F*-patterns. The f_o levels of synthesis corresponded to the f_o

contours of the original sounds for both synthesis types. (Note that the utterance subset included tokens of men, women and children with an even distribution of 30–36–34%. Roughly speaking and corresponding to the musical C-major scale, the corresponding average *f*o levels of the synthesised sounds were approximately 131 Hz for men, 220 Hz for women, and 262 Hz for children.) This study again showed a substantial decrease in the average vowel recognition rate for synthesised sounds compared with natural sounds. However, the decrease also depended on formant contours: The vowel recognition rate of 95.4% for the natural sounds dropped to 88.5% for the synthesised sounds with formant contours and to 73.8% for the synthesised sounds with static formants. However, as was the case for the previous study, recognition rates again proved to be vowel quality-related, with maximum rates of 91.6–89.6% for the sounds of /i/, 96.4–88.6% for the sounds of /ɪ/ and 94.2–96.0% for the sounds of /ɔ/ (formant contours and fixed formants) and minimum rates of 80.8–49.4% for the sounds of /æ/ and 70.4–61.0% for the sounds of /u/.

Besides the question of the importance of static or dynamic properties of the vowel sound for vowel quality recognition (here above all in terms of static or dynamic *F*-patterns and *f*o contours) and disregarding the vowel quality-related differences, the two synthesis studies summarised above indicate that vowel sounds synthesised based on estimated *F*-patterns may be recognisable with a rate above 70%. Therefore, the authors of the two synthesis studies cited concluded that static *F*-patterns represent the primary acoustic characteristics of vowel quality and that the dynamic properties of sounds play a secondary – although quite important – role in vowel quality recognition.

However, this general recognition result was shown only for cases for which the *f*o of the natural and the synthesised sounds were similar, because different *f*o levels for synthesis related to single unchanged *F*-patterns were not investigated. Therefore, the general question arises as to whether the recognised vowel quality is maintained or shifts to another quality if, for a given *F*-pattern of a vowel, *f*o is substantially varied in vowel synthesis. (Note that this question has already been formulated in the early study of Potter and Steinberg, 1950; see also Miller, 1953.)

The question of the effect of *f*o on vowel quality in vowel synthesis when an *F*-pattern is kept unchanged is a matter of debate (for overviews and discussions, see Fujisaki and Kawashima, 1968; Traunmüller, 1981, 1988; Hirahara and Kato, 1992; Maurer and Landis, 1995, 1996; de Cheveigné and Kawahara, 1999; Ménard et al., 2002; Assmann and

Nearey, 2008; Barreda and Nearey, 2012). While some scholars have concluded that f_o has only a marginal or very limited effect on vowel quality and have attributed this effect mainly to differences in age/size/gender of the speakers, other scholars have concluded that it indeed has a substantial impact on vowel recognition, above all referring to studies that take into account f_o variation exceeding 300 Hz.

When we began exploring vowel (re-)synthesis within different experimental approaches, in our first attempts, we experienced that a substantial variation of f_o in (re-)synthesis based on a given unchanged *F*-pattern often resulted in a shift of recognised vowel quality. However, because our work has also revealed many aspects of a nonuniform relation between vowel recognition and the vowel spectrum – which will be exposed extensively in this treatise – we also expected that f_o-related vowel quality shifts might depend on f_o range, f_o difference of sounds compared and the vowel qualities in question. (For additional aspects we observed later, see Chapter M7.)

Against this background, three vowel recognition experiments for synthesised sounds related to statistical *F*-patterns reported by Maurer et al. (1991, hereafter referred to as MA), Pätzold and Simpson (1997, hereafter referred to as PS) and Fant (1959, hereafter referred to as FA) were conducted. The *F*-patterns given by MA were selected because they report values for men, women and children, including different and similar f_o levels for the sounds produced by the speakers of the three different speaker groups; however, values are given for the long Swiss German vowels /i, e, a, o, u/ only. The *F*-patterns given by PS were selected because they report values for men and women of the eight vowels /i, y, e, ø, ɛ, a, o, u/ of Standard German. (Note that, in this study, formant measurement was conducted without editing calculated formant tracks, and the vowel /ɛ/ was investigated as a short vowel.) Finally, the *F*-patterns given by FA were selected because they represent a historical basis within the context of the formulation of the source–filter theory of speech production (Fant, 1970) and, at the same time, the vowels /i, y, e, ø, ɛ, a, o, u/ of Swedish selected for the present study are comparable to variants of long Standard German vowels.

Experiment 1

***F*-patterns investigated:** MA reported formant frequency values for sounds of the long Swiss German vowels /i, e, a, o, u/ (region of the canton of Zurich) produced by seven men, seven women and seven children. The vowel sounds were produced in isolation (V context) at different f_o levels of approximately 131–220–262 Hz (men), 220–262 Hz (women) and 262 Hz (children), according to the musical C-major scale and further adapted to the three synthesis experiments reported here (average f_o levels given by MA were 110–170–220–270 Hz, see Table 1; sounds produced at f_o of 170 Hz were disregarded here). For each speaker group and each f_o level of the sounds, estimated average F_1–F_2 for /a–o–u/ and average F_1–F_2–F_3 for /e–i/ were reported. For the present synthesis experiment, higher formants were added: Concerning the vowels /a–o–u/, according to the speaker groups, F_3 was set to 2600 Hz for men, 2800 Hz for women and 3000 Hz for children in terms of approximations to corresponding values given in PS, FA and also Hillenbrand et al. (1995). For all vowels, F_4 was set to 3500 Hz for men, 4200 Hz for women and 4400 Hz for children in terms of approximations to corresponding values given in Syrdal (1985) and Hillenbrand et al. (1995). F_5 was set to 4500 Hz for men, 5400 Hz for women, and 5700 Hz for children. The F_5 values for men and women were set according to Rabiner (1968), and the F_5 value for children was adapted corresponding to the children's default F_4. Formant bandwidths were set to default values of 90–110–170–400–500 Hz, according to Hillenbrand and Nearey (1999). The resulting F-patterns (limited to F_1–F_2–F_3) are given in Table 1 in the chapter appendix.

Sound synthesis: 1 sec. steady-state vowel sounds (no f_o contour) with a fade in/fade out of 0.05 sec. were produced based on the above F-patterns using a Klatt synthesiser in cascade mode (synthesiser used as implemented in Praat; sampling frequency = 44.1 kHz, resolution = 16 bit). For every single F-pattern, seven sounds were synthesised on seven different f_o levels of 65–131–220–262–330–440–523 Hz. The frequency range of 131–523 Hz covered both the f_o levels of recognisable vowel sounds and the statistical F_1 levels of sounds of close and close-mid vowels; the f_o level of 65 Hz was added in order to imitate creaky phonation. As a result, a sample of 210 synthesised sounds was created.

Listening test: The vowel quality of the synthesised sounds was tested in a listening test according to the standard procedure of the Zurich Corpus and involving the five standard listeners. The listening test was divided into seven subtests, one subtest for each of the seven f_o levels investigated.

Results of experiment 1

The listening test results are shown in Table 2a in the chapter appendix. Considering only those vowel sounds for which a labelling majority for a given vowel quality was indicated by the listeners (recognition rate ≥ 60%), the below vowel quality matches or shifts were found. Note that the direction of the recognised vowel quality shifts in the following discussion of results is always given for a stepwise increase of f_o from a lower to a higher frequency level of comparison.

Vowel recognition for *F*-pattern-related statistical f_o: For all sounds synthesised at an f_o level that corresponded to the average statistical f_o level of the natural sounds and their estimated *F*-patterns (see Table 2a, results marked in green), the recognised vowel quality matched with vowel intention of the natural sounds.

Vowel recognition for other f_o levels not related to statistical f_o: Since, for the other synthesised sounds, shifts caused by an increase or a decrease of f_o were most prominent in the sound series related to *F*-patterns of close-mid vowels, they will be discussed here first, followed by results obtained for close and open vowels.

For a single *F*-pattern of the close-mid front vowel /e/ and synthesis levels of f_o that are substantially below the *F*-pattern-related statistical f_o in question, a shift in vowel quality was found in five out of six sound series. Adopting a general perspective of a stepwise increase of f_o from a lower to a higher level, the f_o increase resulted in shifts in an open–close direction. The same shift direction was found for the remaining sound series with an f_o increase to 330 Hz. Thus, for a single *F*-pattern of the close-mid vowel /e/ and an f_o range of 65–330 Hz, as a general tendency, substantially increasing f_o from a lower to a higher level caused vowel quality shifts in an open–close direction (see Table 2a, results marked in red). However, further increasing f_o and thereby substantially surpassing the levels of statistical average F_1 (and applied F_1) of the close vowels (f_o > 330 Hz) and of statistical average F_1 of some of the close-mid vowels (f_o > 440 Hz), the shift direction was "reverted" in terms of a close–open shift subsequent to an open–close shift in two of the six sound series (see Table 2a, results marked in purple) while for another two series, again, a further open–close shift was observed. For a single *F*-pattern of the close-mid back vowel /o/ and the entire range of f_o variation of 65–523 Hz in synthesis, a substantial increase of f_o from a lower to a higher level resulted in an open–close shift in vowel quality for all sound series (see Table 2a, results marked in red). Particular attention should be given to the occurrence of numerous vowel quality shifts that included adjacent and non-adjacent qualities, that is /ɛ–e–y-i/ or /ɔ–o–u/, respectively.

For a single *F*-pattern of the close front vowel /i/ and a range of f_o variation of 65–330 Hz in synthesis, a substantial increase of f_o from a lower to a higher level caused an open–close shift in two of the six sound series; further increasing f_o to levels above 330 Hz and thereby substantially surpassing the levels of statistical average F_1 (and applied F_1) of the close vowels ($f_o > 330$ Hz) resulted in an inverted or reverted close–open shift for all six sound series. For a single *F*-pattern of the close back vowel /u/ and an f_o range of 65–330 Hz in synthesis, a substantial increase of f_o from a lower to a higher level did not cause any shift in vowel quality; a further f_o increase resulted in a close–open shift in only one case of a single sound related to an *F*-pattern of men.

For all *F*-patterns of /a/, f_o variation in synthesis had no marked effect on vowel recognition.

Experiment 2

***F*-patterns investigated:** PS reported formant frequency values for long and short Standard German vowels produced by 12 men and 12 women in the context of read sentences. For these vowels and speaker groups, estimated average F_1–F_2–F_3 patterns were reported (see Table 1 in the chapter appendix). However, no f_o levels related to measured *F*-patterns and speaker groups were given. For the present experiment, the values for /i–y–e–ø–ɛ–a–o–u/ were selected, and assumed statistical f_o levels were set to 131 Hz for the *F*-patterns of men and 220 Hz for the patterns of women. Values for F_4, F_5 and the formant bandwidths were set as in experiment 1.

Sound synthesis and listening test: The conditions for the sound synthesis (112 synthesised sounds in total) and the listening test corresponded to those of experiment 1.

Results of experiment 2

The listening test results are shown in Table 2b in the chapter appendix. Considering again only those vowel sounds for which a labelling majority for a given vowel quality was obtained (recognition rate ≥ 60%), the below vowel quality matches or shifts were found. (Note again that the direction of the recognised vowel quality shift is always given for a stepwise increase of f_o from a lower to a higher frequency level of comparison.)

Vowel recognition for f_o levels related to estimated *F*-patterns: For seven of 16 sounds synthesised at an f_o level that corresponded to the average statistical f_o level of the natural sounds and their estimated *F*-patterns (see Table 2b, results marked in green), the recognised vowel quality did not match with the vowel intention of the natural sounds: At these f_o levels of synthesis, the sounds related to the *F*-patterns of the vowels /i, y/ were either recognised as /y, ø/ or not recognised clearly (no labelling majority), the sound related to the *F*-pattern of /u/ of women was not recognised clearly, the sound related to the *F*-pattern of /e/ of men was recognised as /ø/ and the sound related to the *F*-pattern of /ɛ/ of men was not recognised clearly (see indications marked with "*").

Vowel recognition for other f_o levels not related to statistical f_o: Besides the above mismatches of vowel intention and recognition, the results of the second synthesis experiment were somewhat in line with the results of the first experiment, above all concerning both occurring vowel quality shifts in an open–close direction related to an increase of f_o in synthesis up to a level of 330 Hz and some inverted or reverted shifts in a close–open direction for synthesis levels of f_o above 330 Hz. (Note also the open-mid–close shift for the sounds of /ɛ/ of the women.)

Experiment 3

***F*-patterns investigated:** FA reported formant frequency values for sounds of Swedish vowels produced by seven men and seven women at f_o levels of approximately 131 Hz (men) and 220 Hz (women) in the context of keywords with prolonged vowel duration. For these vowels and speaker groups, estimated average F_1–F_2–F_3–F_4 were reported (see Table 1 in the chapter appendix). However, for the sounds of /e–ɛ–a/ of women, the F_4 measurements seemed to be problematic since no values are given. For the present experiment, the values for /i–y–e–ø–ɛ–a–o–u/ were selected. Missing values for F_4 and values for F_5 and the formant bandwidths were set as in experiment 1.

Sound synthesis and listening test: Sound synthesis (112 synthesised sounds in total) and listening test conditions corresponded to those of experiment 1.

Results of experiment 3

The listening test results are shown in Table 2c in the chapter appendix. For the vowel sounds with a labelling majority for a given vowel quality (recognition rate ≥ 60%), the below vowel quality matches or shifts were found.

Vowel recognition for f_o levels related to estimated *F*-patterns: Concerning the sounds synthesised at an f_o level that corresponded to the average statistical f_o level of the natural sounds and their estimated *F*-patterns (see Table 2c, results marked in green), the sounds related to the *F*-patterns of /y, ø/ of women were recognised as /i, e/, and the sounds related to both *F*-patterns of /ɛ/ of men and women were recognised as /ø/ or as a quality in between /ɛ/ and /e/ (see indications marked with "*"). For the remaining vowel qualities, vowel recognition accorded to vowel intention.

Vowel recognition for other f_o levels not related to statistical f_o: Besides the above mismatches of vowel intention and recognition, the results of the third synthesis experiment were again in line with the results of the first two experiments, above all with regard to occurring vowel quality shifts in an open–close direction for increasing f_o levels in synthesis from 65 Hz up to 330 Hz for the *F*-patterns of close-mid vowels, and occurring inverted or reverted shifts in a close–open direction for synthesis f_o levels above 330 Hz for *F*-patterns of close and close-mid vowels and, in this study, also of /a/. Further, shifts in an open–close direction related to an increase of f_o in synthesis also occurred for one *F*-pattern of the open-mid vowel /ɛ/. (Note an open-mid–close shift for the sounds of /ɛ/ of women again.)

General discussion

In three experiments, the effect of f_o variation in a source–filter synthesis based on statistical *F*-patterns was investigated. Patterns and pattern-related f_o levels of speakers different in age and gender were included in the experiments. During synthesis, stepwise increasing f_o variation was set to 65–131–220–262–330–440–523 Hz for each single *F*-pattern. This range of f_o variation covered the range of statistical F_1 for close and close-mid vowels and included the f_o level of 65 Hz to imitate creaky phonation.

In sum, f_o variation was found to have two types of effects on vowel recognition: (i) If, for a given statistical *F*-pattern, f_o was increased incrementally from a lower level of 65 Hz to a higher level of 330 Hz in synthesis, this variation either had no effect on the recognised vowel quality or the quality shifted in an open–close direction (sometimes including an additional unrounded–rounded shift). Open–close shifts were pronounced above all for sounds of the close-mid vowels. (ii) If f_o was further increased and thereby substantially surpassed the levels of statistical F_1 of close vowels and, subsequently, also of some close-mid vowels (for the corresponding F_1 of the statistics investigated, see Table 1), in some cases, the open–close shifts newly occurred or continued, while in other cases, they were inverted or reverted to close–open shifts. For the rest of the cases, no shifts were found. Table 3 in the chapter appendix provides a compilation of exemplary sound series illustrating these findings.

In these terms, vowel recognition did not relate to statistical *F*-patterns independent of the f_o levels of the sounds. This main finding supports earlier claims that, for (re-)synthesised vowel sounds, f_o variation substantially affects vowel recognition if *F*-patterns are kept unchanged (see the chapter introduction). At the same time, in many cases, not only the spectral peak pattern but also the entire filter curve used for sound synthesis proved to be ambiguous in that a single peak pattern and filter curve represented different vowel qualities. Indeed, the finding of a direct impact of f_o on vowel recognition in vowel (re-)synthesis, keeping the filter unchanged, indicates that the entire spectral envelope of a vowel sound is an ambiguous representation of vowel quality.

Inverted or reverted shifts in a close–open direction for sounds synthesised at f_o of 440 and 523 Hz may have to be considered with regard to two different aspects: On the one hand, these shifts may have been related to the specific fine structure of the harmonic spectra of the sounds in question resulting from the relation between the LPC filter curve and the f_o level applied in synthesis (consider, above all, the frequency distance of the harmonics and the resulting sampling of the filter curve including the match or mismatch of harmonics and filter frequencies). But on the other hand, above all for sounds of close front vowels for which these inverted shifts mainly occurred in the present investigation, the change in the relation between lower and higher spectral energy may have been the cause of the inverted shifts. Notably, if sounds of close (and sometimes also of close-mid) vowels were HP filtered applying CFs below c. 1 kHz, which in its turn resulted in a change in the relation between the lower and higher spectral energy, close–open shifts were observed (see Chapter 8.2). As we will argue below, this supports the thesis of the vowel being a kind of foreground–background phenomenon.

For a Klatt synthesis including extensive f_o variation, special attention should be given to the following observation: If two sounds, S1 and S2, are synthesised based on an identical filter curve but at two very different f_o levels markedly below and above the first filter frequency of an applied *F*-pattern, the acoustic analysis thereof often does not allow for identification of the common filter curve for both sounds as used in sound production. This finding for vowel synthesis points to a possible dissociation of vowel production and vowel recognition: While the filter pattern used in synthesis can be assessed reasonably well in the acoustic analysis (LPC analysis) of the radiated sound S1, this may not be the case for the radiated sound S2, that is, the calculated filter pattern for the radiated sound S2 may differ substantially from the filter pattern of its production. It may be argued that calculating the LPC filter curve for sounds at higher f_o is not substantiated methodologically. However, as a counterargument, a resynthesis of S2 based on the calculated LPC filter curve and f_o for that radiated sound may, in many cases, produce a third sound very similar to S2, including its vowel quality. Such consideration leads to the expectation that for a given higher f_o level, vowel synthesis with two substantially different filter curves may result in sounds with very similar vowel qualities, a topic that is addressed directly in Chapter M9. (For an illustration of the sounds investigated here, see Table 3 in the chapter appendix, Series 10 and 11; for a re-examination of the synthesis, use the Klatt synthesiser in the Zurich Corpus.)

In this type of experimentation, spectral representation of vowel quality was again indicated to be nonuniform: The occurrence and the extent of the vowel quality shifts and their associated levels of f_o and ranges of f_o variation depended on the vowel qualities of the natural reference sounds, the different studies and their statistical *F*-patterns as well as on the speaker groups and the related statistical f_o. Therefore, most importantly, if a small set of statistical *F*-patterns related to only a few vowels is investigated in a synthesis experiment of this type, not including all long vowels of a given language, the obtained results will not predict results for the other vowel groups. However, even if all long vowels of a language with corresponding *F*-patterns and f_o levels are investigated, the corresponding results will not predict in detail the results of a similar experiment based on a different method of *F*-pattern estimation or a similar experiment investigating vowels of a different language. Furthermore, production parameter variation for natural reference sounds, such as e.g. phonation, vocal effort, extensive f_o variation, and different speaking and singing styles, can be expected to often result in different *F*-patterns for a given vowel, and this also has to be accounted for when investigating vowel synthesis of this type and interpreting results of vowel recognition.

However, some experimental limitations and relativisations of results also have to be considered. The listening test was performed by the same five listeners who performed the listening tests for the Zurich Corpus (which only concerns Standard German vowels), even though the investigated *F*-patterns concerned Swiss German, Standard German and Swedish. Also, the investigated *F*-patterns relate to studies with different methods of formant estimation. Both aspects may have influenced the vowel recognition results: Above all, concerning the FA study, the vowel recognition mismatches for synthesised sounds based on statistical *F*-patterns and related f_o values may be a result of the listeners not being native Swedish speakers, and concerning the PS study, the mismatches may be a result of automatic formant calculation with no spectral crosscheck of calculated formant tracks. (This may explain the differences in vowel recognition for the synthesised sounds at pattern-related f_o levels found between MA and PS, as well as the differences found within speaker group-related patterns of PS.) However, here, emphasis is given to the observation and discussion of vowel quality shifts as an effect of f_o variation, with no further investigation into the impact of language differences and differences in *F*-pattern estimation. From the perspective of this investigation, similar general tendencies of f_o-variation-related vowel quality shifts were found for *F*-patterns of all three studies.

Further, the sound and vowel quality of the samples produced by the Klatt synthesiser based on static *F*-patterns at various f_o levels was often poor. New synthesis tools are needed to improve the sound quality while, at the same time, relating the sounds to a single filter curve of production in order to demonstrate the ambiguity of spectral peaks and spectral envelopes concerning vowel recognition.

Chapter appendix

Table 1. Source–filter synthesis based on statistical *F*-patterns, including variation of f_o: *F*-patterns investigated. F_1–F_2–F_3 patterns are given for the MA and PS studies, and F_1–F_2–F_3–F_4 patterns are given for the FA study. For additional higher filter frequencies applied in synthesis, see the experiments sections. Column 1 = vowel quality (V) of the natural sounds for which *F*-patterns were estimated. Column 2 = speaker group (SG, where m = men, w = women and c = children). Columns 3 and 4 = statistical average f_o related to the *F*-patterns (where fo c = calculated values in Hz, and fo m = values in Hz according to musical C-major scale). Columns 5–7 or 5–8 = *F*-patterns, in Hz.
[M-03-01-T01]

Table 2. Source–filter synthesis based on statistical *F*-patterns, including variation of f_o: Synthesised sounds and vowel recognition results. Column 1 = *F*-pattern-related study. Columns 2–4 = see Table 1, Columns 1, 2 and 4. Columns 5–11 = vowel recognition results for the synthesised sounds with a recognition rate ≥ 60%, per f_o level applied in synthesis. Extended online table: Columns 12 ff. = details of the vowel recognition results (labelling of the five listeners, given in phonetic open–close order; note ns = no vowel specified, txt = free comment). Colour code: Green = synthesis with statistical f_o related to the *F*-patterns in question; no colour = synthesis at f_o not related to the *F*-pattern in question, but with recognised vowel quality which matches the quality that the statistical *F*-pattern was related to, or no labelling majority; red = synthesis at f_o not related to the *F*-pattern in question and associated with a vowel quality shift in an open–close direction with increasing f_o from lower to higher levels; purple = synthesis at f_o not related to the *F*-pattern in question and associated with a vowel quality shift in a close–open direction with increasing f_o from lower to higher levels.
[M-03-01-T02]

Table 3. Source–filter synthesis based on statistical *F*-patterns, including variation of f_o: Different types of effects of f_o variation on vowel recognition. Column 1 = series number and sound links (S/L). Columns 2–6 = reference *F*-pattern of sound synthesis (ST = *F*-pattern-related study; V = vowel quality the statistical *F*-pattern was related to; SG = speaker group; fo m = average statistical f_o related to the reference *F*-pattern the sound synthesis was based on, values in Hz according to C-major scale; N = number of sounds documented). Column 7 = synthesised sounds with statistical f_o level (S1; f_o, in Hz) and the results of vowel recognition (recognition rate ≥ 60%). Columns 8–10 = synthesised sounds with varied f_o levels (S2, S3, S4), f_o (in Hz) and results of vowel recognition (V). Column 11 = frequency range (FR) of f_o variation. Column 12 = illustrated aspects in the sound series.
[M-03-01-T03]

Table 1. Source–filter synthesis based on statistical F-patterns (in Hz), including variation of fo (in Hz): F-patterns investigated. [M03-01-T01]

Maurer et al. (1992; MA)						
V	**SG**	**fo c**	**fo m**	**F1**	**F2**	**F3**
i	m	120	131	247	2165	2953
	m	235	220	268	2292	2906
	m	274	262	288	2404	2941
	w	237	220	258	2435	3507
	w	277	262	287	2490	3355
	c	277	262	325	2868	3497
u	m	117	131	263	736	2600
	m	238	220	268	736	2600
	m	272	262	292	795	2600
	w	239	220	287	712	2800
	w	273	262	293	789	2800
	c	275	262	308	727	3000
e	m	119	131	335	2050	2633
	m	232	220	444	2090	2752
	m	271	262	468	2120	2797
	w	236	220	457	2310	2964
	w	271	262	504	2327	2889
	c	277	262	551	2569	3431
o	m	117	131	346	700	2600
	m	235	220	468	840	2600
	m	267	262	506	881	2600
	w	235	220	464	887	2800
	w	268	262	533	873	2800
	c	274	262	547	1098	3000
a	m	116	131	691	1099	2600
	m	234	220	739	1160	2600
	m	264	262	746	1184	2600
	w	236	220	738	1217	2800
	w	263	262	757	1214	2800
	c	268	262	797	1366	3000

Pätzold and Simpson (1997; PS)							
V	**SG**	**fo c**	**fo m**	**F1**	**F2**	**F3**	**F4**
i	m	–	131	290	1986	2493	–
	w	–	220	329	2316	2796	–
y	m	–	131	310	1505	2205	–
	w	–	220	342	1667	2585	–
u	m	–	131	309	961	2366	–
	w	–	220	350	1048	2760	–
e	m	–	131	372	1879	2486	–
	w	–	220	431	2241	2871	–
ø	m	–	131	375	1458	2220	–
	w	–	220	434	1646	2573	–
o	m	–	131	380	907	2415	–
	w	–	220	438	953	2835	–
ɛ	m	–	131	498	1639	2451	–
	w	–	220	592	1944	2867	–
a	m	–	131	639	1225	2477	–
	w	–	220	779	1347	2785	–

Fant (1959; FA)							
V	**SG**	**fo c**	**fo m**	**F1**	**F2**	**F3**	**F4**
i	m	128	131	256	2066	2960	3400
	w	218	220	278	2520	3450	3900
y	m	128	131	257	1928	2421	3300
	w	215	220	270	2480	2920	3575
u	m	127	131	307	730	2230	3300
	w	222	220	340	690	2900	4000
e	m	124	131	334	2050	2510	3400
	w	215	220	365	2540	2950	4200
ø	m	126	131	363	1690	2200	3390
	w	215	220	372	2000	2610	3650
o	m	132	131	402	708	2460	3150
	w	223	220	433	815	2840	3600
ɛ	m	125	131	438	1795	2385	3415
	w	214	220	545	2140	2860	4200
a	m	124	131	680	1070	2520	3500
	w	215	220	860	1195	2830	4200

Table 2. Source–filter synthesis based on statistical F-patterns, including variation of fo (in Hz): Synthesised sounds and vowel recognition results. [M03-01-T02] Extended online table: ☐↗

Sounds				Vowel recognition per fo (in Hz)							
ST	V	SG	fo m	65	131	220	262	330		440	523
2a: Experiment 1, MA study	i	m	131	i	**i**	i	i	i	Statistical F1 limit (sounds of close vowels)	e	–
			220	i	i	**i**	i	i		e	e
			262	e	i	i	**i**	i		e	e
		w	220	i	i	**i**	i	i		e	e
			262	i	i	i	**i**	i		e	e
		c	262	e	e	i	**i**	i		i	e
	u	m	131	u	**u**	u	u	u		u	u
			220	u	u	**u**	u	u		u	u
			262	u	u	u	**u**	u		o	u
		w	220	u	u	**u**	u	u		u	u
			262	u	u	u	**u**	u		u	u
		c	262	u	u	u	**u**	u		u	u
	e	m	131	e	**e**	–	–	y		e	–
			220	ɛ	ɛ	**e**	e	e		–	–
			262	ɛ	ɛ	e	**e**	e		–	y
		w	220	ɛ	ɛ	**e**	e	e		i	e
			262	ɛ	ɛ	e	**e**	e		e	–
		c	262	ɛ	ɛ	–	**e**	e		e	i
	o	m	131	–	**o**	o	u	u		u	u
			220	ɔ	ɔ	**o**	o	o		u	u
			262	ɔ	ɔ	o	**o**	o		–	u
		w	220	ɔ	ɔ	**o**	o	o		u	u
			262	ɔ	ɔ	o	**o**	o		–	u
		c	262	ɔ	ɔ	ɔ	**o**	o		u	u
	a	m	131	a	**a**	a	–	o		a	–
			220	a	a	**a**	a	–		a	a
			262	a	a	a	**a**	a		a	a
		w	220	a	a	**a**	a	–		a	a
			262	a	a	a	**a**	a		a	–
		c	262	a	a	a	**a**	a		a	a

Table 2 (continuation). [M03-01-T02]

Sounds				Vowel recognition per fo (in Hz)							
ST	V	SG	fo m	65	131	220	262	330		440	523
2b: Experiment 2, PS study	i	m	131	y	**y***	y	y	y	Statistical F1 limit (sounds of close vowels)	–	–
		w	220	e	e	**–***	–	i		e	e
	y	m	131	ø	**ø***	y	y	y		ø	ø
		w	220	ø	ø	**ø***	ø	y		ø	ø
	u	m	131	u	**u**	u	u	u		–	u
		w	220	–	u	**–***	u	u		u	o
	e	m	131	ø	**ø***	ø	ø	y		y	–
		w	220	ɛ	e	**e**	e	e		–	–
	ø	m	131	ø	**ø**	ø	ø	øy		ø	ø
		w	220	ø	ø	**ø**	ø	ø		y	ø
	o	m	131	o	**o**	o	u	u		o	u
		w	220	o	o	**o**	o	u		u	u
	ɛ	m	131	–	**–***	ø	ø	ø		ø	ø
		w	220	ɛ	ɛ	**ɛ**	–	–		–	y
	a	m	131	–	**a**	a	–	–		a	–
		w	220	a	a	**a**	a	a		a	a
2c: Experiment 3, FA study	i	m	131	i	**i**	i	i	i	Statistical F1 limit (sounds of close vowels)	e	–
		w	220	i	i	**i**	i	i		e	e
	y	m	131	y	**y**	y	y	y		ø	–
		w	220	i	i	**i***	i	i		e	e
	u	m	131	u	**u**	u	u	u		u	u
		w	220	–	o	**u**	u	u		u	u
	e	m	131	e	**e**	y	y	y		e	–
		w	220	e	e	**e**	e	i		i	e
	ø	m	131	ø	**ø**	ø	ø	y		y	ø
		w	220	e	ø	**e***	–	y		e	–
	o	m	131	o	**o**	o	o	u		u	u
		w	220	o	o	**o**	o	–		u	u
	ɛ	m	131	ø	**ø***	ø	ø	ø		y	–
		w	220	ɛ	ɛ	**ɛe***	e	e		e	y
	a	m	131	a	**a**	–	aɔ	o		ɔ	–
		w	220	a	a	**a**	a	a		ɔ	a

Table 3. Source–filter synthesis based on statistical F-patterns, including variation of fo: Different types of effects of fo variation on vowel recognition. [M-03-01-T03]

S/L	Reference F-pattern					Synthesis and vowel recognition					Illustration
	ST	V	SG	fo m Hz	N	S1 Hz–V	S2 Hz–V	S3 Hz–V	S4 Hz–V	FR Hz	Vowel quality shifts as a result of fo variation
1 ↗	MA	e	m	131	3	131–e	65–e	330–y		65–330	Close-mid–close (and unrounded–rounded) shift with no change of the spectral peak structure.
2 ↗	FA	ø	m	131	3	131–ø	330–y	440–y		131–440	Close-mid–close shift with no substantial change of the spectral peak structure and close-mid–close shift with a change of the first spectral peak frequency or related prominent harmonic frequency.
3 ↗	MA	o	m	131	3	131–o	330–u	440–u		131–440	Close-mid–close shift with no change of the spectral peak structure and close-mid–close shift with a change of the first two spectral peak frequencies or related prominent harmonic frequencies.
4 ↗	FA	o	m	131	2	131–o	440–u			131–440	Close-mid–close shift with very limited change of the spectral peak structure.
5 ↗	PS	y	m	131	3	131–ø	65–ø	330–y		65–330	Close-mid–close shift with a very limited spectral change > 1 kHz in terms of “smeared” peaks.
6 ↗	PS	ɛ	w	220	3	220–ɛ	131–ɛ	523–y		131–523	Open-mid–close shift with a very limited spectral change > 1 kHz in terms of “smeared” peaks.
7 ↗	FA	a	m	131	3	131–a	65–a	330–o		65–330	Open–close-mid shift with no change of the spectral envelope (see also Series 8).
8 ↗	MA	a	m	131	3	131–a	65–a	330–o		65–330	Open–close-mid shift with no change of the spectral envelope (see also Series 7).
9 ↗	PS	a	w	220	4	220–a	65–a	262–a	440–a	65–440	No shift and no change of the spectral envelope, in contrast to Series 7 and 8.
10 ↗	MA	u	w	220	4	220–u	131–u	330–u	523–u	131–523	No shift despite marked fo variation and spectral changes < 1 kHz.
11 ↗	FA	i	w	220	2	220-i	523–e			220–523	Inverted close–close-mid shift effected by fo surpassing the first filter frequency of synthesis.

M3.2 Source–Filter Resynthesis Based on Estimated Spectral Envelopes of Single Natural Sounds, Including f_o Variation

Introduction

Pursuing the investigation of vowel recognition for synthesised and resynthesised sounds (see the general Introduction to this treatise regarding this terminological differentiation) based on supposed vowel quality-related spectral characteristics but varying f_o, in a further study, a resynthesis was conducted relating to spectral envelopes of single natural vowel sounds instead of using estimated statistical *F*-patterns.

Hillenbrand et al. (2006) reported an experiment in which, besides word and consonant recognition, vowel recognition was assessed comparing three types of source-filter synthesis (understood here as resynthesis): (i) the filter being a harmonic envelope of a single natural utterance (further detailed in Paul, 1981; Hillenbrand and Houde, 2003), (ii) the filter being a sum of exponentially damped sine waves at frequencies and amplitudes that correspond to the spectral peaks of a single natural utterance (further detailed in Hillenbrand and Houde, 2002) and (iii) the filter being an estimated *F*-pattern of a single natural utterance. A 300-utterance subset (hVd syllables) of the sound database of Hillenbrand et al. (1995) was selected and served as a basis for vowel resynthesis (for details of the subset investigated, see Hillenbrand and Nearey, 1999). The f_o levels of the source applied in resynthesis corresponded to the f_o contours of the original sounds. (Roughly speaking and indicated according to the musical C-major scale, the corresponding f_o levels were comparable to approximately 131 Hz for men, 220 Hz for women, and 262 Hz for children.) Vowel recognition was assessed in a listening test involving 13 listeners. The results of the test showed that the average vowel recognition rate was nearly identical for natural sounds (95.2%) and sounds resynthesised using the spectral envelope (95%), with the rate dropping somewhat for peak-related resynthesis (90% for Klatt synthesis, 89.3% for damped sine wave synthesis). (Note that some vowel quality-related differences in recognition results occurred, which are not discussed further here.) The results thus indicated that vowel sounds resynthesised based on the spectral envelopes of single utterances might be recognisable in a similar way to natural utterances.

However, this recognition result was again shown only for cases for which the f_o of the natural and the resynthesised sounds were similar because different f_o levels for resynthesis related to single unchanged filters were not investigated. Consequently, f_o variation was again not included in the experiment, neither concerning the production of natural vowel sounds nor concerning resynthesis. Therefore, a new and extended resynthesis experiment was conducted in the framework of this treatise: Spectral envelopes of natural vowel sounds of single speakers produced at very different f_o were used in resynthesis, and for every single envelope, the resynthesis in its turn included different f_o levels. Notably, within this experimental setting, vowel recognition was not mainly examined for f_o variation of the resynthesised replicas, as was the case in the previous experiment related to statistical average *F*-patterns, but recognition was examined in parallel and to the same extent for f_o variation of the natural reference sounds and their replicas. As a consequence, f_o-related vowel quality shifts for resynthesised sounds related to an unchanged spectral envelope were not examined mainly for increasing f_o from a lower to a higher frequency level when compared with the f_o level of the natural reference sound, but in most cases for both directions of f_o variation.

Experiment

Reference sample of natural vowel sounds: Based on the Zurich Corpus, sounds of the eight long Standard German vowels produced by three untrained speakers (nonprofessionals), one man, one woman and one child, were selected. The selection criteria consisted of a large vocal range on the part of the speakers and a high vowel recognition rate (matching vowel intention) for all sounds produced by a speaker, as found in the standard listening test conducted when creating the Zurich Corpus. The sounds investigated were produced in nonstyle mode with voiced phonation and medium vocal effort in V context at different intended f_o levels of 220–262–330–440–523–659 Hz (all speakers), 165 Hz (man and woman) and 131 Hz (man only). As a result, a reference sample of 168 natural vowel sounds (64 sounds of the man, 56 sounds of the woman, and 48 sounds of the child) was created. The vowel recognition rates obtained in the listening test conducted when creating the corpus were 100% for 158 sounds, 80% for six sounds and 60% for three sounds, with one sound not being recognised. (The sounds with a recognition rate of 60% concerned the utterance of /o/ produced by the man at an intended f_o of 659 Hz, the utterance of /y/ produced by the woman at an intended f_o of 262 Hz and the utterance of /ɛ/ produced by the child at intended f_o of 440 Hz. The utterance of /o/ produced by the child at an intended f_o of 659 Hz was not recognised.)

Estimation of spectral envelopes and f_o: The source contour and the course of the spectral envelope of every single sound were calculated according to the methods described in Hillenbrand et al. (2006). For each pair of these two dynamic features related to a single natural utterance (sounds normalised to the same RMS level 0.2 relative to maximum), eight resynthesised sounds were produced based on an unchanged spectral envelope but varying f_o in terms of applying eight different f_o levels as approximations to average levels of 131–165–220–262–330–440–523–659 Hz: The f_o variation in resynthesis was controlled by applying a shift factor to the f_o contour of the natural reference sound, the factor being equal to the quotient of f_o of resynthesis and f_o of the natural reference sound, both calculated f_o levels approximately corresponding to the C-major scale. (Note that the low f_o levels of 131–165 Hz were also applied to the sounds of the child and the lowest f_o level of 131 Hz was also applied to the sounds of the woman.) As a result, a sample of 1344 configurations of spectral envelopes and f_o levels was created.

The f_o level of 65 Hz was not applied in the present study because no vowel quality shifts were observed for the f_o variation of 65–131 Hz in the previous experiment (see Chapter M3.1), except for two sounds and vowel boundary recognition. Conversely, the high f_o level of 659 Hz was added to extend the frequency range of f_o variation.

Resynthesis: Based on these configurations, a resynthesis was conducted using the spectral envelope synthesiser described in Hillenbrand et al. (2006). The sound duration of a resynthesised sound corresponded with the duration of its natural reference sound.

Listening test: Vowel recognition of the natural and resynthesised sounds was assessed in a listening test according to the standard procedure of the Zurich Corpus and involving the five standard listeners, with the test divided into eight subtests, each subset presenting sounds at similar f_o. The natural and resynthesised sounds were not separated in the test. (Note that even though vowel recognition of the natural vowel sounds used in this experiment had already been tested when creating the database of the Zurich Corpus, they were tested anew here in a different sound context. Hence, the vowel quality recognition results given below concern the results of this experiment-specific test.)

Results

The listening test results are shown in Table 1 in the chapter appendix. Considering only the vowel sounds for which a labelling majority of the listeners for a given vowel quality was obtained (recognition rate ≥ 60%), the below vowel quality matches or shifts were found compared to vowel intention. Note that, in this chapter's results and discussion sections, the direction of the recognised vowel quality shifts is again given for a stepwise increase of f_o from a lower to a higher frequency level of comparison. Note also the following perspective and terminology: If the vowel qualities of sounds resynthesised at f_o levels below the level of the natural reference sound in question were recognised as more open than the vowel quality of the reference sound, then this finding is not further discussed concerning different degrees of vowel openness, and a general shift in an open–close direction with increasing f_o levels is stated. In contrast, if the vowel qualities of sounds resynthesised at f_o levels equal to or above the level of the natural reference sound in question were recognised to be more open than the vowel quality of the reference sound, this finding is assigned as inverted or reverted (if previously a more closed vowel quality occurred in the series for sounds resynthesised at f_o levels equal to or above the level of the natural reference sound). The colour code applied in Table 1 accords with this perspective.

Vowel recognition of the natural sounds: The vowel qualities of all natural vowel sounds were recognised by the listeners (labelling majority) according to vowel intention, except for two high-pitched sounds of /o/ at 659 Hz produced by the woman and the child (see Table 1, extended table online, Column 12).

Vowel recognition of the resynthesised sounds at f_o of the natural reference sounds: For f_o levels up to 523 Hz, with the exception of four sounds only, the recognised vowel quality of all resynthesised sounds matched the intended quality of the related natural sound (see Table 1, vowel recognition results marked in green). However, at the highest f_o level of 659 Hz, the corresponding vowel recognition rate for the resynthesised sounds dropped substantially.

In detail: For the sounds produced by the man, the intended vowel qualities of the sounds of /ø/ at f_o of 440 Hz and of /i, y, e, ø, u/ at 659 Hz were not recognised as matching the intended vowel quality. The same held true for the sounds of /o/ produced by the woman at intended f_o of 165 and 659 Hz. As for the sounds of the child, the sounds of /o/ at 220 Hz, /i/ at 523 Hz and /i/ and /ø/ at 659 Hz were not recognised as matching the intended vowel quality.

Vowel recognition of the resynthesised sounds at an f_o level other than the level of the natural reference sounds: Since, in contrast to the previous synthesis experiment discussed in Chapter M3.1, shifts were prominent for sounds related to spectral envelopes of both close and close-mid vowels, the order of the below presentation of the results follows a close–open direction of vowel qualities.

For the majority of the natural sounds of the close vowels /i, y, u/, a resynthesis based on a single spectral envelope at f_o levels below the level of the natural reference sound resulted in a close–open shift or, from the perspective of a stepwise increase of f_o from lower to higher levels, a shift in an open–close direction with increasing f_o (see Table 1, results marked in red). It is noteworthy that the shifts exceeded two adjacent vowel qualities for several sound series of these vowels. Furthermore, for the sounds of these three vowels, a resynthesis based on a single spectral envelope at f_o levels above the level of the natural reference sound did not affect the recognised vowel quality, except for four singular and interfering cases of reverted close–open shifts (see Table 1, results marked in purple). Exceptions to these general results concerned one series of resynthesised sounds related to the natural sound of /i/ produced by the child at an intended f_o of 523 Hz, for which a front–back confusion and an open–close back shift were found, one additional case of a resynthesised sound related to the natural sound of /y/ produced by the child at an intended f_o of 659 Hz and four cases of resynthesised sounds related to the natural sound of /u/ produced by the woman at an intended f_o of 165 Hz, for which front–back confusions occurred (see Table 1, results marked in grey and with "*"; note also single identifications indicating front–back confusions in the listener-specific details of the vowel recognition test).

For the natural sounds of the close-mid vowels /e, ø, o/, a resynthesis based on a single spectral envelope and a stepwise increase of f_o from the lowest to the highest level investigated caused a general open–close shift. This shift concerned either the resynthesis of low f_o levels up to the f_o level of the natural reference sound, as was the case for close vowels, or the resynthesis from the f_o level of the natural reference sound to higher levels, or both kinds of f_o variation. Again, the shifts exceeded adjacent vowel qualities in several sound series related to a single spectral envelope. Exceptions to these general results were two interfering cases of inverted close–open shifts and three resynthesised sounds related to the natural sounds of /ø/ produced by the man at an intended f_o of 659 Hz and the child at an intended f_o of 262 Hz, for which front–back confusions were found.

For approximately half of the natural sounds of the open-mid vowel /ɛ/ produced by the man and the woman, a resynthesis based on a single spectral envelope and stepwise increasing f_o from the level of the natural reference sound to the highest level investigated resulted in an open–close shift. However, this shift concerned only resynthesised sounds related to natural vowel sounds produced at f_o of 131–330 Hz. No marked shifts were found for the natural sounds of /ɛ/ produced by the child. Additional aspects concern rare singular and interfering cases of inverted close–open shifts and one case of a front–back confusion (see the sound of /ɛ/ produced by the child at 440 Hz).

Except for two single cases, an f_o variation in the resynthesis of sounds of the open vowel /a/ did not cause a vowel quality shift.

Discussion

In general terms, the vowel recognition results indicated that a resynthesis of natural vowel sounds based on related estimated spectral envelopes and with a stepwise f_o level variation from low to high caused (i) marked vowel quality shifts in an open–close direction associated with an increase of f_o from low levels up to the levels of the natural reference sounds for sounds of the close vowels /i, y, u/ if the natural references were produced at middle or higher levels of f_o, (ii) general open–close shifts for sounds of the close-mid vowels /e, ø, o/ if f_o was substantially varied in resynthesis, (iii) nonuniform open–close shifts for sounds of the open-mid vowel /ɛ/ if f_o resynthesis levels surpassed the level of the natural reference sounds and if the natural reference sounds were produced at lower or middle levels of f_o, and (iv) marginal or no shifts for sounds of the vowel /a/. In these terms, and extending the findings of the previous experiment related to statistical *F*-patterns, this study showed in an exemplary manner that vowel recognition does not relate to a measured spectral envelope of a natural sound independent of f_o levels. Thus, the spectral envelope *per se* of a natural sound indeed proves to be an ambiguous representation of vowel quality due to its relation to f_o.

Most importantly, the two findings of a general open–close vowel quality shift direction with increasing f_o and, within the limit of this general shift direction, of the nonuniform character of individual shifts – that is, their dependence on the vowel qualities of the natural reference sounds, the f_o levels of these reference sounds, the f_o levels and the range of f_o variation of the resynthesised replicas as well as the individual course of the spectral envelope due to intra- or inter-speaker differences of sound production – confirmed the corresponding findings discussed in the previous chapters.

We dispense with a detailed discussion of occurring inverted or reverted close–open shifts or front–back or back–front confusions as the number of such cases was small, and no general interpretation can be derived. Above all, possible artefacts of the applied synthesis technique were not investigated. We interpret that these cases are not a reason for a substantial relativisation of the general indications we are arguing for.

A comparison of the results of this resynthesis experiment (based on single natural sounds and their spectral envelopes) with the results of the previous synthesis experiment (based on statistical *F*-patterns) highlights two main differences. The first difference concerns resynthesis related to the spectral envelopes of the natural vowel sounds produced at f_o below 330 Hz: If f_o was raised in resynthesis from a lower to a higher level and thereby substantially surpassed the frequency level commonly assumed as the statistical F_1 of the vowel quality in question, inverted or reverted close–open shifts following previous open–close shifts occurred very rarely, in contrast to the synthesis related to statistical *F*-patterns. The second difference concerns the extent of the vowel quality shifts: These shifts were far more pronounced in the present than in the previous experimentation.

Both kinds of recognition differences may be due either to the difference in the f_o ranges of the natural vowel sounds and their replicas investigated or to the difference between vowel synthesis based on averaged LPC curves and vowel resynthesis based on a spectral envelope of a single natural reference sound. In addition, the methodological problem of *F*-pattern and spectral envelope estimation with increasing levels of f_o may also have an impact on the recognition of resynthesised sounds as investigated here: With rising f_o, as is true for formant estimation, spectral smoothing becomes in its turn problematic because of spectral undersampling and interrelated distortions, the estimation problem being severe for $f_o \geq 300$ Hz (de Cheveigné and Kawahara, 1999; Hillenbrand and Houde, 2003). Thus, from a theoretical perspective, spectral envelope estimation for sounds produced at an f_o level markedly surpassing 300 Hz is methodologically unsubstantiated. This methodological issue may be understood as limiting the validity of the results of the present experiment, and it may explain some vowel quality shifts associated with decreasing f_o levels from high (f_o of the natural reference sound > 300 Hz) to low. However, the recognised vowel quality for almost all resynthesised sounds of natural reference sounds with f_o levels up to 523 Hz, resynthesised applying the f_o levels of these reference sounds, matched the intended vowel quality of the natural sound.Moreover, increasing f_o levels in resynthesis for envelopes related to natural reference sounds at lower f_o and decreasing f_o levels in resynthesis for envelopes related to natural reference sounds at higher f_o resulted in the same general vowel quality shift direction: an open–close direction with increasing f_o in resynthesis, which corresponded to the close–open direction with decreasing f_o. This result supported the validity of the experimental design and the results obtained.

Nevertheless, for a comparison of the results of this resynthesis experiment related to single natural sounds and their harmonic envelopes as well as the results of the previous synthesis experiment related to statistical *F*-patterns, both the principal difference between an LPC filter curve and a harmonic envelope and the methodological problems of LPC filter curve and harmonic envelope estimation have to be taken into account. Indeed, vowel recognition in source–filter resynthesis may depend on the type of filter applied.

As in the previous study, some further methodological limitations and relativisations regarding the results also have to be considered here. Although the sound quality for sounds produced with this type of spectral envelope synthesiser, including dynamic sound characteristics, is much better than for sounds produced with a Klatt synthesiser related to static *F*-patterns and static f_o, some artefacts still occurred, sometimes making listening to the sounds unpleasant to the ear. Although not under further investigation here, these artefacts must be considered when interpreting the results. Improved (re-)synthesis tools will be needed for future research. Also, note again that the spectral envelopes investigated in the present experiment did not mirror spectral changes due to variations in phonation, vocal effort and additional production modes, e.g. speaking and singing styles.

A final consideration concerns the possible dissociation between a filter curve applied in vowel sound production and the filter curves measured for the produced sound, as discussed in the previous chapter. Similarly to using LPC filter curves, if sounds are resynthesised using the same spectral envelope characteristics but applying different f_o levels, the envelopes measured for the radiated sounds can manifest substantial differences when compared with the envelope of sound production. (For exemplary illustration, see Table 2, Series 7 and 8; see also Series 6 and 11.)

Chapter appendix

Table 1. Source–filter resynthesis based on estimated spectral envelopes of single natural reference sounds, including f_o variation: Vowel recognition results. Columns 1–3 = natural reference sounds (SP = speaker; fo = f_o in Hz of the natural reference sounds according to musical C-major scale; V = intended and recognised vowel quality, with vowel recognition according to the standard listening test conducted when creating the Zurich Corpus). Columns 4–11 = vowel recognition results with a recognition rate of ≥ 60% per f_o level applied in resynthesis. Extended online table: Column 12 = details of the vowel recognition results for the natural reference sounds (experiment-specific listening test; labelling of the five listeners; note ns = no vowel specified, txt = free comment). Columns 13 ff. = details of the vowel recognition results for the resynthesised sounds. Colour code and "*" marks: Green = resynthesis at f_o of the natural reference sound; no colour = resynthesis at f_o differing from f_o of the natural reference sound but with recognised vowel quality matching vowel intention of the reference sound, or no labelling majority; red = resynthesis at f_o differing from f_o of natural reference sound associated with a vowel quality shift in an open–close direction with increasing f_o from lower to higher levels compared to vowel intention; purple = resynthesis at f_o differing from f_o of natural reference sound associated with a vowel quality shift in a close–open direction with increasing f_o from lower to higher levels; grey and/or "*" marks = front–back or back–front confusions.
[M-03-02-T01]

Table 2. Source–filter resynthesis based on estimated spectral envelopes of single natural reference sounds, including f_o variation: Exemplary illustrations. Columns 1–5 = natural reference sounds (S/L = series number and sound link; V = intended and recognised vowel quality, with vowel recognition according to the experiment-specific listening test; SP = speaker, where m = man, w = woman and c = child; fo = f_o in Hz of the natural reference sounds according to musical C-major scale; N = number of sounds documented). Columns 6 and 7 = resynthesised sounds (fo var = range of f_o variation in Hz, according to musical C-major scale; Illustration = aspects illustrated in the sound series). In Column 7, single sounds are given in terms of f_o–V, with f_o of resynthesis and vowel quality recognised (labelling majority).
[M-03-02-T02]

Table 1. Source–filter resynthesis based on estimated spectral envelopes of single natural reference sounds, including fo variation (in Hz): Vowel recognition results. [M03-02-T01]
Extended online table:

Natural sounds			Vowel recognition of resynthesis, per fo (in Hz)							
SP	**fo**	**V**	**131**	**165**	**220**	**262**	**330**	**440**	**523**	**659**
man	131	**i**	**i**	i	i	i	i	i	i	i
	165		e	**i**	e	i	i	i	y	i
	220		e	i	**i**	i	i	i	i	i
	262		e	i	e	**i**	i	i	i	i
	330		–	e	e	e	**i**	i	i	i
	440		ɛ	–	–	e	e	**i**	i	i
	523		–	ɛ	ɛ	e	e	e	**i**	i
	659		ɛ	ɛ	ɛ	ɛ	ɛ	e	ɛ	–
woman	165		e	**i**	e	i	i	i	i	i
	220		e	e	**i**	i	i	i	i	i
	262		i	i	i	**i**	i	i	i	i
	330		e	e	e	e	**i**	i	i	i
	440		e	e	e	e	e	**i**	i	i
	523		e	ɛ	ɛ	e	e	e	**i**	i
	659		ɛ	–	ɛ	–	ɛ	e	–	**i**
child	220		i	i	**i**	i	i	i	i	i
	262		e	i	i	**i**	i	i	i	i
	330		i	i	i	i	**i**	i	i	i
	440		–	–	–	e	e	**i**	i	–
	523		–*	ɔ*	o*	o*	o*	o*	u*	u*
	659		ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	e	–
man	131	**y**	**y**	y	y	y	y	y	y	y
	165		–	**y**	e	y	i	y	y	y
	220		e	y	**y**	e	–	y	y	y
	262		–	y	y	**y**	y	y	y	y
	330		e	e	e	e	**y**	i	y	y
	440		ɛ	–	ø	–	e	**y**	y	y
	523		ɛ	ɛ	ɛ	ɛ	e	ɛ	**y**	y
	659		ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	–	–
woman	165		ø	**y**	y	y	y	y	y	y
	220		–	y	**y**	y	y	y	y	y
	262		ø	y	ø	**y**	y	y	y	y
	330		ø	ø	ø	ø	**y**	y	y	y
	440		–	–	ø	ø	ø	**y**	y	y
	523		ɛ	ɛ	–	e	e	e	**y**	–
	659		a	–	ɛ	–	ɛ	ɛ	ɛ	**y**
child	220		ø	y	**y**	y	y	y	y	y
	262		ø	–	y	**y**	–	y	y	y
	330		–	ø	ø	y	**y**	y	y	–
	440		–	–	–	–	–	**y**	y	u*
	523		–	–	–	–	–	–	**y**	y
	659		a	–	–	ɔ*	ɛ	ɛ	–	**y**

Natural sounds			Vowel recognition of resynthesis, per fo (in Hz)							
SP	**fo**	**V**	**131**	**165**	**220**	**262**	**330**	**440**	**523**	**659**
man	131	**u**	**u**	u	u	u	u	u	u	u
	165		u	**u**	u	u	u	u	–	u
	220		u	u	**u**	u	u	u	u	u
	262		u	u	u	**u**	u	u	u	u
	330		u	u	u	u	**u**	u	u	u
	440		–	–	ɔ	o	o	**u**	u	u
	523		–	–	ɔ	–	o	o	**u**	u
	659		a	a	a	–	a	a	a	–
woman	165		u	**u**	u	u	i*	i*	i*	i*
	220		o	u	**u**	u	u	u	u	u
	262		u	u	–	**u**	u	u	u	u
	330		–	–	–	u	**u**	u	u	u
	440		o	o	o	u	o	**u**	u	u
	523		–	ɔ	–	ɔ	o	o	**u**	u
	659		ɔ	ɔ	ɔ	o	o	o	o	**u**
child	220		u	u	**u**	u	u	u	u	u
	262		u	u	u	**u**	u	u	u	u
	330		o	–	u	u	**u**	u	u	u
	440		o	ɔ	o	o	o	**u**	u	u
	523		–	–	o	–	o	–	**u**	u
	659		ɔ	ɔ	ɔ	ɔ	o	o	o	**u**
man	131	**e**	**e**	e	e	e	i	y	i	y
	165		e	**e**	e	e	e	y	y	y
	220		e	e	**e**	e	e	i	y	y
	262		–	–	–	**e**	e	y	y	y
	330		ɛ	ɛ	ɛ	ɛ	**e**	e	e	y
	440		ɛ	ɛ	ɛ	ɛ	ɛ	**e**	ɛ	y
	523		ɛ	ɛ	ɛ	ɛ	e	e	**e**	e
	659		–	–	–	ɛ	ɛ	ɛ	ɛ	–
woman	165		e	**e**	e	e	i	i	i	i
	220		e	e	**e**	e	i	i	i	i
	262		ɛ	e	ɛ	**e**	e	i	i	i
	330		–	–	e	e	**e**	e	e	i
	440		e	ɛ	e	e	e	**e**	e	–
	523		ɛ	ɛ	ɛ	ɛ	–	e	**e**	e
	659		ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	**e**
child	220		e	e	**e**	e	e	i	i	i
	262		–	e	e	**e**	e	–	–	i
	330		ɛ	ɛ	ɛ	ɛ	**e**	e	e	i
	440		e	–	e	–	e	**e**	e	–
	523		ɛ	–	ɛ	–	ɛ	e	**e**	e
	659		ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	e	**e**

Table 1 (continuation). [M03-02-T01]

Natural sounds			Vowel recognition of resynthesis, per fo (in Hz)							
SP	**fo**	**V**	**131**	**165**	**220**	**262**	**330**	**440**	**523**	**659**
man	131	**ø**	**ø**	ø	ø	y	y	y	y	y
	165		ø	**ø**	ø	ø	–	y	y	–
	220		ø	ø	**ø**	ø	ø	y	y	y
	262		–	ø	ø	**ø**	ø	y	y	–
	330		–	–	ø	ø	**ø**	ø	y	–
	440		ε	–	ε	–	–	**–**	ø	–
	523		–	ε	ε	–	ø	ø	**ø**	ø
	659		a	–	a	ɔ*	ɔ*	–	–	**–**
woman	165		ø	**ø**	ø	ø	ø	y	y	y
	220		ø	ø	**ø**	ø	–	y	y	y
	262		–	ø	ø	**ø**	ø	y	y	y
	330		ø	ø	ø	ø	**ø**	ø	y	–
	440		ø	ε	ø	–	ø	**ø**	–	y
	523		–	–	–	ε	–	–	**ø**	–
	659		a	–	–	ε	ε	ε	ε	**ø**
child	220		ø	ø	**ø**	ø	y	y	y	y
	262		o*	–	ø	**ø**	ø	y	y	y
	330		–	–	ø	–	**ø**	ø	y	y
	440		–	–	ø	–	ø	**ø**	–	y
	523		ε	ε	ε	ε	ε	ø	**ø**	–
	659		a	a	a	–	–	ø	–	–
man	131	**o**	**o**	o	o	o	–	–	–	–
	165		o	**o**	–	u	u	u	u	u
	220		–	o	**o**	o	o	u	u	u
	262		–	o	o	**o**	o	o	u	u
	330		ɔ	ɔ	ɔ	ɔ	**o**	o	–	u
	440		a	ɔ	ɔ	ɔ	o	**o**	o	u
	523		a	a	a	–	–	ɔ	**o**	–
	659		ɔ	a	–	ɔ	ɔ	ɔ	a	**o**
woman	165		u	**u**	u	u	u	u	-	u
	220		o	o	**o**	o	u	u	u	u
	262		o	o	o	**o**	o	u	u	u
	330		o	o	o	–	**o**	–	u	u
	440		a	ɔ	ɔ	ɔ	ɔ	**o**	–	u
	523		a	a	ɔ	–	–	o	**o**	ɔ
	659		a	a	a	a	a	a	a	–
child	220		–	o	**u**	u	u	u	u	u
	262		o	o	o	**o**	u	u	u	u
	330		–	–	–	ɔ	**o**	u	–	u
	440		ɔ	ɔ	ɔ	ɔ	o	**o**	o	u
	523		–	ɔ	–	–	ɔ	o	**o**	–
	659		a	a	a	a	a	a	–	**o**

Natural sounds			Vowel recognition of resynthesis, per fo (in Hz)							
SP	**fo**	**V**	**131**	**165**	**220**	**262**	**330**	**440**	**523**	**659**
man	131	**ε**	**ε**	ε	ε	ε	e	e	e	–
	165		ε	**ε**	ε	–	–	–	–	y
	220		ε	ε	**ε**	ε	ε	ε	ε	e
	262		ε	ε	ε	**ε**	ε	ε	–	y
	330		ε	ε	ε	ε	**ε**	ε	ε	–
	440		ε	ε	ε	ε	ε	**ε**	ε	–
	523		ε	ε	ε	ε	ε	ε	**ε**	ε
	659		ε	–	ε	ε	ε	ε	ε	**ε**
woman	165		ε	**ε**	e	e	–	i	–	y
	220		ε	ε	**ε**	ε	e	e	e	i
	262		ε	ε	ε	**ε**	e	e	e	e
	330		ε	ε	ε	ε	**ε**	ε	e	e
	440		ε	a	ε	ε	ε	**ε**	ε	ε
	523		a	–	ε	ε	ε	ε	**ε**	ε
	659		ε	–	ε	–	ε	ε	ε	**ε**
child	220		ε	ε	**ε**	ε	ε	ε	ε	ε
	262		ε	ε	ε	**ε**	ε	ε	ε	–
	330		ε	ε	ε	ε	**ε**	ε	ε	ε
	440		a	–	–	–	ε	**ε**	ɔ*	ε
	523		aε	a	aε	–	ε	ε	**ε**	ε
	659		a	–	–	ε	ε	ε	ε	**ε**
man	131	**a**	**a**	a	a	a	a	a	a	a
	165		a	**a**	a	a	a	a	a	a
	220		a	a	**a**	a	a	a	a	–
	262		a	a	a	**a**	a	a	a	a
	330		a	a	a	a	**a**	a	a	a
	440		a	a	a	a	a	**a**	a	a
	523		–	a	a	a	a	a	**a**	a
	659		–	a	a	a	a	–	ε	**a**
woman	165		a	**a**	–	a	a	a	a	o
	220		a	a	**a**	a	a	a	a	a
	262		a	a	a	**a**	a	a	–	a
	330		a	a	a	a	**a**	a	a	a
	440		a	a	a	a	a	**a**	a	a
	523		a	a	a	a	a	a	**a**	a
	659		a	a	a	a	a	a	a	**a**
child	220		a	a	**a**	a	a	a	a	a
	262		a	a	a	**a**	a	a	a	a
	330		a	a	a	a	**a**	a	a	a
	440		a	a	a	a	a	**a**	a	a
	523		a	a	a	a	a	a	**a**	a
	659		a	a	a	a	a	a	a	**a**

Table 2. Source–filter resynthesis based on estimated spectral envelopes of single natural reference sounds, including fo variation (in Hz): Exemplary illustrations. [M-03-02-T02]

Natural sounds					Resynthesis	
S/L	**V**	**SP**	**fo**	**N**	**fo var**	**Illustration**
2a: Sounds of the close vowels. Series 1 to 3=resynthesised sounds with fo equal to or lower than fo of the original sound. The fo decrease in synthesis effected vowel quality shifts in a close–open direction or, looked at from the perspective of low to high fo levels, in an open–close direction with rising fo.						
1 ↗	i	w	440	5	220–440	S1=natural reference sound, 440–i; S2=resynthesis, 440-i; S3–S5=resynthesis, 330-e, 262-e, 220-e; a close-mid–close shift effected by an fo increase from 220 to 440 Hz.
2 ↗	y	w	330	3	220–330	S1=natural reference sound, 330–y; S2=resynthesis, 330-y; S3=resynthesis, 220-ø; a close-mid–close shift effected by an fo increase from 220 to 330 Hz.
3 ↗	u	m	440	5	220–440	S1=natural reference sound, 440–u; S2=resynthesis, 440-u; S3–S5=resynthesis, 330-o, 262-o, 220-ɔ; an open-mid–close shift effected by an fo increase from 220 to 440 Hz.
2b: Sounds of the close-mid vowels. Series 4 to 7=resynthesised sounds with fo equal to or lower or higher than fo of the original sound. The fo variation effected again vowel quality shifts in an open–close direction with increasing fo.						
4 ↗	e	c	330	4	220–659	S1=natural reference sound, 330–e; S2=resynthesis, 330-e; S3=resynthesis, 220-ɛ; S4=resynthesis, 659-i; an open-mid–close shift effected by an fo increase from 220 to 659 Hz.
5 ↗	e	w	220	3	220–440	S1=natural reference sound, 220–e; S2=resynthesis, 220-e; S3=resynthesis, 440-i; a close-mid–close shift effected by an fo increase from 220 to 440 Hz.
6 ↗	ø	m	131	7	131–659	S1=natural reference sound, 131–ø; S2=resynthesis, 131–ø; S3–S7=resynthesis, 262-y, 330-y; 440-y; 523-y; 659-y; a close-mid–close shift effected by an fo increase from 131 to 262 Hz and then further up to 659 Hz; fo of the resynthesised sounds at higher levels surpass the first spectral peak of the natural reference sound at lower fo.
7 ↗	o	w	220	6	220–659	S1=natural reference sound, 220–o; S2=resynthesis, 220-o; S3–S6=resynthesis, 330-u, 440-u; 523-u; 659-u; a close-mid–close shift effected by an fo increase from 220 to 330 Hz and then further up to 659 Hz; fo of the resynthesised sounds at higher levels surpass the first spectral peak of the natural reference sound at lower fo; inversion of lower spectral maxima and minima for the first two sounds and the last sound of the series.
8 ↗	o	m	165	6	165–659	S1=natural reference sound, 165–o; S2=resynthesis, 165-o; S3–S6=resynthesis, 330-u, 440-u; 523-u; 659-u; a close-mid–close shift effected by an fo increase from 165 to 659 Hz; fo of the resynthesised sounds at higher levels surpass the first spectral peak of the natural reference sound at lower fo.

Table 2 (continuation). [M-03-02-T02]

Natural sounds					Resynthesis	
S/L	V	SG	fo	N	fo range	Illustration
2c: Sounds of the open-mid vowel. Series 9 and 10=resynthesised sounds with fo equal to or lower or higher than fo of the original sound. The fo variation effected nonuniform vowel quality shifts in an open–close direction with rising fo, that is, different effects of the fo variation were found for sound series of the same vowel an comparable ranges of fo variation.						
9 ↗	ε	w	220	5	165–659	S1=natural reference sound, 220–ε; S2=resynthesis, 220-ε; S3–S5=resynthesis, 165–ε, 440-e, 659-i; an open-mid–close shift effected by an fo increase from165 to 659 Hz; fo of the resynthesised sounds at higher levels surpass the first spectral peak of the natural reference sound at lower fo.
10 ↗	ε	c	330	7	220–659	S1=natural reference sound, 330–ε; S2=resynthesis, 330-ε; S3 and S4=resynthesis, 262-ε, 220-ε; S5–S7=resynthesis, 440-ε; 523-ε; 659-ε; lacking effect of fo variation on vowel quality.
2d: Sounds of the open vowel. Series 11 and 12=resynthesised sounds of /a/ with fo equal to or lower or higher than fo of the original sound. In these examples, the fo variation effected no vowel quality shift.						
11 ↗	a	w	220	9	131–659	S1=natural reference sound, 220–a; S2=resynthesis, 220–a; S3 and S4=resynthesis, 165–a, 131–a; S5–S9=resynthesis, 262–a; 330–a; 440-a; 523-a; 659-a; lacking effect of fo variation on vowel quality.
12 ↗	a	c	440	7	220–659	S1=natural reference sound, 440–a; S2=resynthesis, 440–a; S3–S7=resynthesis, 220–a; 262–a; 330-a; 523-a; 659-a; lacking effect of fo variation on vowel quality.
2e: Additional illustration. Resynthesised sounds based on a single spectral envelope but varying fo can manifest different spectral envelopes when measured from the radiated sounds in question.						
13 ↗	i	m	131	7	131–659	S1=natural reference sound, 131–i; S2=resynthesis, 131–i; S3–S7=resynthesis, 262–i; 330–i; 440-i; 523-i; 659-i; a lacking effect of fo variation on vowel quality; different spectral envelopes of production and acoustic measurement (see sounds at fo > 300 Hz).

M3.3 Source–Filter Synthesis Based on Model Filter Patterns of Long Standard German Vowels, Including Variation of f_o

Introduction

In a third type of experimentation, we attempted to create a model synthesis that allows for a straightforward replication and testing of vowel quality shifts that occur with rising f_o. The basic idea of the model synthesis was to create interrelated filter patterns and f_o levels in terms of all filter frequencies of a pattern being multiples (in whole numbers) of f_o for two or three f_o levels. These models represent "ideal" cases of filter curves and harmonic spectra, the dominant harmonics always coinciding with the filters, and the only acoustic differences between the sounds being f_o (and pitch) and frequency distances of harmonics.

Three experiments were conducted. In the first synthesis experiment, the f_o variation for model *F*1–*F*2–*F*3 patterns (in terms of approximations to observed spectral envelopes of natural vowel sounds) was investigated, the f_o range being 200–600 Hz. The approximations were based on an extensive analysis of the sounds of the Zurich Corpus. The second synthesis experiment investigated the same model *F*-patterns but with f_o levels halved to 100–300 Hz, keeping the range of f_o variation in semitones unchanged. In a third experiment, the first synthesis was repeated, but with the levels of the first two filters varied. (For an earlier report on the experimental design and vowel recognition results of synthesised sounds, see the conference contributions and additional online materials of Maurer et al., 2017; Kathiresan et al., 2018).

Experiment 1

***F*-patterns and f_o variation investigated:** In the first experiment, model *F*1–*F*2–*F*3 patterns were created, which approximately related to observed *F*-patterns of natural sounds of the back vowels /o, ɔ/ and the front vowels /e, ø, ɛ/ produced at an f_o level of c. 200 Hz. For sound synthesis related to natural sounds of the close-mid vowels /o, ø, e/, F_1–F_2–F_3 were set as multiples of 400 Hz, and the two f_o levels of 200 and 400 Hz were investigated in synthesis. For sound synthesis related to natural sounds of the open-mid vowels /ɔ, ɛ/, F_1–F_2–F_3 were set as multiples of 600 Hz, and the three f_o levels of 200, 300 and 600 Hz were investigated in synthesis. The bandwidths and formant levels were set to bring the resulting sound spectra into line with the spectra of the natural sounds. For all *F*-patterns, F_4–F_5 of 4200–5400 Hz with 200 Hz bandwidths and low levels of 50 dB were added to smoothen the higher frequencies > 3.5 kHz. Table 1 in the chapter appendix shows the resulting 17 configurations of model *F*-patterns and f_o levels.

Synthesis: For all configurations of *F*-patterns and f_o levels, steady-state sounds of 1 sec. were synthesised using the Klatt synthesiser as implemented in the Praat software (parallel mode, sampling frequency SF = 44.1 kHz, fade in/out = 0.03 sec.).

Listening test: Vowel recognition was assessed according to the standard test procedure of the Zurich Corpus and involving the five standard listeners, with each synthesised sound presented twice in the test.

Results 1

The listening test results are shown in Table 1 in the chapter appendix (see the results for the main experiment, confusion matrix and labelling majority in terms of a recognition rate of ≥ 50% for vowel openness). For all *F*-patterns investigated, increasing the level of f_o resulted in vowel quality shifts in an open–close direction. An f_o difference of seven semitones or one octave caused a change in vowel quality to an adjacent quality, and an f_o difference of c. 19 semitones resulted in a change to a non-adjacent quality.

Experiment 2

In a second experiment, based on the same configuration of model *F*-patterns as in experiment 1, synthesis was repeated by applying $f_o = f_o/2$ of the previous experiment (lowering f_o by one octave) in order to determine the role of the f_o range in this type of investigation. The vowel quality recognition of the sounds was examined in a separate listening test, according to the procedure of experiment 1.

Results 2

The listening test results are shown in Table 1 in the chapter appendix (see the results for experiment 2; for the confusion matrix, see Kathiresan et al., 2018). In contrast to experiment 1, an f_o difference of seven semitones from 100 to 150 Hz and of an octave from 100 to 200 Hz caused no vowel quality shift in an open–close direction. However, note that a synthesis based on an *F*-pattern related to natural sounds of /ɔ/ produced /a/. A more pronounced f_o difference of c. 19 semitones resulted in the vowel shifts /a–o/, /ɛ–e/ and /ɛ–ø/.

Experiment 3

In a third experiment, based on the same configuration of model *F*-patterns and f_o variation as described for experiment 1, synthesis was repeated with two different level variations, L_{F1}= -10 dB and L_{F2} = +10 dB, and L_{F1}= -20 and L_{F2} = +10 dB, in order to investigate the role of filter levels for this type of investigation. Vowel quality recognition of the sounds was again examined in a separate listening test, according to the procedure of experiment 1.

Results 3

The listening test results are shown in Table 1 in the chapter appendix (see the results for experiment 3; for the confusion matrix, see Kathiresan et al., 2018). In contrast to experiment 1, for both variations of the first two filter levels, increasing f_o in synthesis by one octave from 200 to 400 Hz had no effect for the *F*-pattern related to natural sounds of /o/. Also, for the *F*-pattern related to natural sounds of /ɔ/, increasing f_o in synthesis from 200 to 300 and to 600 Hz did not result in an /ɔ–o/ or /ɔ–o–u/ shift but in an /a–o/ shift. Concerning the recognition of front vowels, a comparison of the results of experiments 1 and 3 showed no pronounced differences concerning open–close vowel quality shifts, except for one sound in the last sound series.

General discussion

The results of the first synthesis experiment showed consistent vowel quality shifts in an open–close direction with increasing f_o for all sound pairs and sound triples tested. Shifts to adjacent vowel qualities /o–u/, /e–i/ and /ø–y/ were found for *F*-patterns as multiples of 400 Hz and f_o variation of 200–400 Hz, and shifts to adjacent and non-adjacent vowel qualities /ɔ–o–u/, /ɛ–e–i/ and /ɛ-ə–ø–y/ were found for *F*-patterns as multiples of 600 Hz and f_o variation of 200–300–600 Hz. Based on these model *F*-patterns and f_o level variations, in a paradigmatic way, the relation of spectral envelope representation of vowel quality to f_o was made evident by the vowel recognition results.

However, as highlighted in the various previous experiments, the vowel quality shifts due to f_o variation were again found to be nonuniform in the present experiment. Above all, contrary to the seven-semitone or one-octave f_o variation of 200–300 Hz or 200–400 Hz in experiment 1, no vowel quality shifts were found in the second experiment for the seven-semitone or one-octave f_o variation of 100–150 Hz or 100–200 Hz. This result indicated that not only the extent of f_o variation but also the frequency range of this variation impacted vowel quality shifts. (Note here the observation of Traunmüller, 1981, that an f_o variation below 150 Hz did not affect the recognised vowel openness, in contrast to an f_o variation above 150 Hz, keeping an *F*-pattern unchanged in vowel synthesis. For a subsequent discussion of this “end of scale effect”, see Fahey and Diehl, 1996.)

In addition, the level ratio of the investigated filters also influenced the vowel recognition results – although to a limited degree – above all for synthesised sounds related to *F*-patterns of natural sounds of back vowels. Note in this context that the role of formant amplitude variation in synthesis for vowel recognition is a matter of debate (for an overview, see Kiefte et al., 2010). However, the studies published on this matter are not discussed here in detail. It is difficult to relate them to each other since they applied different *F*-patterns, f_o levels and formant amplitudes or formant levels.

As said, the general aim of this type of experimentation was to contribute to the creation of model experiments that allow for a simple verification of a perceptual change in recognised vowel quality caused by f_o changes only, keeping the resonance curve of vowel production unchanged and applying different f_o levels in a way that allows for an exact spectral sampling of the resonance frequencies for all sounds. At the same time, this type of experimentation also highlights the nonuniform character of the effect of f_o variation. For the present context, note again that a Klatt synthesiser is implemented in the Zurich Corpus in order to allow for a straightforward replication of the experiment.

Chapter appendix

Table 1. Source–filter synthesis based on model filter patterns of Standard German back and front vowels, including f_o variation in synthesis: *F*-patterns and f_o variation investigated and vowel recognition results (details). Columns 1–13 = model *F*-patterns and f_o variation investigated (VO = vowel openness; S/L = series of sound pairs or triplets, and sound links for the main experiment 1; fo = f_o levels applied in synthesis in experiment 1, in Hz; Δfo = f_o level differences in reference to the first sound of a series, in semitones, ST; $F_{(i)}$, $L_{(i)}$ and $B_{(i)}$ = formant frequencies, levels and bandwidths applied in synthesis). Columns 14–21 = vowel recognition results in terms of the confusion matrix and the labelling majority for vowels and vowel openness obtained (Maj = values for recognition rates ≥ 50%). Column 22 = vowel recognition results of experiment 2 (recognition rates ≥ 50%). Columns 23 and 24 = vowel recognition results of experiment 3 (3a = labelling majority for level variation of L_{F1}–L_{F2} of -10 and +10 dB, 3b = labelling majority for level variation of L_{F1}–L_{F2} of -20 and +10 dB, with recognition rates ≥ 50%). Colour code: Red = vowel quality shifts in an open–close direction related to f_o variation. Note that the indication “b/a” in the confusion matrix denotes “back vowel or /a/”. Recognition results in parenthesis indicate a labelling majority for two adjacent vowels. For replication, use the KlattSyn tool in the Zurich Corpus; see also Figures C-03-03-F01 and C-03-03-F02 in the main text.
[M-03-03-T01]

Table 1. Source–filter synthesis based on model filter patterns of Standard German back and front vowels, including fo variation in synthesis: F-patterns and fo variation investigated and vowel recognition results (details). [M-03-03-T01]

Vowel synthesis													Vowel recognition (experiments 1–3)										
Model F-patterns and fo variation													Main experiment 1								Exp. 2 and 3		
VO	S	fo	Δfo	F1	L1	B1	F2	L2	B2	F3	L3	B3	Confusion matrix							Maj	2	3a	3b
		Hz	ST	Hz	dB	Hz	Hz	dB	Hz	Hz	dB	Hz	–	–	–	a	ɔ	o	u	V			
back	1	200	ref	400	100	100	800	105	100	2800	90	200						9	1	o	o	o	o
		400	12																10	u	o	o	o
	2	200	ref	600	100	100	1200	95	100	3000	85	200				3	7			ɔ	a	a	a
		300	7															10		o	a	o	o
		600	19																10	u	o	–	o
													ə	ε	ø	e	y	i	b/a				
front	3	200	ref	400	100	100	2400	100	200	2800	100	200			1	9				e	e	e	e
		400	12													1	3	6		i	e	i	i
	4	200	ref	400	100	100	2800	100	200	3200	100	200				9			1	e	e	e	e
		400	12														2	8		i	e	i	i
	5	200	ref	400	100	100	2000	100	150	2800	100	200			8	2				ø	ø	ø	ø
		400	12														10			y	ø	y	y
	6	200	ref	600	100	100	2400	100	200	3000	100	200		8					2	ε	ε	ε	ε
		300	7											4		6				e	ε	e	e
		600	19													4	1	5		i	e	i	(y-i)
	7	200	ref	600	100	100	1800	100	150	3000	100	200	5	5						(ə-ε)	ε	ε	ε
		300	7											2	8					ø	ε	ø	ø
		600	19												2		8			y	ø	y	(ø-y)

M3.4 Source–Filter Synthesis Based on Model Filter Patterns of Half-Open Tubes, Including Variation of f_o

Introduction

Extending the previous experimentation, vowel synthesis based on model F-patterns was also related to three half-open tube resonance patterns commonly attributed to the three average vocal tract lengths of men, women and children, respectively. In the literature, it is common to associate voiced vowel sounds produced with half-open tube resonances to the "neutral" vowel quality schwa (see e.g. Fant, 1970, pp. 54–57; Laver, 1994, p. 410; Pickett, 1999, pp. 38–40; Rendall et al., 2005). However, to the best of our knowledge, it has not yet been demonstrated in the literature that this association is independent of f_o variation. Therefore, a new experiment was designed, applying the same procedure as for the first experiment described in the previous chapter. This new experiment was conducted in the context of an Interspeech Show and Tell conference presentation (Maurer et al., 2019; see also Maurer et al., 2017), and it is integrated into the context of the present treatise and its line of argument.

Experiment

Half-open tube resonance patterns and f_o variation: Three F_1–F_2–F_3–F_4–F_5 patterns commonly attributed to the three average vocal tract lengths and related open-tube resonance patterns of men, women and children (for adults, see e.g. Pickett, 1999, p. 39) were configured, F_1 set to 500, 600 or 700 Hz, respectively, and the higher frequencies set to odd multiples of F_1. All bandwidths were set to 100 Hz. For each of the three patterns, three f_o levels of 1/3, 1/2 and 1/1 of the first filter frequency were investigated. For the resulting synthesis, again, the frequencies of the dominant harmonics always coincided with the filter frequencies, and the sounds differed only in f_o (and pitch) and the frequency distances of the harmonics. Table 1 shows the resulting nine configurations of F-patterns and f_o levels.

Synthesis: For all nine configurations of F-patterns and f_o levels, steady-state sounds of 1 sec. were synthesised using the Klatt synthesiser as implemented in the Praat software (cascade mode, sampling frequency = 44.1 kHz, fade in/out = 0.05 sec.).

Listening test: Vowel recognition was assessed in three subtests performed by the five standard listeners of the Zurich Corpus.

In subtest 1, all of the nine synthesised sounds were presented in random order. In the first round, the listeners were asked to listen to these nine sounds to familiarise themselves with the sounds and their quality. In a second round, they were asked to vocally repeat (imitate) each of the presented nine vowel sounds (ignoring the pitch) and then to assign a vowel quality or a boundary of adjacent vowel qualities as shown on the standard test screen of the Zurich Corpus. During the test, the listeners were allowed to repeat sound playback for each vowel as many times as they wanted if they felt unsure about their choice of vowel quality.

Note that, in addition to the vowel qualities and quality boundaries labelled according to the standard procedure of the Zurich Corpus, the vowel /œ/ was also accepted as an additional labelling option because two listeners insisted on the recognition of this vowel quality or a vowel boundary including this quality. Note, therefore, the difference between /œ/ (denoting a single vowel quality) and /øe/ (denoting the vowel boundary between the two adjacent vowels /ø/ and /e/).

In subtest 2, for each *F*-pattern separately, the related sound triplets were presented in ascending order of f_o level. In the first round, the listeners were asked to listen to each single sound triplet in order to, again, become familiar with the sounds and their quality in the context of this new presentation form. In the second round, they were asked to vocally repeat (imitate) all three sounds of a triplet (ignoring the pitch) and then to assign them to three vowel qualities or vowel boundaries as shown on the standard test screen of the Zurich Corpus. Again, the listeners were allowed to repeat sound playback as many times as they wanted if they felt unsure about their choice of vowel qualities.

In subtest 3, the second test was repeated with the sounds of the triplets presented in descending order of f_o level.

Results

Table 1 in the chapter appendix shows the *F*-patterns and f_o variation investigated and the listening test results. In the table, the overall vowel recognition results are given in Columns 9 and 10, including the results of all subtests, and listener-specific details of vowel recognition and subtest-specific vowel qualities with a recognition rate of ≥ 60% are given in Columns 11 to 17. According to the recognition results and the obtained labelling majorities for the sounds investigated, for the *F*-pattern commonly attributed to men, the recognised vowel quality for the f_o level of 125 Hz applied in synthesis was schwa. However, when f_o was increased by one octave, the recognised vowel quality shifted to the adjacent quality of /ø/ in all three listening conditions. Increasing f_o by two octaves caused a shift to the non-adjacent quality of /y/ for the second and third listening conditions and a shift to the vowel boundary of /ø/ and /y/ for the first condition. For the *F*-pattern of women, the same results were found for both the initial f_o level of 200 Hz applied in synthesis as well as for an f_o variation of approximately seven and 19 semitones. For the *F*-pattern of children, the recognised vowel quality for the f_o level of 233 Hz was /ɛ/–like. An f_o increase of approximately seven semitones only resulted in a vowel boundary confusion. However, increasing f_o by 19 semitones resulted in a shift to the non-adjacent quality of /y/. Minor differences related to either the different listening test conditions or to listener-specific labelling.

Discussion

While vowel synthesis based on half-open tube resonance patterns related to approximated average vocal tract lengths of adults combined with speaker group-specific f_o levels as given in formant statistics did produce a schwa sound, comparable to the results of experiment 1 in the previous chapter, the recognised vowel quality shifted in an open–close direction with increasing f_o levels in synthesis, thereby involving adjacent and non-adjacent vowel qualities. Thus, the ambiguity of *F*-patterns and spectral envelopes in their representation of vowel quality also markedly affected half-open tube resonance patterns: The findings indicated that these resonance patterns, commonly assumed to relate to neutral or centralised articulatory configurations, are not recognised consistently as neutral schwa vowels.

Some differences were found for the *F*-pattern of children: Above all, the sound synthesised at the low f_o level was not recognised as a clear schwa but rather as an /ɛ/–like sound, and only an increase of f_o by 19 semitones caused a pronounced vowel quality shift in an open–close direction. This finding again points to the nonuniform character of the relation between vowel quality and spectral characteristics since the results depended on the *F*-patterns and the frequency range of f_o variation.

Apart from these aspects, as stated, vowel recognition proved to be somewhat dependent on the vowel context in the listening test and on single listeners.

Chapter appendix

Table 1. Source–filter synthesis based on model filter patterns of half-open tubes, including *f*ₒ variation: *F*-patterns and *f*ₒ variation investigated and vowel recognition results (details). Columns 1–8 = vowel synthesis (SG/L = speaker group commonly related to the resonance pattern investigated and link to the synthesised sounds; fo = *f*ₒ levels applied in synthesis, in Hz; Δfo = *f*ₒ level differences in reference to the first sound of a series, in semitones, ST, with approximations given in parentheses; F(i) = *F*-patterns investigated). Columns 9 and 10 = overall vowel recognition results (VR, where V = vowel quality, % = recognition rate including results of all subtests). Columns 11–17 = listener-specific details of vowel recognition and subtest-specific vowel qualities with a recognition rate of ≥ 60% (vowel boundaries given in parentheses).
[M-03-04-T01]

M3.4 Source–Filter Synthesis Based on Model Filter Patterns of Half-Open Tubes, Including Variation of *f*ₒ 485

Table 1. Source–filter synthesis based on model filter patterns of half-open tubes, including fo variation: F-patterns and fo variation investigated and vowel recognition results (details). [M-03-04-T01]

Vowel synthesis								VR		Vowel recognition (VR, details)						
SG/	f_o	Δf_o	F_1	F_2	F_3	F_4	F_5	Maj		Subtests	Listeners					Maj
L	Hz	ST	Hz	Hz	Hz	Hz	Hz	V	%		1	2	3	4	5	
men	125	ref	500	1500	2500	3500	4500	ə	93	single	ə	ə	ə	œ	ə	ə
										triplets, *f*o ↑	ə	ə	ə	ə	ə	ə
										triplets, *f*o ↓	ə	ə	ə	ə	ə	ə
	250	12						ø	80	Single	ø	ø	ø	ø	œ	ø
										triplets, *f*o ↑	ø	ø	ø	ø	œ	ø
										triplets, *f*o ↓	ø	ø	ø	ø	œ	ø
	500	24						y	67	Single	øy	ø	øy	y	øy	(ø–y)
										triplets, *f*o ↑	y	y	y	y	y	y
										triplets, *f*o ↓	y	y	y	y	øy	y
women	200	ref	600	1800	3000	4200	5400	ə	93	Single	ə	ə	ə	œ	ə	ə
										triplets, *f*o ↑	ə	ə	ə	ə	ə	ə
										triplets, *f*o ↓	ə	ə	ə	ə	ə	ə
	300	(7)						ø	67	Single	ø	ə	ø	ø	œ	ø
										triplets, *f*o ↑	ø	əø	ø	ø	œ	ø
										triplets, *f*o ↓	ø	ø	ø	ø	ə	ø
	600	(19)						y	60	Single	øy	ø	ø	y	øy	(ø–y)
										triplets, *f*o ↑	y	y	y	y	øy	y
										triplets, *f*o ↓	y	y	y	y	øy	y
children	233	ref	700	2100	3500	4900	6300	ɛ	60	Single	ɛ	ɛ	ə	ɛ	ə	ɛ
										triplets, *f*o ↑	ɛ	ɛ	ə	ɛ	ə	ɛ
										triplets, *f*o ↓	ɛ	ɛ	ə	ɛ	ə	ɛ
	350	(7)						ɛ-œ- ə-ø-øe		Single	øe	əø	əɛ	œ	ə	ɛ-œ- ə-ø-øe
										triplets, *f*o ↑	øe	ə	ø	ɛ	ə	
										triplets, *f*o ↓	øe	ø	ø	ɛ	œ	
	700	(19)						y	73	Single	øy	y	y	øy	y	y
										triplets, *f*o ↑	y	y	y	øy	ø	y
										triplets, *f*o ↓	y	y	y	y	y	y

M3.5 Paradigmatic Examples of Formant Pattern and Spectral Shape Ambiguity in Natural Vocalisations and Their Resynthesised Replicas

Introduction

Neither *F*-patterns nor spectral envelopes were found to be stable within the vowel-related frequency ranges when looking at natural sounds of single vowels produced at different f_o levels, even for sounds produced by a single speaker (see Chapter M2). In a vowel synthesis or resynthesis that was based on one single unchanged *F*-pattern and related LPC filter curve or one single unchanged spectral envelope, depending on the experimental setting, the recognised vowel quality for a substantial portion or even the majority of sounds changed if f_o was substantially altered (see Chapters M3.1 to M3.4). These findings led to the conclusion that *F*-patterns and spectral envelopes *per se* are ambiguous acoustic representations of vowel quality.

In the preceding Chapters M3.1 to M3.4, the ambiguity became evident in synthesis and resynthesis experiments. In the present chapter, the ambiguity is addressed as a core phenomenon of natural vowel sounds. In two earlier studies (Maurer et al., 2000; Maurer et al., 2019, based on the sounds of the first version of the Zurich Corpus) and in the Preliminaries (pp. 64–65 and 187–216), we have already described and documented the ambiguity phenomenon. For the present treatise, the documentation of Maurer et al. (2019) was revisited and renewed, and selected examples of documentation are transferred to this treatise and integrated into the main line of argument.

Experiment

Vowels and speakers: Voiced sounds of the eight long Standard German vowels /i, y, e, ø, ɛ, a, o, u/ and of /ɔ/ produced in V context at various levels of f_o as documented in the Zurich Corpus were investigated, disregarding vocal effort and production style. The vowel /ɔ/ was also included in the investigation in order to examine possible *F*-pattern and spectral shape ambiguity for the non-adjacent back vowels /ɔ/ and /u/. Production, recording and vowel recognition of these sounds accorded to the standard procedures when creating the corpus.

Selection of sound pairs or sound triplets – general procedure: Against the background of earlier studies and the experiments described above, paradigmatic sound series were compiled that provide evidence for formant pattern and spectral shape ambiguity in natural vocalisations.

Sound pairs of two adjacent vowels /e–i/, /ø–y/ and /o–u/ were compiled according to the following criteria:

- Both sounds of a pair were produced by the same speaker.
- For both sounds of a pair, the upper limit of calculated f_o was 400 Hz for men, 450 Hz for women and 500 Hz for children, reflecting the average everyday vocal range of women and children and the chest and “mixed” voice for men.
- A 100% vowel recognition rate matching vowel intention was obtained in the standard listening test conducted when creating the Zurich Corpus.
- Vowel-related peaks of the harmonic spectra, spectral envelopes and estimated *F*-patterns of both sounds were appraised as being comparable; above all, differences in estimated vowel quality-related spectral peaks and estimated *F*-patterns were considered to remain within the commonly assumed range of variation for sounds of a single vowel quality.

For the entire f_o range of recognisable vowel sounds, sound triplets of adjacent and non-adjacent vowels /ɛ–e–i/ or /ɛ–e–y/, /a–o–u/ and /ɔ–o–u/ were compiled according to the following criteria:

- All three sounds were produced by speakers of a single speaker group (children, women or men).
- An 80–100% vowel recognition rate matching vowel intention was obtained in the standard listening test conducted when creating the Zurich Corpus. (Note that for these adjacent and non-adjacent vowels produced by speakers of a single age- and gender-related speaker group, the compilation of sound triplets with similar vowel-related spectral peaks, *F*-patterns and spectral envelopes proved to be much more difficult than merely compiling sound pairs of adjacent vowels; therefore, f_o was not limited and vowel sounds with a vowel recognition rate of 80% were also investigated).
- Vowel-related peaks of the harmonic spectra, spectral envelopes and estimated *F*-patterns of all three sounds were appraised as comparable.

For each comparison of vowel pairs or triplets, several sound pairs or triplets were thus created and further investigated.

Acoustic analysis: Vowel-related *F*-patterns (F_1–F_2–F_3 for sounds of front vowels and F_1–F_2 for sounds of back vowels and /a/) and f_o levels were calculated according to the standard procedure of the Zurich Corpus, applying speaker group-specific parameters. Subsequently, for every single sound, the *F*-pattern was crosschecked on the basis of the spectrum, spectrogram and formant tracks. If a calculated

F-pattern was not confirmed in the crosscheck – above all, if formant tracks were scattered and/or the *F*-pattern did not match the harmonic spectrum and its peaks – the default parameter of the maximum LPC filter number was changed accordingly. The final LPC filter values applied are shown in Tables 1 and 2 in the chapter appendix, with six, five or four formants at a maximum for a frequency range of 5.5 kHz (indicated as default or P6, P5 and P4, according to the parameters of Zurich Corpus; see the Introduction). Exceptions occurred and are discussed in the results section when presenting the corresponding sounds.

Spectral comparison: The similarity of vowel-related spectral peaks and spectral envelopes of the sounds of a pair or a triplet were investigated by the author in a direct visual comparison of the acoustic analysis results and their graphic illustration. In general, patterns of spectral peaks and/or estimated *F*-patterns and/or spectral envelopes were considered similar if their differences were estimated as smaller than the differences commonly attributed to the spectra of two adjacent vowels. In the results section, this knowledge-based criterion is objectivated by means of numerical indications.

For sounds of front vowels, the spectral comparison was related to a frequency range of up to 3 kHz for adults and up to 3.5 kHz for children. For sounds of back vowels and /a/, the spectral comparison was related to a frequency range of up to 2 kHz for all speakers.

For one sound of /ɛ/, none of the three parameter settings produced F_1 values matching the lower spectral peak, and the peak frequency was assessed in a visual examination of the harmonic spectrum (see Table 2, Series 1). For two sounds of /e/ and /i/ (see Table 2, Series 3), the calculated F_3 values were somewhat low and spectral similarity was again assessed in a visual examination of the harmonic spectrum and its peaks. Furthermore, as an exception to the general procedure of *F*-pattern estimation, for one triplet of sounds of back vowels with only a single spectral peak < 2 kHz, the vowel-related spectral peak frequency was assessed by means of a visual examination of the harmonic spectrum only (see Table 2, Series 7).

First sound compilation and subsequent selection of sound pairs and triplets for documentation: The first sound sample compiled consisted of several sound comparisons related to the vowel pairs and triplets investigated. For exemplary documentation in this treatise, the entire sample was reduced: For each of the pairs of the adjacent vowels /e–i/, /e–y/ and /o–u/ and each of the triplets of the adjacent and non-adjacent vowels /ɛ–e–i/ or /ɛ–e–y/, /a–o–u/ and /ɔ–o–u/, three sound comparisons were selected, resulting in a total of nine sound pairs and nine sound triplets of natural sounds or 45 sounds in total.

Resynthesis: All natural sounds in this final selection were resynthesised as steady-state sounds of 1 sec. using the Klatt synthesiser as implemented in the Praat software (cascade mode, sampling frequency = 44.1 kHz, fade in/out = 0.05 sec.): For every single sound of a pair or triplet and the related estimated *F*-pattern as given in Tables 1 and 2 (including also the higher formants of the corresponding LPC analysis not given in the table), the f_o level of that sound and the f_o level(s) of the opposing sound(s) were applied for the resynthesis. As a result, four synthesised sounds were produced for each pair of natural sounds, and nine synthesised sounds were produced for each natural sound triplet, resulting in a total of 117 resynthesised sounds.

The main idea of this experimental design was to validate the estimated similarity of the spectral peaks, spectral envelopes and estimated *F*-patterns of the natural sounds that stand in comparison: For opposing natural sounds produced at very different f_o levels, the literature does not stipulate an objective criterion for assessing the similarity of spectral peak patterns, *F*-patterns and spectral envelopes in vowel-related frequency ranges. Nevertheless, suppose a vowel quality of a natural reference sound is maintained in a resynthesis that is based on both its estimated LPC curve and its average f_o level. In this case, the maintained vowel quality can be considered a possible validation criterion for the LPC curve and, thus, for the related estimated *F*-pattern and spectral envelope. Moreover, suppose an increase or decrease of the f_o level applied in resynthesis causes a vowel quality shift while keeping the LPC curve unchanged. In that case, the LPC curve is indicated to be an ambiguous representation of vowel quality. Finally, suppose the same vowel quality is recognised for the two or three sounds of a sound pair or triplet, the sounds resynthesised based on the two or three LPC curves but applying equal f_o levels. In that case, vowel recognition validates an assessment of spectral peak pattern, *F*-pattern and spectral envelope similarity to the natural reference sounds of different vowels that stand in comparison.

Listening test: As mentioned, the vowel qualities of the original natural sounds presented in this chapter were assessed within the standard listening test procedure when building up the Zurich Corpus, involving the five standard listeners. In addition, in a separate listening test involving the five standard listeners of the Zurich Corpus, vowel recognition of the resynthesised sounds was investigated according to the standard procedure of the corpus, except for the condition of sound presentation: Instead of featuring single sounds, each test item consisted of two sounds, the natural reference sound (first sound) and one of the resynthesised replicas of the sound pair or triplet (second sound), resulting in two test items per pair and three items per triplet. The two sounds of a test item were presented (separated by a 1 sec. pause), and the listeners were asked to assign the recognised vowel quality of the second sound only. All sounds investigated were presented in one listening test (no speaker-blocked test condition).

Results

Tables 1 (sound pairs) and 2 (sound triplets) in the chapter appendix show the sound comparisons and the results of acoustic analysis and vowel recognition test for the resynthesised sounds.

Natural reference sounds – vowel recognition: For all natural reference sounds, the vowel recognition rate (matching with intended vowel quality, see Tables 1 and 2, Column 4) was 100%. The only exception concerns a sound of /e/ with a rate of 80% (see Table 2, Series 3, sound link).

Sound pairs of adjacent vowels – similarity of the *F*-patterns of the natural sounds compared: For all natural sounds of the sound pairs (see Table 1), the F_1 difference between the two estimated *F*-patterns was < 35 Hz, and the F_2 difference was < 100 Hz, except for one pair (see Series 1). The F_3 difference for the pairs of front vowel sounds was < 140 Hz. However, for all pairs of front vowel sounds, the difference in either F_2 or F_3 was < 60 Hz.

Sound pairs of adjacent vowels – vowel recognition of the resynthesised replicas based on the *F*-patterns and f_o levels of their respective natural reference sounds: According to the vowel recognition results for the resynthesised sounds based on the estimated *F*-patterns and f_o levels of their respective natural reference sounds, ignoring unrounded–rounded confusions, a recognition rate of ≥ 80% matching vowel intention or vowel openness was found for all sounds (see Table 1, VR of resynthesis with f_o unchanged). Thus, a vowel resynthesis based on estimated *F*-patterns and LPC curves of the natural sounds, applying their average f_o level, did not affect vowel recognition substantially. This result supported the validity of the estimation and comparison of the *F*-patterns of the natural reference sounds as given in Table 1.

Sound pairs of adjacent vowels – vowel recognition of the resynthesised replicas with f_o variation: As for the vowel recognition of the resynthesised sounds based on a single estimated F-pattern but applying the two different f_o levels of a sound comparison, an increase of f_o resulted in a close-mid–close vowel quality shift and, vice versa, a decrease of f_o resulted in a close–close-mid shift, with shift recognition rates ≥ 80%. If a general perspective of increasing f_o from a lower to a higher level is adopted, thus, a general open–close shift direction with increasing f_o levels was found (see Table 1, VR of resynthesis with f_o unchanged and exchanged).

Sound triplets of adjacent and non-adjacent vowels – similarity of the F-patterns of the natural sounds compared: For all natural sounds of the sound triplets (see Table 2), the F_1 difference (or the difference of the single peak frequencies, see Series 7) between the three F-patterns was < 65 Hz, the F_2 difference was in the range of 28–146 Hz for the triplets of sounds of /a/ or /ɔ/ compared with /o/ and /u/, and the differences in F_2–F_3 for the three triplets of front vowel sounds were 22–16 Hz, 214–145 Hz and 170–144 Hz, respectively.

Sound triplets of adjacent and non-adjacent vowels – vowel recognition of the resynthesised replicas based on the F-patterns and f_o levels of their respective natural reference sounds: According to the vowel recognition results for the resynthesised sounds based on the estimated F-patterns and f_o levels of their respective natural reference sounds, ignoring unrounded–rounded confusions, a recognition rate of ≥ 80% matching vowel intention or vowel openness was found for eight of the nine sounds of /ɛ, e, i/ and /ɛ, e, y/, for four out of the nine sounds of /a, o, u/ (if /a/ and /aɔ/ boundaries are taken as quasi-equal in vowel openness) and for eight out of the nine sounds of /ɔ, o, u/ (if /a/ and /aɔ/ boundaries and back–front or front–back confusions with unchanged vowel openness are included in the estimate of vowel openness). For all other sounds, the identification rate for a corresponding match was 60% (see Table 2, VR of resynthesis with f_o unchanged).

Sound triplets of adjacent and non-adjacent vowels – vowel recognition of the resynthesised replicas with f_o variation: As for the vowel recognition for the resynthesised sounds based on a single estimated F-pattern but applying the three different f_o levels of sound comparison, an increase in f_o from the lowest to the highest level of comparison resulted in a shift from an open-mid to a close vowel quality for all sounds and all F-patterns of /ɛ, e, i/ and /ɛ, e, y/, with shift recognition rates ≥ 80% (see Table 2, VR of resynthesis with f_o exchanged and the resulting entire shift). This also held true for seven out of the nine *F*-patterns of /ɔ, o, u/ (if back–front confusions and corresponding vowel openness are included). For the sounds of the two remaining patterns, the identification rate for the corresponding shifts dropped to 60%.

Concerning the /a, o, u/ sound triplets, shifts from an open or an open-mid to a close vowel quality were found for six out of the nine *F*-patterns, with a shift recognition rate of 60–80%. For the remaining patterns, shifts from /a/ to the vowel boundary of /u/ and /o/, from /a/ to /o/ and /a/ to /ɔ/ were found, with 60–100% recognition rates.

Occurring back–front confusion in resynthesis: One listener recognised five resynthesised sounds related to the *F*-patterns of the back vowels /ɔ, o, u/ in Series 7 and one sound of the vowel triplet in Series 8 as front vowels (see Table 2, extended online version). For the sounds in Series 7, this may be mainly due to the specific spectral shape of the natural reference sounds manifesting only one distinct spectral peak below 2 kHz. This kind of spectral shape, often found for sounds of back vowels (see Chapters M7.1 and M7.2; see also the discussion on the matter in the Preliminaries), resembles the spectral shapes of natural sounds of front vowels. In addition, the extensive f_o variation also has to be considered.

For replication and crosschecking of the resynthesised sounds, use the Klatt synthesiser integrated into the Zurich Corpus, relating to the LPC parameters as given in Tables 1 and 2.

Discussion

In this experiment, natural sound pairs of adjacent vowels and natural sound triplets of adjacent and non-adjacent vowels were compiled for which the estimated *F*-patterns and the visual inspection of the spectral envelopes and spectrograms of comparison were appraised to be similar for the vowel-related frequency range.

Concerning the spectral similarity of sound pairs and sound triplets, firstly, the corresponding evaluation was based on a direct comparison of *F*-patterns, spectrograms, harmonic spectra and spectral peaks of the sounds in question, and the links to the sounds and their spectra provided in the tables allow for a re-examination of our appraisal. Secondly, the numerical differences of the estimated formant frequencies, listed in the results section of this chapter, supported the notion of spectral similarity: Above all, the F_1 differences for sounds of the two or three different vowels compared were markedly smaller than the differences of statistical F_1 of these two or three vowels generally given in the literature. Further, if the actual range of F_2 and F_3 variation occurring for sounds of a single vowel quality produced at a given, single f_o level is taken into account (for exemplary illustration, see Peterson and Barney, 1952, Figure 8), the differences of the entire F_1–F_2–F_3 patterns were found to be within a variation range of sounds of one of the vowels of comparison. Finally, resynthesis also confirmed the estimated similarity, above all for the sound pairs of close-mid and close vowels and the sound triplets of open-mid, close-mid and close vowels: For these sounds and vowels, in a resynthesis based on the estimated *F*-patterns and calculated f_o of the natural reference sounds, the recognition rate for vowel quality or vowel openness, matching vowel intention, was ≥ 80%, with only two exceptions for which the rate dropped to 60%, as was also the case for the majority of the sounds of the /a, o, u/ triplets. In contrast, in a resynthesis based on the estimated *F*-patterns of the natural reference sounds but varying f_o in terms of applying all two or three levels of a sound pair or triplet and then testing vowel recognition of the resynthesised replicas, recognised vowel qualities shifted for almost all replicas of the natural reference sounds investigated, with the shifts being very pronounced for the sound pairs and triplets of close-mid and close and of open-mid, close-mid and close vowels. Again, a general open–close shift direction resulting from an increase in f_o from a low to a high level was found.

In these terms, the sound comparisons of natural sounds presented in this study once again provide exemplary evidence for the formant pattern and spectral shape ambiguity, this ambiguity being interpreted here as a core phenomenon of the acoustics of the vowel. Note in this context that the ambiguity shown for adjacent vowels concerned sounds produced by a single speaker and that the ambiguity shown for adjacent and non-adjacent vowels concerned sounds produced by speakers of a single speaker group (women or men). This limitation of sound production strongly indicates that the ambiguity shown directly resulted from f_o variation. Note also that the ambiguity could be demonstrated despite the static character and mediocre sound quality of the synthesised sounds produced with the Klatt synthesiser.

The somewhat weaker support for the ambiguity thesis found for the /a, o, u/ sound triplets may be primarily due to the nonuniform character of the vowel spectrum: As discussed in Chapter M2.1, no marked indication of a relation of the lower vowel spectrum to f_o was found for sounds of /a/, in strong contrast to sounds of close and close-mid vowels. Correspondingly, vowel quality shifts in vowel synthesis based on statistical *F*-patterns but including f_0 variation were rare for sounds of /a/ (see Chapter M3.1), as was true for shifts in vowel resynthesis based on the spectral envelopes of natural vowel sounds including f_0 variation (see Chapter M3.2). These findings stood in strong contrast to the findings for sounds of all other vowels.

In addition to the vowel quality recognition results discussed above, a few cases of back–front confusion also occurred in this experiment, as was the case for previous synthesis and resynthesis experiments. We interpret these cases in the context of the vowel sound being a foreground–background phenomenon (for details and further discussion, see Chapter M8).

Differences regarding style and vocal effort when producing natural sounds were ignored. However, as the vowel recognition results of resynthesis indicated, these differences did not relativise the general finding of the ambiguity phenomenon.

Chapter appendix

Due to sound editing and recalculation of the patterns when updating the corpus, marginal differences between the values for *F*-patterns given in the two tables and the corresponding values given in the online corpus may occur (see the Introduction). The values in the two tables correspond to the calculated values used for resynthesis at the time of investigation. With one exception, occurring differences are ≤ 5 Hz and are neglectable. The exception concerns F_2 of the first sound of Series 1 in Table 2, the difference being 13 Hz. Further, a 1 Hz difference in f_o occurs between the value given in Table 1 for the first sound of Series 9 and the value given in the online corpus.

Table 1. Formant pattern and spectral shape ambiguity in natural vocalisations: Sound pairs investigated and results of acoustic analysis and vowel recognition. Columns 1–6 = natural reference sounds (S/L = sound pairs and sound links; SP = speaker ID in the Zurich Corpus; SG = speaker group, where c = children, w = women, m = men; V = intended and recognised vowel quality; PS = production style; VE = vocal effort). Columns 7–14 = results of acoustic analysis (fo = calculated f_o, in Hz; F(i) = formant frequencies, in Hz; ΔF(i) = formant frequency differences, in Hz; Par = parameter setting of LPC analysis according to the procedure of the Zurich Corpus, where def = speaker group-related default setting, and P4, P5, P6 = applied settings overriding the default). Columns 15–17 = vowel recognition results for the resynthesised sounds based on the *F*-patterns and f_o levels of their respective natural reference sounds (fo unchanged = f_o level applied, in Hz; V = recognised vowel quality according to the labelling majority; Maj = labelling majority, in %). Columns 18–21 = vowel recognition results for the resynthesised sounds based on the *F*-patterns of their respective natural reference sounds but applying the levels of the opposing natural sounds of a pair, and the resulting vowel quality shifts with increasing f_o from the lower to the higher level of comparison (fo = exchanged f_o level applied, in Hz; V = recognised vowel quality according to the labelling majority; shift = resulting vowel quality shift when compared with the resynthesis at unchanged f_o; Maj = labelling majority for the vowel quality shift, in %). Columns 22ff. = details of the vowel recognition results. Colour code: Blue = recognised vowel quality matching vowel intention of the natural reference sound for resynthesis based on its *F*-pattern and unchanged f_o level; red = recognised vowel quality mismatching vowel intention of the natural reference sound for resynthesis based on its *F*-pattern but with exchanged f_o, in terms of a vowel quality shift in an open–close direction with increasing f_o.
[M-03-05-T01]

Table 2. Formant pattern and spectral shape ambiguity in natural vocalisations: Sound triplets investigated and results of acoustic analysis and vowel recognition. Columns 1–17 = see Table 1. Columns 18–25 = vowel recognition results for the resynthesised sounds based on the *F*-pattern of their respective natural reference sounds but applying the levels of the opposing natural sounds of a triplet, and the resulting vowel quality shift with increasing f_o from lower to higher levels of comparison. (Note that Columns 24 and 25 show the vowel quality shifts for the qualities with the greatest difference in vowel openness.) Column specification accords to Table 1 (note F(i) given in parentheses = assessed in a visual examination of the harmonic spectrum; M as parameter setting for LPC analysis = manual assignment). Extended online table: Columns 26ff. = details of the vowel recognition results (note vowel qualities in parentheses and marked with "*" for occurring back–front confusions; for details see text). Colour code = see Table 1.
[M-03-05-T02]

Tabelle 1. Formant pattern and spectral shape ambiguity in natural vocalisations: Sound pairs investigated and results of acoustic analysis and vowel recognition. [M-03-05-T01]

Natural reference sounds						Acoustic analysis								VR of resynthesis									VR resynthesis (details)			
S/L	SP	SG	V	PS	VE	fo	F1	ΔF1	F2	ΔF2	F3	ΔF3	Par	fo unchanged			fo exchanged					fo unchanged		fo exchanged	
						Hz	Hz	Hz	Hz	Hz	Hz	Hz		fo	V	Maj	fo	V	Maj	shift	Maj	fo	V	fo	V
1 ↗	1052	w	e	N	low	167	322	1	2499	139	3158	27	def	167	e	100	310	i	100	e – i	100	167	e e e e e	310	i i i i i
			i	ST	med	310	323		2360		3131		def	310	i	100	167	e	80	e – i	80	310	i i i i i	167	e e e e ei
2 ↗	1033	m	e	N	med	165	339	33	2378	15	2730	69	def	165	e	80	346	i-y	100	e – i-y	80	165	ei e e e e	346	i i i i iy
			i	N	high	346	306		2393		2799		P5	346	y–i	80	165	e	80	e – i-y	80	346	i i y y e	165	i e e e e
3 ↗	1056	c	e	N	med	217	455	26	3073	82	3701	23	def	217	e	80	429	i-y	100	e – i-y	80	217	e e e e o	429	i i i i y
			i	N	med	429	429		2991		3724		def	429	i	100	217	e	100	e – i	100	429	i i i i i	217	e e e e e
4 ↗	1002	m	ø	N	high	164	335	13	1696	56	2289	54	def	163	ø	80	330	y	100	ø – y	80	164	ø ø ø ø øy	330	y y y y y
			y	CS	med	330	322		1640		2343		def	330	y	100	164	ø	80	ø – y	80	330	y y y y y	164	yø ø ø ø ø
5 ↗	1066	w	ø	N	med	221	438	16	1687	31	2710	94	def	221	ø	100	423	y	100	ø – y	100	221	ø ø ø ø ø	423	y y y y y
			y	N	med	423	422		1656		2616		def	423	y	80	221	ø	100	ø – y	80	423	y y y y u	221	ø ø ø ø ø
6 ↗	1091	c	ø	N	med	226	453	17	1790	55	2599	133	def	226	ø	100	436	y	100	ø – y	100	226	ø ø ø ø ø	436	y y y y y
			y	N	med	436	436		1735		2732		def	436	y	100	226	ø	100	ø – y	100	436	y y y y y	226	ø ø ø ø ø
7 ↗	1005	w	o	N	med	215	409	17	775	32	–	–	def	215	o	100	401	u	100	o – u	100	215	o o o o o	401	u u u u u
			u	N	med	401	392		807		–		def	401	u	100	215	o	100	o – u	100	401	u u u u u	215	o o o o o
8 ↗	1042	m	o	N	low	195	380	14	753	92	–	–	def	195	o	100	392	u	80	o – u	80	195	o o o o o	392	o u u u u
			u	EC	low	392	394		845		–		P5	392	u	100	195	o	80	o – u	80	392	u u u u u	195	o o o o uo
9 ↗	1036	w	o	N	med	186	370	28	743	60	–	–	def	186	o	80	395	u	100	o – u	80	186	o o o o ou	395	u u u u u
			u	N	low	395	398		803		–		def	395	u	80	186	o	80	o – u	80	395	ou u u u u	186	o o o o ou

Tabelle 2. Formant pattern and spectral shape ambiguity in natural vocalisations: Sound triplets investigated and results of acoustic analysis and vowel recognition. [M-03-05-T02]. Extended online Table: ↗

Natural ref. sounds					Acoustic analysis									VR of resynthesis										
S/L	SG	V	PS	VE	fo	F1	SP	ΔF1	F2	ΔF2	F3	ΔF3	Par	fo unchanged			fo exchanged							
					Hz	Hz	Hz	Hz	Hz	Hz	Hz	Hz		fo	V	Maj	fo	V	Maj	fo	V	Maj	shift	Maj
		ɛ	CS	low	99	(500)			2120		2731		def	99	ɛ	80	263	e	100	539	i-y	100	ɛ – i-y	80
1 ↗	m	e	N	med	263	512		50	2142	22	2744	16	P5	263	e	100	539	i-y	80	99	ɛ	80	ɛ – i-y	80
		i	ST	med	539	550			2123		2728		def	539	i-y	100	99	ɛ	100	263	e	100	ɛ – i-y	100
		ɛ	N	med	165	526			2246		2867		def	165	ɛ	100	326	e	100	547	i-y	80	ɛ – i-y	80
2 ↗	w	e	CS	high	326	581		55	2347	214	2800	145	def	326	e	100	547	i-y	80	165	ɛ	100	ɛ – i-y	80
		y	N	high	547	546			2133		2722		def	547	i-y	100	165	ɛ	100	326	e-ø	100	ɛ – i-y	100
		ɛ	N	high	256	728			2459		3247		def	256	ɛ	100	359	e	60	663	i	80	ɛ – i	80
3 ↗	w	e	N	high	359	706		61	2509	170	(3162)	(144)	def	359	e	60	663	i	100	256	ɛ	100	ɛ – i	100
		i	N	low	663	667			2629		(3103)		def	663	i	100	256	ɛ	100	359	e	100	ɛ – i	100
		a	N	med	107	626			1083		–		def	107	a	100	327	o	60	595	conf	–	a – o	60
4 ↗	m	o	N	high	327	653		63	1082	41	–	–	def	327	o	60	595	u-o	100	107	a	100	a – u-o	100
		u	N	med	595	590			1123		–		def	595	u	60	107	a-ɔ	100	327	o	80	a-ɔ – u	60
		a	N	med	131	599			1036		–		def	131	a-ɔ	100	291	ɔ-o	80	575	u	60	a-ɔ – u	60
5 ↗	m	o	CS	high	291	583		23	1108	103	–	–	def	291	o	80	575	u	60	131	a-ɔ	100	a-ɔ – u	60
		u	N	med	575	576			1139		–		def	575	u	60	131	a-aɔ	100	291	o	60	a-aɔ – u	60
		a	N	med	163	624			1081		–		def	163	a	60	300	ɔ	60	622	conf	–	a – ɔ	60
6 ↗	w	o	N	high	300	602		22	1121	66	–	–	def	300	o	60	622	u	60	163	a	80	a – u	60
		u	N	high	622	624			1147		–		def	622	u	80	163	a	80	300	o	60	a – u	80

Tabelle 2 (continuation). [M-03-05-T02]

Natural ref. sounds					Acoustic analysis									VR of resynthesis										
S/L	SG	V	PS	VE	fo	F1	SP	ΔF1	F2	ΔF2	F3	ΔF3	Par	fo unchanged			fo exchanged							
					Hz	Hz	Hz	Hz	Hz	Hz	Hz	Hz		fo	V	Maj	fo	V	Maj	fo	V	Maj	shift	Maj
		ɔ	N	med	80	–	500		–		–		M	80	ɔ (ε)	100	250	o (e)	100	508	u (i)	100	ɔ – u (ε – i)*	100*
7 ↗	m	o	N	high	250	–	500	0	–	–	–	–	M	250	o (e)	100	508	u (i)	80	80	ɔ	100	ɔ – u (i)*	80*
		u	N	med	508	–	500		–		–		M	508	u	100	80	a-ɔ	100	250	o	100	a-ɔ – u	100
		ɔ	N	med	89	493	–		823		–		def	89	a-ɔ	100	238	o	80	495	u (i)	80	a-ɔ – u (i)*	80*
8 ↗	m	o	N	high	238	479	–	15	940	146	–	–	def	238	o	80	495	u	80	89	ɔ	60	ɔ – u	60
		u	N	low	495	494	–		969		–		def	495	u	60	89	ɔ	80	238	o	100	ɔ – u	60
		ɔ	ST	med	160	497	–		981		–		def	160	ɔ	100	253	o	100	488	u	80	ɔ – u	80
9 ↗	w	o	N	high	253	496	–	7	1009	28	–	–	def	253	o	100	488	u	100	160	ɔ	100	ɔ – u	100
		u	EC	high	488	490	–		988		–		def	488	u	100	160	ɔ–ɔo	80	253	o	100	ɔ-ɔo – u	80

M4 Vowel Spectrum and Age and Gender of the Speakers

M4.1 Similar Lower Spectral Peak Frequencies and Estimated Formant Frequencies for Vowel Sounds Produced by Children, Women and Men at a Similar f_o

Introduction

Vowel quality-related *F*-patterns, as given in formant statistics for voiced vowel sounds, differ according to the age and gender of the speakers: Generally, statistical average *F*-patterns were reported as highest for children, intermediate for women and lowest for men. These differences are commonly understood to be primarily due to the different average vocal tract sizes of these age- and gender-related groups. (For some relativisations, see the Preliminaries, pp. 114–117.)

However, to the best of our knowledge, almost all studies of statistical average *F*-patterns for voiced sounds of men, women and children were conducted under laboratory conditions, with the investigated vowel sounds produced with medium vocal effort in isolation at f_o levels comparable to those of relaxed speech (sounds produced in V context or extracted from citation-form words or from read sentences). Most importantly, such f_o levels fall into a lower vocal range of the speakers in general and into a lower f_o range of recognisable vowel sounds in particular (see Chapter M2). Therefore, existing formant statistics do not provide empirical evidence of whether or not supposed age- and gender-related spectral differences remain if f_o is substantially varied. (Note also the possible additional effect of vocal effort variation on estimated *F*-patterns.) In this context, two types of comparisons are of specific interest: Firstly, vowel sounds produced by speakers different in age and gender at similar f_o levels; secondly, vowel sounds produced by adults at f_o levels that are higher than the levels of those produced by children, and vowel sounds produced by men at f_o levels that are higher than the levels of those produced by women. The first aspect is addressed in this chapter; the second will be addressed in the following chapter in terms of children–adult sound comparisons.

In an earlier study, Maurer et al. (1991; see also Chapter M3.1) investigated average *F*-patterns of men, women and children producing sounds of the long Swiss German vowels /i, e, a, o, u/ at f_o levels of approximately 131 Hz (men only), 175–220 Hz (men and women) and 262 Hz (all speakers; approximate levels given according to musical C-major scale and further adapted, as noted in Chapter M3.1). According to the results, when the sounds were produced by men at f_o of c. 131 Hz, by women at f_o of c. 220 Hz and by children at f_o of c. 262 Hz, the estimated F-patterns differed comparably to the differences generally reported in the literature. However, when men and women produced the sounds at the same levels of f_o of 175–220 Hz, and when adults and children produced the sounds at the same level of f_o of 262 Hz, (i) differences in F_1 disappeared for the sound comparison of all three speaker groups and all five vowels, (ii) differences in F_2 also disappeared for the sound comparison of men and women for the patterns of back vowels and /a/, as was true for the sound comparison of adults and children for the vowel /u/, and (iii) differences in F_2 substantially decreased for the sound comparison of adults and children for the vowels /o/ and /a/. Interpreting these results, we concluded that the variation of f_o has a much stronger effect on the vowel-related lower formants < 1.5–2 kHz than the assumed average vocal tract size of the speakers does. In contrast to the results for lower formants, the results for formants above 2 kHz for front vowels were difficult to interpret. However, in that study, we found a tendency for the frequency values of the vowels /e/ and /i/ to be higher for children than for adults and higher for women than for men.

During the creation of the Zurich Corpus, we revisited the matter based on viewing the spectral characteristics of vowel sounds produced by men, women, and children with regard to three conditions of comparison concerning the f_o levels of the sounds: (i) f_o levels comparable to the levels as given in formant statistics, that is lower levels for adults than for children and lower levels for men than for women; (ii) similar f_o levels for men and women and similar levels for adults and children; (iii) higher f_o levels for adults than for children and higher levels for men than for women. In a conference paper, we presented observational results and exemplary sound series (Maurer and Landis, 2015): For the first condition, with the sounds produced at intended f_o of 131 Hz for men, 220 Hz for women and 262 Hz for children, F_1 and F_2 again were found to differ according to the differences reported in the literature on statistical average F-patterns. However, exceptions occurred, especially for sounds of /i/, /y/ and /u/. For the second condition, with f_o of the sounds at c. 220 Hz for both men and women and with f_o of the sounds at c. 262 Hz for both adults and children, we observed in many direct sound comparisons a decrease or even a disappearance of the expected speaker-group differences in the formant frequencies < 1.5kHz. (The observational results for the third condition are discussed in the next chapter.)

In the Preliminaries (see pp. 217–237), we have documented these observations in more detail in terms of spectral comparisons of sounds of the vowels /i, e, a, o, u/ produced by a man, a woman and a child at various levels of f_0 and with similar upper f_0 ranges for all three speakers. The documentation of the spectra showed in its turn that age- and gender-related spectral peak and formant frequency differences ≤ 1.5–2 kHz, as generally given in formant statistics, decreased or even disappeared if f_0 of the vocalisations corresponded to the sounds of the three speakers different in age or gender. (Documented cases of sounds produced at higher f_0 levels by adults than children and at higher levels by men than women are also discussed in the next chapter.)

During the creation of the Zurich Corpus, we have also made an attempt at a statistical analysis of age- and gender-related differences in *F*-patterns on the basis of a large sample of sounds of the eight long Standard German vowels produced in V context by men, women and children at various f_0 levels up to 523 Hz. This approach was taken despite our awareness of the general methodological problem of formant estimation (see Bladon, 1982; Hillenbrand et al., 1995; Preliminaries, pp. 47–51 and 118–126). Indeed, we were not then able to create an objective formant estimation that would generate reliable values for a large sound sample that included f_0 levels up to 523 Hz, not only because of the f_0 variation but also because of general difficulties in applying the method of crosschecking, correcting and validating formant tracks based on the spectrogram, as this method is given as a reference in the literature (see Hillenbrand et al., 1995). However, this “failure” of investigation also casts doubt concerning the interpretation of existing statistical formant patterns as providing evidence for the lower vowel spectrum *generally* being related to the age or gender of the speakers, independent of the f_0 levels of the sounds.

For the present course of argument, the documentation given in the Preliminaries was renewed and extended based on the new Zurich Corpus. It is integrated here in this treatise, including sound playback functionality: Firstly, based on the vocalises presented and discussed in Chapter M2.1, vowel sounds produced by the three speakers at an intended f_0 range of 220–330 Hz manifesting similar spectral characteristics < 1.5 kHz were selected; secondly, based on the entire Zurich Corpus, vowel sounds produced by men, women and children at an intended f_0 of 262 Hz were selected which also manifested similar spectral characteristics < 1.5 kHz.

Experiment 1

Sound selection and sound comparison: From the sample of vocalises shown in Chapter M2.1, for each of the eight long Standard German vowels, one sound of the child, one of the woman and one of the man produced at f_o within a range of 220–330 Hz was selected, the three sounds of the three speakers manifesting similar spectral peak frequencies (indicated by similar prominent harmonics) and similar estimated formant frequencies < 1.5 kHz. All sounds were fully recognised in the standard listening test conducted when creating the corpus (100% recognition rate matching vowel intention). As a result, a sample of 24 sounds in total was created.

The higher part of the vowel spectrum relevant for front vowels was only investigated with regard to first indications for possible similarities or even inversions of expected differences in formant frequencies comparing sounds of children with those of adults or sounds of women with those of men. However, this aspect is discussed in more detail in the next chapter.

Formant frequency estimation: F_1–F_2–F_3 patterns for sounds of front vowels and F_1–F_2 patterns for sounds of back vowels and /a/ were estimated, commonly considered vowel-related. Formant frequency calculation accorded to the standard procedure of the Zurich Corpus (see the Introduction). In addition, the calculated frequency values were visually crosschecked based on the harmonic spectrum, spectrogram and formant tracks of the sounds. If, for a sound, the formant tracks calculated for age- and gender-specific standard parameter settings of LPC analysis did not match the harmonic spectrum and the spectrogram, the parameter setting was changed (for this procedure, see again Hillenbrand et al., 1995). If no match was found for all three parameter settings investigated for a higher formant ($F2$ and/or $F3$), no value was taken except for one sound, which manifested a clear spectral peak structure. For this sound, the higher formants were estimated based on the harmonic spectrum and the spectrogram (see the sound of /y/ of the child, the higher formant frequencies given in parentheses).

Estimation of similarity of formant frequencies: For all sounds of all vowels investigated, a range of F_1 differences < 70 Hz was considered marginal, and the F_1 values compared within this range were appraised as similar. The same held true for F_2 of the sounds of back vowels and /a/. This setting was based on the following consideration: The statistical formant frequency differences for sounds of different vowels generally exceed 70 Hz, even within the same speaker group (see e.g.

Hillenbrand et al., 1995), and occurring formant frequency differences for sounds of a given vowel produced within an f_o range of 220–330 Hz (the frequency range of f_o investigated in this experiment) also exceed 70 Hz, as can be verified for the sounds documented in the Zurich Corpus.

Results 1

Table 1 in the chapter appendix shows the compilation of the sound series and estimated *F*-patterns. The range of ΔF_1 and ΔF_2 are given in Columns 6 and 8. If the F_1 or F_2 difference of only two of the three sounds of a vowel fell within the range of 70 Hz, in the table, the F_1 or F_2 levels of the remaining sound are indicated as higher or lower than the levels of the other sounds. For the sounds of /o/, F_2 varied strongly, and the levels are indicated as low, middle and high. Similar estimated formant frequencies are coloured green. Note that only the lower vowel spectrum < 1.5 kHz was the focus of investigation.

According to the results of acoustic analysis, for all sounds of all vowels, F_1 of the sounds of the woman was either similar to F_1 of the sounds of the child or higher (see the sounds of /a/), and F_1 of the sounds of the man was similar to F_1 of the sounds of the child, except for the sounds of /ɛ/. Thus, F_1 of the sounds of the woman and the man were similar for the six vowels /i, y, e, ø, o, u/. For the sounds of the vowels /a, o/, F_2 for the child was either lower than those of the two adult speakers or lower than the F_2 of the man. (Note that, for the sounds of the vowel /o/, the F_2 value for the man was higher than for the woman.) For the sounds of the vowel /u/, F_2 values were comparable for all three speakers. In this regard, the spectra of the sounds compared in this first experiment did not indicate a general age- and gender-related difference in estimated spectral peaks and formant frequencies < 1.5 kHz.

Besides, the estimated F_2 of the sounds of front vowels produced by the woman exceeded the F_2 of the corresponding sounds of the child in three cases (see Table 1, sounds of /i, ø, ɛ/ with the corresponding values marked with "*"), and the estimated F_2 of a sound of a front vowel produced by the man corresponded to the F_2 of the sound of the woman in one case (see Table 1, the sound of /y/ with the corresponding value marked with "*").

Experiment 2

Sound selection and sound comparison: Based on the Zurich Corpus, for each of the eight Standard German vowels /i, y, e, ø, ɛ, a, o, u/ and sounds produced in V context with medium vocal effort in nonstyle mode at intended f_o of 262 Hz, two sounds of children, two sounds of women and two sounds of men were selected which manifested similar spectral peak frequencies (indicated by similar prominent harmonics) and similar estimated formant frequencies < 1.5 kHz and which were fully recognised in the standard listening test conducted when creating the corpus (100% recognition rate matching vowel intention). The same procedure as in the previous experiment was applied to the higher part of the vowel spectrum. As a result, a second sample with a total of 48 sounds was created.

Formant frequency estimation: Formant frequency estimation accorded to the procedures described in experiment 1. Note that for one sound of /i/ produced by a child, F_3 was manually assigned according to the third peak in the spectrum (see the value given in parentheses in Table 2). Further, for one sound of /y/ and one sound of /ɛ/ produced by men, F_3 could not be estimated.

Estimation of similarity of formant frequencies: The estimation of similarity of formant frequencies accorded to the procedures described for experiment 1.

Results 2

Table 2 in the chapter appendix shows the compilation of the sound series and estimated F-patterns, the table corresponding to the structure of Table 1 in the chapter appendix. Except for five single sounds, the estimated F_1 values for all sounds of all vowels and all speakers were within a range of < 70 Hz. The five sounds for which F_1 fell outside that range did not indicate general age- and gender-related differences. Therefore, they are considered here as demonstrating possible frequency variations independent of age and gender. Concerning the sounds of /a/ and /o/, as stated, F_2 values varied strongly. However, again, no general age- or gender-related differences were indicated. For the sounds of /u/, F_2 values were comparable for all speakers except for one sound of a man, for which markedly higher F_2 values were found than for the sounds of children and women. In these terms, for the sounds compared in this second experiment, the sound spectra again did not indicate a general age- and gender-related difference in estimated spectral peaks and formant frequencies < 1.5 kHz. Besides, for a few sounds of front vowels, similar estimated F_2 values for adult speakers or higher F_2 for men than for women also occurred (see Table 2, values marked with "*"). Furthermore, for five of the six sounds of /ø/, similar estimated F_2 for the speakers of all three speaker groups or higher F_2 for adults than for children (including one sound with higher F_2 for the man than for the women) were found.

Discussion

The sound comparisons investigated and documented here demonstrate cases of vowel sounds which do not manifest age- and gender-related differences in estimated spectral peaks and formant frequencies < 1.5 kHz, the sounds documented being produced by all speakers at similar f_o levels. The documentation provides further evidence for our earlier interpretation that f_o variation has a much stronger effect on lower formants < 1.5–2 kHz than the assumed average vocal tract size of the speakers. (Here, with regard to a general appraisal and future research, a frequency range of < 1.5–2 kHz is given to include the consideration of sounds of /a, o, u/ produced at high f_o levels above 500 Hz.)

Obviously, the finding of a lack of general age- and gender-related *F*-pattern differences < 1.5–2 kHz independent of f_o of the compared sounds calls for a reconsideration of the relation between *F*-patterns and vocal tract sizes (above all, if occurring inversions of commonly expected age- and gender-related *F*-pattern differences are also taken into account; see the next chapter). However, this question is left open here, as is also the case for the question of whether or not speaker-related age or gender differences can be recognised when investigating sounds produced by speakers of all three speaker groups at similar f_o and manifesting similar vowel-related *F*-patterns.

Chapter Appendix

Table 1: Comparison of sounds and sound spectra of the long Standard German vowels produced by children, women and men at a similar f_o level: Sound series investigated and estimated *F*-patterns of experiment 1. Columns 1–3 = sounds (S/L = sound series and sound links; V = intended and recognised vowel quality; SG = speaker and speaker group, where c = child, w = woman, m = man). Columns 4–9 = f_o and estimated *F*-patterns (fo = calculated f_o, in Hz; F(i) = estimated formant frequencies; ΔF(i) = range or levels of formant frequencies for the sounds compared; for level indications, see text). Column 10 = parameter setting of LPC analysis for *F*-pattern calculation (Par; def = age- and gender-related standard default setting; P4, P5 and P6 = altered settings; for setting details, see Introduction). Colour code: Green = similar estimated formant frequencies for the sounds of the child, the woman and the man, or higher frequencies for the adults than the child. "*" = similar or higher F_2 for sounds of front vowels for adults compared to the child or for the man compared to the woman. Note that for the manual estimation of F_2 and F_3 for one sound produced by the child (see series for /y/, values given in parenthesis), the parameter setting is marked with "(M)".
[M-04-01-T01]

Table 2. Comparison of sounds and sound spectra of the long Standard German vowels produced by children, women and men at a similar f_o level: Sound series investigated and estimated *F*-patterns of experiment 2. Columns, colour code and "*" marks conform to those in Table 1. Note also that for the manual estimation of F_3 for one sound produced by a child (see series for /i/, value given in parenthesis), the parameter setting is marked with "(M)".
[M-04-01-T02]

Table 1. Comparison of sounds and sound spectra of the long Standard vowels German produced by children, women and men at a similar fo level: Sound series investigated and estimated F-patterns of experiment 1. [M-04-01-T01]

Sounds			fo and F-patterns (Hz)						Par
S/L	**V**	**SG**	**fo**	**F1**	**ΔF1**	**F2**	**ΔF2**	**F3**	
1 ↗	i	c	262	343	15	2926	–	–	def
		w	221	355		3079*		4018	P4
		m	221	340		2402		3550	P5
2 ↗	y	c	250	313	4	(2200)	–	(2850)	def (M)
		w	268	316		1796		2919	P4
		m	262	312		1826*		2333	def
3 ↗	e	c	243	493	20	2731	–	–	P5
		w	248	473		2693		3187	def
		m	250	492		2203		2504	def
4 ↗	ø	c	245	493	24	1917	–	3146	P5
		w	245	469		2033*		3068	def
		m	245	483		1757		2181	def
5 ↗	ε	c	260	868	44	2417	–	4235	def
		w	268	912		2706*		3666	def
		m	260	735	lower	2172		3167	P5
6 ↗	a	c	271	844	13	1204	lower	–	def
		w	246	1064	higher	1343	3	–	def
		m	257	831	13	1340		–	def
7 ↗	o	c	259	520	45	934	middle	–	P5
		w	256	476		775	low	–	def
		m	264	521		1003	high	–	def
8 ↗	u	c	297	312	68	711	65	–	def
		w	297	315		756		–	def
		m	332	380		776		–	def

Table 2. Comparison of sounds and sound spectra of the long Standard German vowels produced by children, women and men at a similar fo level: Sound series investigated and estimated F-patterns of experiment 2. [M-04-01-T02]

Sounds			fo and F-patterns (Hz)						Par
S/L	V	SG	fo c	F1	ΔF1	F2	ΔF2	F3	
1 ↗	i	c	249	297	31	3267	–	4370	def
		c	259	309		3038		(4150)	def (M)
		w	254	315		2296		3368	def
		w	263	292		2586		3447	def
		m	260	286		2576*		3614	P4
		m	263	284		2368		2685	def
2 ↗	y	c	254	290	61	2024	–	2479	def
		c	257	336		2162		2892	P5
		w	261	297		1814		2480	def
		w	268	340		1819		2480	def
		m	252	351		1738		2491	def
		m	262	316		1536		–	def
3 ↗	e	c	257	511	17	3046	–	3753	def
		c	267	510		2622		3379	def
		w	265	527		2562		3289	P4
		w	273	523		2347		3062	def
		m	264	524		1996		2453	def
		m	260	526		2310*		2888	P4
4 ↗	ø	c	260	456	60	2076	–	3552	def
		c	256	512		1536*		2641	P5
		w	258	496		1558*		2520	def
		w	267	452		1638*		2702	def
		m	255	508		1552*		2536	def
		m	258	502		1775*		2176	def

Table 2 (continuation). [M-04-01-T02]

Sounds			fo and F-patterns (in Hz)						Par
S/L	V	SG	fo c	F1	ΔF1	F2	ΔF2	F3	
5 ↗	ε	c	266	628	lower	2348	–	3142	def
		c	263	716	66	2321		3262	def
		w	241	768		2183		2923	def
		w	261	707		2089		3201	def
		m	258	773		1810		2715	P5
		m	264	658	lower	2104*		–	P5
6 ↗	a	c	262	792	39	1467	high	–	def
		c	258	824		1156	low	–	def
		w	257	739	lower	1283	middle	–	def
		w	258	974	higher	1209	middle	–	def
		m	257	831	39	1340	high	–	def
		m	257	815		1143	low	–	def
7 ↗	o	c	262	521	37	767	low	–	P5
		c	259	520		934	middle	–	P5
		w	267	506		1047	high	–	def
		w	261	531		941	middle	–	P6
		m	265	543		986	middle	–	def
		m	262	517		1036	high	–	def
8 ↗	u	c	265	341	39	778	58	–	def
		c	265	324		782		–	def
		w	253	339		742		–	def
		w	260	303		785		–	def
		m	265	380	higher	800		–	def
		m	261	302	39	836	higher	–	def

M4.2 Cases of Inverted Vowel-Related Spectral Differences for Sounds Produced by Children and Adults

Introduction

In the previous chapter, cases of vowel sounds are shown that do not manifest age- and gender-related differences in estimated spectral peaks and formant frequencies < 1.5 kHz, the compared sounds produced by all speakers at similar f_o levels. In pursuit of the question of whether vocal tract size stands in an imperative, direct relation to the entire vowel spectrum, a further experiment was devised to investigate the occurrence of inversions of commonly expected age-related *F*-pattern differences, that is higher vowel-related formant frequencies for sounds of adults than of children.

In the two earlier studies on *F*-patterns for sounds produced by men, women and children at similar and different f_o levels mentioned in the previous chapter (Maurer et al., 2015; Preliminaries, pp. 217–237), we have already reported and documented spectral characteristics < 1.5–2 kHz for sounds of children, women and men as sometimes being inverted with regard to vocal tract size-related expectations if the f_o levels of the compared sounds were highest for men, intermediate for women and lowest for children: Numerous cases with higher formant frequencies < 1.5–2 kHz for men than for women occurred, as was true for numerous sounds with higher formant frequencies < 1.5–2 kHz for adults than for children. Moreover, referring to the vocalises presented in Chapter M2.1 and comparing sounds of the adults produced at higher f_o levels with sounds of the child produced at lower f_o levels, and similarly also comparing the sounds produced by the man at higher f_o levels with the sounds produced by the woman at lower f_o levels, inversions of assumed age- and gender-related differences < 1.5–2 kHz as indicated in formant statistics also occurred, above all for sounds of close and close-mid vowels.

Furthermore, considering the large variation of observable higher formants for sounds of front vowels (for an illustration, see again Peterson and Barney, 1952), the question of the occurrence of inverted spectral age and gender differences can be extended to the comparison of F_1–F_2 patterns for sounds of all vowels, independent of the frequency range of F_2, and it can even be extended to the comparison of F_1–F_2–F_3 for sounds of front vowels. (Note in this context the corresponding indications given in the previous chapter.) Thereby, sound comparisons between children and adults are of particular interest because of the pronounced differences in their vocal tract sizes.

Against this background, to highlight the inversion phenomenon and to document and embed it in the context of our line of argument, an attempt was made to compile sound series manifesting inverted F_1 or F_1–F_2 or even F_1–F_2–F_3, that is, F_1 or F_1–F_2 or even F_1–F_2–F_3 standing in contrast to vocal tract size-related assumptions. The attempt was based on the above indications and on the assumption that the effect of f_o variation on the lower part of the vowel spectrum is more important than the vocal tract size of the speaker in question. Vocal effort and production style were disregarded for the sound selection.

Experiment

Sound selection and sound comparison: On the basis of the Zurich Corpus, for each of the eight long Standard German vowels, six sounds produced in V context at various f_o levels by two children, two women and two men were selected for which children–adult inversions for one or more estimated spectral peaks and formant frequencies were manifest. According to the standard listening test results when creating the corpus, all sounds were fully recognised (100% vowel recognition rate matching vowel intention). As mentioned, vocal effort and production style were disregarded. As a result, a sample of 48 sounds in total was created.

Formant frequency estimation: Formant frequency estimation accorded to the procedure described in the previous chapter.

Estimation of inverted formant frequencies: For each sound comparison, the occurring levels of F_1 or F_2 were assigned as low, middle and high, or low and high, to highlight the occurring inversions of commonly expected age-related *F*-pattern differences.

Formant frequency estimation for sounds at higher f_o levels is methodologically unsubstantiated. Therefore, the estimated values given here are only meant to indicate a spectral peak structure of the sounds in question, and only the direct comparison of the spectra can confirm the interpreted inversion of the commonly assumed age- and gender-related spectral differences.

Results

Table 1 in the chapter appendix shows the compilation of the sound series and estimated *F*-patterns. Inversions of commonly expected age- and gender-related spectral peaks and/or estimated formant frequencies in terms of lower F_1 or F_1–F_2 or F_1–F_2–F_3 for sounds of children when compared with sounds of adults are coloured green. According to the results of the acoustic analysis, the F_1 levels of the sounds of /i, y, o, u/ of all adults exceeded those of the children by more than 100 Hz. For the sounds of /e, ø, ɛ, a/, this also held true for one sound of a woman and one sound of a man at a minimum. The F_2 levels of the sounds of /y, ø, o, u/ of all adults exceeded those of the children by more than 100 Hz. For the sounds of /i, ɛ, a/, this also held true for one sound of a woman and one sound of a man at a minimum. Besides, note the occurring cases of children–women inversions of F_1–F_2–F_3 for sounds of /i/, /ø/ and /ɛ/ and of F_1–F_2 for sounds of /e/. In sum, pronounced children–adult inversions of F_1 were found and demonstrated here for all sounds of all vowels. Except for two sounds produced by men, the same held true for pronounced children–adult inversions of F_1–F_2.

Discussion

In the previous chapter, we have presented cases of vowel sound comparisons which do not manifest age- and gender-related differences in estimated spectral peaks and formant frequencies < 1.5 kHz, the sounds documented being produced by all speakers at similar f_o levels. We have concluded that cases of this kind support our earlier interpretation that f_o variation affects lower formants < 1.5–2 kHz much more than the assumed average vocal tract size of the speakers does. The present demonstration of cases of vowel sound comparisons, for which commonly assumed age-related children–adult differences in estimated spectral peaks and formant frequencies did not only decrease or disappear but were inverted, strengthens this conclusion. Notably, even for sounds of front vowels, this inversion was not limited to F_1 but could be shown to occur also for F_1–F_2 and, in rare cases, even for F_1–F_2–F_3.

Vocal effort, production style and register changes were disregarded in the experiment. However, no systematic speaker group-related vocal effort was manifest in the sound compilation, and only seven of the 32 selected sounds of the adults were produced in a particular style (six sounds in CS style, one sound in ST style). Changes in register and associated changes in articulation must be taken into account when interpreting the inversions, but they can hardly explain the phenomenon as such. (On this matter, see also Chapter 4.3 in the main part of this treatise.)

Inversions of commonly assumed gender-related women–men differences in formant frequencies were not investigated. However, much more pronounced inversions can be expected for women–men comparisons than for the documented children–adult comparisons (see also Chapter M2.1 and the Preliminaries, pp. 217–237).

Chapter Appendix

Table 1: Comparison of sounds and sound spectra of the long Standard German vowels produced by children and adults at different *f*ₒ levels and/or with different vocal efforts: Sound series investigated and estimated *F*-patterns. Columns 1–4 = sounds (S/L = sound series and sound links; V = intended and recognised vowel quality; SG = speaker and speaker group, where c = child, w = woman, m = man; VE = vocal effort). Columns 5–11 = *f*ₒ and estimated *F*-patterns (fo = calculated *f*ₒ, in Hz; F(i) = estimated formant frequencies, in Hz; F(i)-C = frequency levels of the formants within a sound series in terms of low, middle or high levels of the sounds compared). Column 12 = parameter setting of LPC analysis for *F*-pattern calculation (Par; def = default, age- and gender-related standard setting; P4, P5 and P6 = altered settings; for setting details, see Introduction). Colour code: Green = inverted formant frequency differences (higher levels for adults than children).
[M-04-02-T01]

Table 1. Comparison of sounds and sound spectra of the long Standard German vowels produced by children and adults at different fo levels and/or with different vocal efforts: Sound series investigated and estimated F-patterns. [M-04-02-T01]

Sounds				fo and F-patterns (Hz)							Par
S/L	V	SG	VE	fo	F1	F1-C	F2	F2-C	F3	F3-C	
1 ↗	i	c	med	222	251	low	2645	middle	3676	low	def
		c	high	380	388	low	2637	middle	4105	high	def
		w	high	486	518	middle	2917	high	3886	middle	P4
		w	high	581	558	middle	2937	high	3703	low	def
		m	med	584	603	high	2894	high	–	–	P4
		m	low	586	590	high	2347	low	–	–	P4
2 ↗	y	c	med	294	308	low	1793	low	–	–	def
		c	med	300	317	low	1646	low	3072	high	def
		w	med	504	502	high	2024	high	2688	middle	P4
		w	low	524	524	high	2097	high	–	–	def
		m	high	514	518	high	2053	high	2568	low	P5
		m	high	511	511	high	2030	high	2669	middle	def
3 ↗	e	c	low	397	431	low	2486	middle	–	–	def
		c	med	322	477	low	2361	low	–	–	def
		w	high	323	627	high	2587	high	3050	high	P4
		w	high	330	673	high	2596	high	3009	high	def
		m	high	264	526	middle	2435	middle	2931	high	P5
		m	med	297	583	midle	2323	low	2795	low	def
4 ↗	ø	c	med	389	434	low	1546	low	–	–	def
		c	med	209	421	low	1483	low	2679	low	P5
		w	med	296	587	high	1822	high	2946	high	def
		w	med	318	612	high	1893	high	2701	low	def
		m	med	360	656	high	1809	high	–	–	def
		m	high	266	516	middle	1815	high	2631	low	def
5 ↗	ε	c	high	197	600	low	2236	low	3222	middle	def
		c	med	228	637	low	2084	low	3024	low	def
		w	high	310	921	high	2420	high	3155	middle	def
		w	low	480	967	high	2483	high	3737	high	P4
		m	high	259	707	middle	2301	middle	–	–	def
		m	med	291	780	middle	2358	middle	–	–	def
6 ↗	a	c	high	198	739	low	1265	low	–	–	def
		c	med	247	786	low	1283	low	–	–	def
		w	high	165	880	high	1313	middle	–	–	def
		w	high	237	950	high	1410	high	–	–	def
		m	high	288	920	high	1419	high	–	–	P5
		m	high	319	966	high	1492	high	–	–	def
7 ↗	o	c	low	223	442	low	881	low	–	–	def
		c	high	223	442	low	800	low	–	–	def
		w	med	271	554	high	985	middle	–	–	def
		w	med	287	572	high	1143	high	–	–	def
		m	high	283	559	high	1192	high	–	–	def
		m	high	291	583	high	1108	high	–	–	def
8 ↗	u	c	low	250	281	low	743	low	–	–	def
		c	med	297	312	low	711	low	–	–	def
		w	med	518	519	high	1052	high	–	–	def
		w	med	438	470	middle	905	middle	–	–	def
		m	med	501	503	high	1002	middle	–	–	P5
		m	high	443	467	middle	1143	high	–	–	def

M5 Vowel Spectrum, Phonation Type and Vocal Effort

M5.1 Vowel Spectrum and Phonation Type I – Comparing the Spectra of Natural Vowel Sounds Produced With Voiced, Whispered, Creaky and Breathy Phonation, Excluding f_o Variation of the Voiced Sounds

Introduction

Four main observations and experimental results regarding the differences between whispered and voiced sounds have been reported in the literature (for an overview, see Swerdlin et al., 2010; Sharifzadeh et al., 2012; Houle and Levi, 2020): Whispered sounds tend to be longer compared to voiced sounds (for a detailed discussion, see Houle and Levi, 2020); their spectra exhibit higher estimated spectral peak frequencies ($P_{(i)}$ and P-patterns, see Introduction) and estimated F-patterns (with most pronounced differences for P_1 and F_1, and with widened formant bandwidths) as well as a different spectral energy distribution (lower acoustic power, relatively flat noise-like spectrum, see McLoughlin et al., 2013; note that these spectral differences were reported as also being related to vowel qualities, see Sharifzadeh et al., 2012); their vowel recognition has proven to be somewhat impaired (Kallail and Emanuel, 1984a, 1984b, 1985; Tartter, 1989; Eklund and Traunmüller, 1997); they have been found to have a pitch although their spectra do not manifest a harmonic structure and the sounds lack f_o (for details and references, see the introduction to Chapter M6.1).

The reported spectral differences related to whispered and voiced phonation are generally understood as being mainly a consequence of changes in the geometry of the vocal tract around the glottis and the difference in the source characteristics. (The reported differences in sound duration and vowel recognition are not investigated further here.)

The periodicity of vowel sounds produced with creaky phonation is reported as irregular, but the sound spectrum is reported to show no or only marginal spectral peak differences when compared with voiced sounds. (For an overview, see Swerdlin et al., 2010; Keating et al., 2015; note that Swerdlin et al. reported small but significant frequency shifts from voiced to creaky phonation for the first and third resonances and associated these shifts with a decrease in the ratio of the glottal length to the glottal area. However, as a general estimate, they noted: "The average values of the resonance frequencies $R1$–$R4$ for creak phonation […] are similar to the average values for the normal [voiced] voice.")

The spectral characteristics of vowel sounds produced with breathy phonation (for an overview, see Hillenbrand and Houde, 1996) are reported to manifest an increased amplitude of the first harmonic, a lessened sound periodicity and weak levels of the harmonics in the upper frequency part of the spectrum combined with increased aspiration noise when compared with voiced vowel sounds. These phonation-related acoustic differences are understood as a consequence of the tendency of the glottal source function to be sinusoidal-like combined with a high level of aspiration noise due to non-simultaneous closure along the length of the vocal folds.

However, as is true for numerous other studies on vowel acoustics, the appraisal of the above studies and their results has to take into consideration that

- the reported observations were generally related to comparisons with voiced sounds produced with a medium vocal effort at f_o levels as given in formant statistics, that is, for voiced vowel sounds in the lower vocal range of the speakers investigated; above all, no further differentiation is commonly given in the literature for comparisons including f_o variation for voiced vowel sounds and also vocal effort variation for both voiced and whispered vowel sounds (however, for a study comparing whispered sounds produced with vocal effort variation and voiced sounds produced with the three voice qualities breathy, normal and pressed, see Konnai et al., 2017);
- with few exceptions, the reported observations also did not address the question of phonation types and their subtypes (see e.g. Keating et al., 2015, for subtypes of creaky phonation, and Konnai et al., 2017, for subtypes of whispered phonation);
- the general problem of spectral peak and formant estimation, further aggravated for breathy and whispered sounds, relativises direct numerical comparisons of *F*-patterns for sounds produced with different phonation types;
- the studies published on the matter differed in method, vowel sets of investigation and details of reported results.

To re-examine the reported spectral variation due to to phonation types but taking into account the relation of the vowel spectrum to f_o for natural voiced sounds, three experiments were conducted: In the first experiment, sounds of all eight long Standard German vowels produced by three speakers (a man, a woman and a child) with voiced, whispered, creaky and breathy phonation were compiled, and their spectra were examined concerning the main characteristics reported in the literature, f_o of the voiced sounds corresponding to speaker-group average f_o levels as given in formant statistics. The aim of this first study was limited to the documentation of a set of vowel sounds comparable to sets investigated in the literature to provide direct and accessible sound comparisons and a starting point and literature-related reference for the subsequent experiments. In the second experiment, based on sounds of the same vowels and the same three speakers as in the first experiment, whispered sounds were compared with voiced sounds produced at various levels of f_o to investigate the question of whether reported and observed phonation-related spectral differences in the lower frequency range (the range generally attributed to the frequency range of F_1) decrease or disappear or are even inverted if the f_o of the voiced sounds of comparison is markedly increased from a lower to a higher level. In the third experiment, based on an inspection of whispered, creaky and voiced sounds of the eight Standard German vowels produced by two speakers (a man and a woman), phonation-related *F*-patterns and source characteristics were derived, and sounds were synthesised for different combinations of phonation-related *F*-patterns and source characteristics, the voiced source characteristics including different levels of f_o.

The results of the first experiment are reported in this chapter. The experiment was conducted in the context of an Interspeech Show and Tell conference presentation (Maurer et al., 2019, Chapter 6.1), and it was transferred to the present treatise, including an extended description of the method and a new discussion. The second and third experiments are discussed in the following two chapters.

Experiment

Selection of speakers and sounds: On the basis of the Zurich Corpus, for each of the eight long Standard German vowels, sound samples of a man, a woman and a child were selected that included sounds for all four types of voiced, whispered, creaky and breathy phonation. All sounds were fully recognised in the standard listening test conducted when creating the corpus (100% recognition rate matching vowel intention). Subsequently, for each speaker, each vowel and each of the four phonation types, one sound produced in V context and in nonstyle mode and, for the voiced sounds, with a medium vocal effort was selected. Note for voiced and breathy vowel sounds:

- In order to relate the sound comparison of the present study to the experimental settings of most of the studies discussed in the literature, only voiced sounds produced at intended f_o of 131 Hz (man), 220 Hz (woman) and 262 Hz (child) were selected.
- According to the standard procedure of the Zurich Corpus, the sounds with breathy phonation were produced by the speakers spontaneously, with speaker-related levels of intended f_o.

As a result of this selection procedure, eight vowel-related comparisons of four sounds were compiled for each speaker, and a sample of 96 sounds (32 sounds per speaker) was investigated.

General spectral analysis, and spectral comparison of voiced and whispered as well as voiced and creaky vowel sounds: Acoustic analysis of all sounds accorded with the standard procedure of the Zurich Corpus. Initially, when comparing voiced and whispered as well as voiced and creaky sounds of a single vowel produced by a single speaker, we tried to assess the main spectral differences or similarities for F_1–F_2 patterns, as reported in the literature, in numerical terms on the bases of calculated F_1–F_2 patterns and a visual crosscheck inspecting the sound spectra and spectral peaks, spectrograms and F_1–F_2 tracks. However, for the majority of the whispered sounds, we were not able to calculate *F*-patterns according to this procedure, and the same held for a substantial part of the creaky sounds. Therefore, we limited ourselves to a visual inspection of the sound spectra and a rough estimation of indications of differences in the spectral peaks. Further, we restricted the examination to P_1 and P_2 so as to limit the *P*-pattern estimation problems. (Note that spectral differences for the comparison of voiced and whispered sounds were reported to concern P_1/F_1 and P_2/F_2 primarily.)

For each comparison of the voiced and whispered sound of a vowel and a speaker, the estimates were made as follows (see Table 1, Columns 5 and 6):

- Higher P_1 and/or P_2 for the whispered sound exceeding P_1 and/or P_2 of the voiced sound by 100 Hz at a minimum, according to the thesis of higher P_1/F_1 and eventually also P_2/F_2 as given in the literature ("y" for "yes"; weak indications with a peak frequency difference < 100 Hz are given in parentheses);
- Similar P_1 and/or P_2 for both sounds ("s" for "similar")
- Estimation of P_1 and/or P_2 highly questionable ("?" for "in question")
- Missing *P*1 and/or *P*2 ("m" for "missing")

For each comparison of the voiced and creaky sound of a vowel and a speaker, the estimates were made as follows (see Table 1, Columns 7 and 8):

- Comparable P_1 and/or P_2 for both sounds (estimated peak frequency difference ≤ 50 Hz), according to the thesis of comparable P_1/F_1 and P_2/F_2 as given in the literature ("y" for "yes")
- Higher or lower P_1 and/or P_2 for the creaky sound ("h" for "higher", "l" for "lower")
- Estimation of P_1 and/or P_2 highly questionable ("?" for "in question")

For the comparison of voiced and breathy sounds (see Table 1, Columns 9 and 10), the hypothesis of either a dominant first harmonic or a markedly steeper spectral slope < 1 kHz ("y" for "yes", "n" for "no") and an increased aspiration noise for breathy sounds ("y" for "yes", "n" for "no") was investigated.

Results

Table 1 in the chapter appendix shows the vowel sounds investigated and the detailed results of their spectral investigation and comparison, including sound links. (Note that these sound links include the voiced sounds of comparison.) Table 2 shows a summary of these results.

Comparison of voiced and whispered sounds: According to the results of the acoustic analysis, 16 of the 24 whispered sounds showed increased P_1 compared to voiced sounds of the same speaker and vowel in question, 12 sounds showed increased P_1 and increased P_2, and three sounds showed increased P_2 only. Otherwise, spectral peaks were either similar for the whispered and the voiced sounds or they were difficult to estimate.

Note in addition:

- The whispered sounds of front vowels in Series 3, 11 and 13 showed two spectral peaks < c. 1.1 kHz. However, in Series 3, P_1 of the voiced sound was found to be clearly below P_1 of the whispered sound, and in Series 13, the dominant spectral peak at c. 1 kHz was interpreted.
- The whispered sounds of front vowels in Series 12 and 20 also showed two spectral peaks < c. 1.1 kHz. BP filtering of the second lower peak < 1.1 kHz did not result in a change in vowel quality, but BP filtering of the first lower peak resulted in a quality shift (author's estimate). The comparison of the voiced and the whispered sounds of these two series was based on this filtering experience.

- The whispered sound in Series 15 showed three spectral peaks < 1.3 kHz. In a direct spectral comparison, no clear indication was found for a P_1–P_2 whispered–voiced difference.
- The whispered sound in Series 7 showed no clear spectral peak structure but a frequency band of higher spectral energy < 1 kHz. However, direct spectral comparison supported the notion of lower P_1–P_2 for the voiced sound compared with the spectral energy < 1 kHz for the whispered sound.
- Deviations from the expected evidence for a clear spectral difference between voiced and whispered vowel sounds were indicated as speaker-related (see Table 2), with fewer deviations for the sounds of the man compared with the sounds of the woman and the child.

Comparison of voiced and creaky sounds: According to the results of the acoustic analysis, for 12 of the 24 creaky sounds investigated, the related spectra manifested approximately similar estimated P_1–P_2 when compared with the voiced sounds of the speaker and the vowel in question. For the remaining 12 creaky sounds, the spectral differences when compared to voiced sounds were either comparable P_1 but higher or lower P_2 (11 sounds), or comparable P_2 but higher P_1 (one sound).

Note also a second additional spectral peak < 1 kHz (with a lower peak level than the first peak) for the sounds of front vowels in Series 17 and 18.

Comparison of voiced and breathy sounds: According to the results of the acoustic analysis, all spectra of the breathy sounds of the two adult speakers showed a dominant first harmonic and/or a markedly steeper spectral slope < 1 kHz combined with increased aspiration noise when compared to voiced sounds. For the child, however, this was the case for only one sound. Concerning the other seven sounds, the spectral investigation indicated:

- increased $H1$ level only (no pronounced increase of aspiration noise; see Series 24; see also below the remark for Series 23);
- increased aspiration noise only (no pronounced level difference for $H1$ or a steeper spectral slope < 1 kHz; see Series 17, 18, 19, 20);
- similar spectral envelopes for breathy and voiced sounds (neither a pronounced level difference for $H1$ or a steeper spectral slope < 1 kHz nor a pronounced increase of aspiration noise; see Series 21)
- difficulty in spectral valuation and comparison of breathy and voiced sounds (see Series 23, the difficulty of the spectral comparison being in part due to substantially higher f_o of the actual breathy sound compared with the f_o level intended; however, also note that

producing breathy sounds with pronounced aspiration noise proved to be difficult for the child).

With regard to the interpretation of these results, the speaker-related differences in f_o of voiced and breathy sounds have to be taken into consideration: Intended f_o = 131 Hz (voiced) and 165 Hz (breathy) for sounds of the man; intended f_o = 220 Hz (voiced) and 349 Hz (breathy) for sounds of the woman; intended f_o = 262 Hz (voiced) and 262 Hz (breathy) for sounds of the child (with high calculated f_o of 330 Hz for the breathy sound of /o/).

Note in addition: For some breathy sounds, the methodological substantiation of *P*-pattern (and *F*-pattern) estimation was weak, and for other sounds, above all for sounds of front vowels, the question arose as to whether a mixture of lower harmonics and additional noise in the higher vowel-related frequency range may play a role in the acoustic representation of recognised vowel quality. (For examples, see Figure 1 in the chapter appendix).

Discussion

All three hypotheses for spectral differences related to differences in phonation types were confirmed for the majority of the sounds examined. However, (i) a substantial number of whispered, creaky and breathy sounds occurred that manifested spectral characteristics not according to the hypotheses exposed in the introduction to this chapter (possibly related to either the age and gender of the speakers or speaker-specific sound production), (ii) the spectra of creaky vowel sounds often manifested a pronounced spectral peak < c. 100 Hz (see, above all, the sounds of the man), and (iii) the general methodological problem of spectral evaluation considerably relativised the spectral estimates.

As mentioned in the introduction to this chapter, the aim of this experiment was limited to the documentation of a set of vowel sounds comparable to sets investigated in the literature, to provide direct and accessible sound comparisons and to create a starting point and literature-related reference for subsequent experiments. Therefore, no advanced discussion and relativisation of the results – relativisations that concern, above all, the small number of sounds and speakers, the lack of an inclusion of phonation subtypes in general and vocal effort variation for voiced and whispered sounds in particular, the lack of f_o variation for voiced and breathy sounds as well as the general methodological problem of spectral comparison – is made here. In these terms, the present study only provided illustrating examples based on an investigation comparable to most of the research published in the literature. With regard to a general appraisal of phonation-related spectral differences of vowel sounds, above all concerning whispered–voiced comparisons, the conclusion of Konnai et al. (2017) is worth noting: Existing assumptions regarding general phonation-related acoustic differences should be viewed with caution, acoustic characteristics being dependent on different phonation subtypes and/or levels of vocal effort. From the perspective of the present treatise and its line of argument, the same holds true for acoustic characteristics being dependent on f_o (and pitch).

Chapter Appendix

Table 1. Comparison of natural vowel sounds produced with voiced, whispered, creaky and breathy phonation, excluding f_o variation of the voiced sounds: Details. Columns 1–4 = sounds (S/L = sound series and sound links, with speaker-related series numbers; V = intended and recognised vowel quality; P = phonation type, where wh = whispered, cr = creaky, br = breathy; fo = intended f_o, in Hz). Columns 5–9 = spectral characteristics examined in comparison to the spectra of the voiced reference sounds (higher P1-P2 = increased P_1 and/or P_2 for whispered sounds compared with voiced sounds; comparable P1-P2 = similar P_1 and/or P_2 for creaky sounds compared with voiced sounds; H1 and AN = either a dominant first harmonic or a markedly steeper spectral slope < 1 kHz (H1) and increased aspiration noise (AN) for breathy sounds in contrast to voiced sounds). Note: “y” for “yes”, “n” for “no”, “l” for “lower”, “h” for “higher”, “s” for “similar”, “?” for “in question”. Colour code: Green = spectral differences or similarities that accorded to the general hypotheses as given in the literature (see text); red = other spectral chararcteristics.
[M-05-01-T01]

Table 2. Comparison of natural vowel sounds produced with voiced, whispered, creaky and breathy phonation, excluding f_o variation of the voiced sounds: Summary. Column 1 = phonation types of comparison. Column 2 = spectral features evaluated. Columns 3–5 = speaker-related results. Column 6 = results for all three speakers. Colour code: Green = spectral differences or similarities that accorded to the general hypotheses as given in the literature.
[M-05-01-T02]

Figure 1. Occurring specific spectral characteristics of breathy sounds. Sounds 1–3 = examples of breathy vowel sounds for which methodological substantiation of P_1 and/or P_2 (and F_1 and/or F_2) estimation is weak. Sounds 4–8 = examples of breathy vowel sounds for which a distinct harmonic structure is weak or absent in the upper vowel-related frequency range.
[M-05-01-F01] ↗

Table 1. Comparison of natural vowel sounds produced with voiced, whispered, creaky and breathy phonation, excluding fo variation of the voiced sounds: Details. [M05-01-T01]

Sounds of a man fo of reference voiced sounds = 131 Hz									
Sounds				Spectral comparison					
S/L	V	P	fo Hz	higher P1-P2		comp. P1-P2		H1	AN
1 ↗	i	wh	–	m	y				
		cr	–			y	l		
		br	165					y	y
2 ↗	y	wh	–	(y)	y				
		cr	–			y	y		
		br	165					y	y
3 ↗	e	wh	–	y	y				
		cr	–			y	h		
		br	165					y	y
4 ↗	ø	wh	–	y	y				
		cr	–			y	y		
		br	165					y	y
5 ↗	ɛ	wh	–	y	y				
		cr	–			y	y		
		br	165					y	y
6 ↗	a	wh	–	y	y				
		cr	–			h	y		
		br	165					y	y
7 ↗	o	wh	–	y	y				
		cr	–			y	(h)		
		br	165					y	y
8 ↗	u	wh	–	y	y				
		cr	–			y	y		
		br	165					y	y

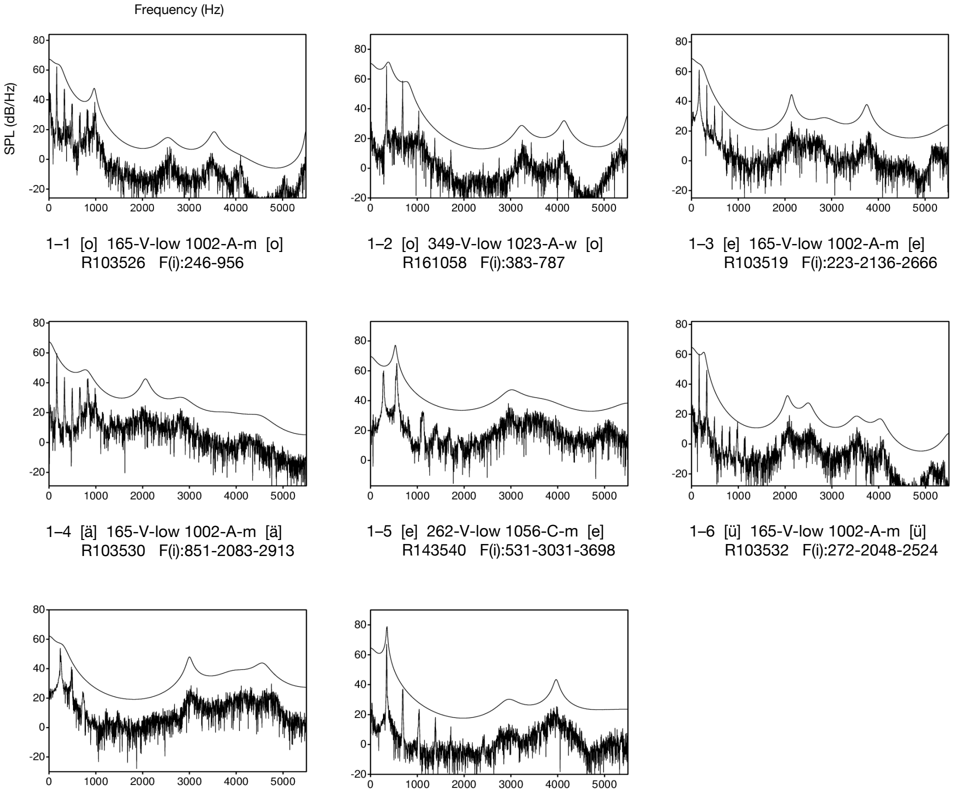
Sounds of a woman fo of reference voiced sounds = 220 Hz									
Sounds				Spectral comparison					
S/L	V	P	fo Hz	higher P1-P2		comp. P1-P2		H1	AN
9 ↗	i	wh	–	y	?				
		cr	–			y	y		
		br	349					y	y
10 ↗	y	wh	–	y	y				
		cr	–			y	y		
		br	349					y	y
11 ↗	e	wh	–	s	s				
		cr	–			y	h		
		br	349					y	y
12 ↗	ø	wh	–	s	s				
		cr	–			y	y		
		br	349					y	y
13 ↗	ɛ	wh	–	y	?				
		cr	–			y	l		
		br	349					y	y
14 ↗	a	wh	–	y	?				
		cr	–			y	y		
		br	349					y	y
15 ↗	o	wh	–	?	?				
		cr	–			y	y		
		br	349					y	y
16 ↗	u	wh	–	y	s				
		cr	–			y	l		
		br	349					y	y

Sounds of a child fo of reference voiced sounds = 262 Hz									
Sounds				Spectral comparison					
S/L	V	P	fo Hz	higher P1-P2		comp. P1-P2		H1	AN
17 ↗	i	wh	–	y	y				
		cr	–			y	h		
		br	262					n	y
18 ↗	y	wh	–	s	y				
		cr	–			y	h		
		br	262					n	y
19 ↗	e	wh	–	s	y				
		cr	–			y	?		
		br	262					n	y
20 ↗	ø	wh	–	s	l				
		cr	–			y	h		
		br	262					n	y
21 ↗	ɛ	wh	–	y	y				
		cr	–			y	h		
		br	262					n	n
22 ↗	a	wh	–	y	y				
		cr	–			y	(y)		
		br	262					y	y
23 ↗	o	wh	–	s	s				
		cr	–			y	y		
		br	262					n	y
24 ↗	u	wh	–	y	y				
		cr	–			y	y		
		br	262					y	n

Table 2. Comparison of natural vowel sounds produced with voiced, whispered, creaky and breathy phonation, excluding fo variation of the voiced sounds: Summary. [M-05-01-T02]

Comparison	Spectral features	Man	Woman	Child	Total
whispered versus voiced	F1–F2 increased	7	1	4	12
	only F1 increased	–	4	–	4
	only F2 increased	1	–	2	3
	F1–F2 similar	–	2	1	3
	miscellaneous	–	1	1	2
	(all sounds)	–	–	–	24
creaky versus voiced	F1–F2 similar	4	5	3	12
	only F1 similar	3	3	5	11
	only F2 similar	1	–	–	1
	F1–F2 different	–	–	–	–
	miscellaneous	–	–	–	–
	(all sounds)	–	–	–	24
breathy versus voiced	increased H1 and aspiration noise	8	8	1	17
	increased H1 only	–	–	1	1
	increased aspiration noise only	–	–	4	4
	similar spectral envelopes	–	–	1	1
	miscellaneous	–	–	1	1
	(all sounds)	–	–	–	24

Figure 1. Occurring specific spectral characteristics of breathy sounds. [M-05-01-F01]



1–1 [o] 165-V-low 1002-A-m [o]
R103526 F(i):246-956

1–2 [o] 349-V-low 1023-A-w [o]
R161058 F(i):383-787

1–3 [e] 165-V-low 1002-A-m [e]
R103519 F(i):223-2136-2666

1–4 [ä] 165-V-low 1002-A-m [ä]
R103530 F(i):851-2083-2913

1–5 [e] 262-V-low 1056-C-m [e]
R143540 F(i):531-3031-3698

1–6 [ü] 165-V-low 1002-A-m [ü]
R103532 F(i):272-2048-2524

1–7 [i] 262-V-low 1056-C-m [i]
R143541 F(i):348-2999-3975

1–8 [i] 349-V-low 1023-A-w [i]
R161054 F(i):344-3046-3960

M5.2 Vowel Spectrum and Phonation Type II – Comparing the First Lower Spectral Peak Frequency of Natural Vowel Sounds Produced With Voiced and Whispered Phonation, Including f_o Variation of the Voiced Sounds

Introduction

As mentioned in the introduction to the previous chapter, according to the literature, the spectra of whispered vowel sounds are assumed to exhibit somewhat higher *F*-patterns in general and a pronounced increase for F_1 in particular when compared with voiced vowel sounds. This phonation-related difference is commonly explained by different vocal tract resonance patterns resulting from a physiological difference at the level of the glottis for the two phonation types. However, in the published studies on the comparison of voiced and whispered sounds, the f_o level of the investigated voiced sounds was generally at lower frequency levels of the vocal range of the speakers (corresponding to levels of citation-form words or relaxed speech), with no substantial intra-speaker variation of f_o. (As also stated, with rare exceptions, the same held true for the lack of variation of phonation subtypes and vocal effort.)

Considering the observed relation of vowel-specific spectral characteristics to f_o for natural sounds, the general question of a possible decrease or disappearance or even inversion of the assumed acoustic voiced–whispered difference arises. This question was investigated in an extension of the previous experiment by including f_o variation for voiced sounds produced with a medium vocal effort and by adding voiced sounds produced with a high vocal effort at middle or higher f_o levels. The investigation was limited to the examination of P_1 since, according to the literature, phonation-related differences between voiced and whispered sounds were reported to be most pronounced for P_1/F_1. In addition, in this approach, the methodological problem of spectral peak and formant frequency estimation was limited to evaluating the lower frequency range < 1 kHz.

Experiment

Selection of speakers and sounds: For each of the eight long Standard German vowels and each of the three speakers of the previous experiment, the voiced and whispered sounds were selected. Subsequently, for each sound comparison of a vowel and on the basis of other voiced sounds produced by the same speakers in V context with medium vocal effort (and in nonstyle mode for the two adult speakers) documented in the Zurich Corpus, two additional voiced sounds at medium and higher levels of f_o were selected for which their spectra showed either comparable or higher estimated P_1/F_1 than for the whispered sound. Finally, applying the same procedure, one sound per vowel and speaker produced in V context with a high vocal effort at a middle or higher f_o level was also added to each of the sound comparisons. Sounds produced with a high vocal effort were included in order to obtain a first indication of a possible additional effect of vocal effort variation for this type of sound comparison. All sounds were fully recognised in the standard listening test conducted when creating the corpus (100% recognition rate matching vowel intention). As a result of this selection procedure, eight vowel-related comparisons of five sounds were compiled for each speaker, and a sample of 120 sounds (40 sounds per speaker) was investigated.

Comparison of estimated P_1: For each comparison of a voiced and the whispered sound of a vowel and a speaker, differences or similarities of P_1 were estimated relating to the procedure of the previous experiment described in Chapter M5.1, as follows: Comparable P_1, or lower or higher P_1 for the voiced than the whispered sound, if the differences in P_1 exceeded 100 Hz at a minimum (weak indications with a peak frequency difference < 100 Hz are given in parentheses).

Results

Table 1 in the chapter appendix shows the results for the comparison of P_1 of voiced and whispered vowel sounds produced by the three speakers, including sound links. Since, in the table, the whispered sounds are taken as the reference for the comparison with the voiced sounds, the results are discussed in whispered–voiced order.

Whispered sounds versus voiced sounds produced at lower f_o levels: For 16 of the 24 comparisons, whispered sounds were observed for which the spectra manifested P_1 at a higher estimated frequency level than P_1 of the voiced sounds produced with a medium vocal effort at lower f_o levels (131 Hz for the man, 220 Hz for the woman, 262 Hz for the child; see Table 1, the first and second rows of a comparison, voiced sounds marked in blue). This finding accorded to the results of the previous experiment. For the remaining six comparisons, either P_1 of the whispered sound was lacking (see Series 1) or P_1 of the whispered sounds were comparable to P_1/F_1 of the voiced sounds.

Whispered sounds versus voiced sounds produced at middle or higher *f*ₒ levels: For all comparisons except one (see Series 1), spectra of voiced sounds produced with medium vocal effort at middle or higher f_o levels could be demonstrated that manifested P_1 comparable to the spectrum of whispered sounds (see Table 1, second to fourth row of a comparison, values marked in green; as mentioned in the previous chapter, the sounds of Series 1 constitute an exception, since no peak comparison was possible because of a lack of a lower spectral peak for the whispered sound). Further, for 14 of the 24 comparisons, voiced sounds produced with a medium vocal effort at middle or higher f_o levels occurred for which the spectra manifested P_1 that surpassed P_1 of the whispered sounds (see Table 1, second to fourth row of a comparison, values marked in red). Finally, for five of the remaining ten comparisons, voiced sounds produced with a high vocal effort at middle or higher f_o levels occurred for which the spectra manifested P_1 that also surpassed P_1 of the whispered sounds. Thus, vocal effort variation from medium to high played a limited additional role in sound comparisons of this type.

Note in addition: Concerning two occurring spectral peaks < 1 kHz for whispered sounds of front vowels and three peaks or a frequency band with prominent spectral energy < 1.3 kHz for whispered sounds of back vowels, see the corresponding note in the previous chapter.

Discussion

Repeating the findings of the previous experiment, if the spectra of whispered and voiced vowel sounds produced by the three speakers were compared, and if the voiced sounds were produced at the lower f_o levels comparable to age- and gender-specific average levels as given in formant statistics, the P_1 of the whispered sounds were in most cases higher than the P_1 of the voiced sounds. However, if the voiced sounds were produced by speakers of all three speaker groups with a medium vocal effort at increased levels of f_o, for all comparisons except Series 1, P_1 of some of the voiced sounds were comparable to P_1 of the whispered sounds. Moreover, for increased fo levels of the voiced sounds, the related P_1 surpassed P_1 of the whispered sounds for 14 of 24 comparisons. Finally, if the vocal effort was also varied in terms of investigating sounds produced with both medium and high vocal effort at middle or higher f_o levels, voiced sounds with P_1 that surpassed P_1 of whispered sounds occurred in 19 of the 24 comparisons.

Thus, to say that the spectra of whispered vowel sounds manifest somewhat higher *P*- and *F*-patterns and a pronounced increase for P_1 and F_1

in particular when compared with voiced vowel sounds may be appropriate for a specific set of voiced sounds of comparison produced at lower levels of f_o (if relativisations, as mentioned in the previous chapter, are taken into account), but to say that whispered vowel sounds *in general* manifest this kind of spectral difference when compared with voiced vowel sounds is empirically contradicted: Even if the study presented here was very limited with regard to the number of speakers and sounds investigated, the results nevertheless indicated that the spectra of many voiced vowel sounds can be expected to manifest P_1/F_1 comparable to or even higher than P_1/F_1 of whispered sounds. This may even apply to the comparison of the entire spectral envelope of voiced and whispered sounds.

The indication that P_1/F_1 (and possibly the entire spectral envelope) of whispered vowel sounds seem to correspond to the P_1/F_1 of voiced sounds produced at intermediate f_o levels of a speaker's vocal range will be addressed and discussed in more detail in the following chapter.

A special note on the whispered sound of the front vowel /e/ in Series 3 is added here. As mentioned, the spectrum of this sound exhibited two peaks < 1 kHz, the first at c. 500 Hz and the second at c. 750 Hz. Deleting the second peak in terms of deleting the spectral energy in the frequency range of 600–1000 Hz (using a corresponding BP filter) did not affect the recognised vowel quality, while deleting the low spectral frequency range < 600 Hz with the first peak (using an HP filter) resulted in a vowel shift towards /ɛ/ (author's estimate). We have observed several other whispered sounds of front vowels with two lower peaks in their spectrum. This manifestation and the lacking effect of either the first or the second lower peak on vowel recognition should be accounted for when considering the relation between vowel quality recognition and its spectral representation. For its part, this observation questions the assumption that vowel quality recognition directly relates to spectral peaks.

Several other aspects of natural whispered and voiced vowel sounds are not discussed here, e.g. different whisper subtypes, speaker-related characteristics and individual vocal abilities, the entire range of vocal effort variation for voiced sounds, and very high-pitched vowel sounds. However, some of these aspects will be addressed in the sixth main chapter.

Chapter Appendix

Table 1. Comparison of natural whispered and voiced vowel sounds and their lower spectral characteristics, including f_o variation of the voiced sounds. Columns 1–5 = sounds (S/L = sound series and sound links, with speaker-related series numbers; V = intended and recognised vowel quality; P = phonation type, where wh = whispered, v = voiced; VE = vocal effort, where med = medium, high = high; fo = intended f_o, in Hz). Column 6 = spectral comparison of P_1 of the whispered sound (reference) and the voiced sounds; indications are given for the voiced sounds in relation to the whispered sound; indications given in parentheses, see text). Colour code: Blue = lower P_1 for the voiced than the whispered sound of comparison; green = comparable P_1 for the voiced and the whispered sound of comparison; red = higher P_1 for the voiced than the whispered sound of comparison.
[M-05-02-T01]

Table 1. Comparison of natural whispered and voiced vowel sounds and their lower spectral characteristics, including fo variation of the voiced sounds. [M-05-02-T01]

Man					
Sounds					Spectrum
S/L	V	P	VE	fo (Hz)	ΔP1
		v	med	131	–
		wh	med	–	lack of P1
1	/i/	–	–	–	–
↗		–	–	–	–
		–	–	–	–
		v	med	131	lower
		wh	med	–	reference
2	/y/	v	med	330	comparable
↗		v	med	494	higher
		v	high	587	higher
		v	med	131	lower
		wh	med	–	reference
3	/e/	v	med	330	comparable
↗		v	med	494	comparable
		v	high	294	(higher)
		v	med	131	lower
		wh	med	–	reference
4	/ø/	v	med	262	comparable
↗		v	med	294	(higher)
		v	high	330	higher

Woman					
Sounds					Spectrum
S/L	V	P	VE	fo (Hz)	ΔP1
		v	med	220	lower
		wh	med	–	reference
9	/i/	v	med	440	comparable
↗		v	med	587	higher
		v	high	587	higher
		v	med	220	lower
		wh	med	–	reference
10	/y/	v	med	440	comparable
↗		v	med	659	higher
		v	high	659	higher
		v	med	220	comparable
		wh	med	–	reference
11	/e/	v	med	392	comparable
↗		v	med	523	higher
		v	high	392	higher
		v	med	220	comparable
		wh	med	–	reference
12	/ø/	v	med	392	comparable
↗		v	med	587	higher
		v	high	440	higher

Child					
Sounds					Spectrum
S/L	V	P	VE	fo (Hz)	ΔP1
		v	med	262	lower
		wh	med	–	reference
17	/i/	v	med	494	comparable
↗		v	med	659	higher
		v	high	784	higher
		v	med	262	comparable
		wh	med	–	reference
18	/y/	v	med	494	comparable
↗		v	med	659	higher
		v	high	698	higher
		v	med	262	comparable
		wh	med	–	reference
19	/e/	v	med	294	comparable
↗		v	med	330	higher
		v	high	330	higher
		v	med	262	comparable
		wh	med	–	reference
20	/ø/	v	med	294	comparable
↗		v	med	330	comparable
		v	high	294	comparable

Table 1 (continuation). [M-05-02-T01]

Man					
Sounds					Spectrum
S/L	V	P	VE	fo (Hz)	ΔP1/ΔF1
		v	med	131	lower
		wh	med	–	reference
5	/ɛ/	v	med	196	comparable
↗		v	med	392	higher
		v	high	349	higher
		v	med	131	lower
		wh	med	–	reference
6	/a/	v	med	494	comparable
↗		v	med	392	higher
		v	high	659	higher
		v	med	131	lower
		wh	med	–	reference
7	/o/	v	med	294	comparable
↗		v	med	392	comparable
		v	high	262	comparable
		v	med	131	lower
		wh	med	–	reference
8	/u/	v	med	392	comparable
↗		v	med	494	(higher)
		v	high	440	(higher)

Woman					
Sounds					Spectrum
S/L	V	P	VE	fo (Hz)	ΔP1/ΔF1
		v	med	220	lower
		wh	med	–	reference
13	/ɛ/	v	med	494	comparable
↗		v	med	523	comparable
		v	high	659	higher
		v	med	220	lower
		wh	med	–	reference
14	/a/	v	med	392	comparable
↗		v	med	523	comparable
		v	high	698	higher
		v	med	220	comparable
		wh	med	–	reference
15	/o/	v	med	247	comparable
↗		v	med	330	comparable
		v	high	494	comparable
		v	med	220	lower
		wh	med	–	reference
16	/u/	v	med	440	comparable
↗		v	med	784	higher
		v	high	880	higher

Child					
Sounds					Spectrum
S/L	V	P	VE	fo (Hz)	ΔP1/ΔF1
		v	med	262	lower
		wh	med	–	reference
21	/ɛ/	v	med	392	comparable
↗		v	med	523	comparable
		v	high	659	higher
		v	med	262	lower
		wh	med	–	reference
22	/a/	v	med	523	comparable
↗		v	med	587	comparable
		v	high	523	comparable
		v	med	262	comparable
		wh	med	–	reference
23	/o/	v	med	330	comparable
↗		v	med	440	comparable
		v	high	349	higher
		v	med	262	lower
		wh	med	–	reference
24	/u/	v	med	494	comparable
↗		v	med	587	(higher)
		v	high	698	higher

M5.3 Vowel Spectrum and Phonation Type III – Vowel Synthesis Based on Estimated Formant Patterns of Single Natural Vowel Sounds With Variation of Source Characteristics in Synthesis, and Related Vowel Recognition

Introduction

Assuming that estimated *P*-patterns and *F*-patterns of natural vowel sounds produced with voiced and creaky phonation are comparable but that the patterns of whispered sounds increase in frequency levels, then both a synthesis of the creaky- and voiced-related patterns with a noise source and a synthesis of the whispered-related patterns with a voiced source should affect vowel quality recognition. (For the use of the term synthesis here, see below.) A corresponding indication was given by Katz and Assmann (2001) in an experiment in which synthesised sounds based on *F*-patterns of natural voiced vowel sounds produced by men, women and children were produced with the two sources of pulse excitation and white noise. Vowel recognition rates proved to be lower for the noise-excited than the pulse-excited sounds, with the differences between noise-excited and pulse-excited vowel sounds being more pronounced for men than women and children. However, as explained, drawing such a straightforward conclusion does not take into account the relation of the lower spectrum of natural voiced vowel sounds to f_o. Again: The phonation-specific differences and similarities of the *P*-patterns and *F*-patterns reported in the literature generally relate to comparisons with voiced sounds at lower f_o levels of relaxed speech only. Yet, as the experiment results discussed in the previous chapter indicated, the assumed differences in vowel quality-related P_1/F_1 can disappear or even be inverted with increasing f_o of the voiced sounds of comparison.

In light of this, a vowel synthesis experiment was conducted addressing two interrelated questions:

– If an *F*-pattern is created for a Klatt synthesiser in such a way that the spectral envelope of the synthesised sound is comparable to the envelope of a natural whispered sound, and if the vowel quality of the synthesised sound with noise as the source is recognised as that of the natural whispered sound imitated, what effect does the exchange of the noise source with a voiced-like source at various levels of f_o (including a very low level imitating creaky phonation) have on vowel recognition?

- Similarly, if an *F*-pattern is created for a Klatt synthesiser in such a way that the spectral envelope of the synthesised sound is comparable to the envelope of a natural creaky sound, and if the vowel quality of the synthesised sound with a very low *f*o as the source is recognised as that of the imitated natural creaky sound, what is the effect of *f*o variation, and what is the effect if the voiced-like source is exchanged with noise?

Note that numerous studies addressed the question of the reconstruction of phonated speech from whispered speech (for overviews, see McLoughlin et al., 2015; Konno et al., 2016). However, to the best of our knowledge, no results are reported for monophthongs produced in V context and *f*o variation in synthesis as is the subject of the present experiment.

Experiment

Selection of speakers and natural sounds as references: Based on the Zurich Corpus, for each of the eight long Standard German vowels, sound triplets of a man and a woman were compiled that included one whispered, one creaky and one voiced sound. All sounds were produced in V context with medium vocal effort and in nonstyle mode, and they were fully recognised according to the results of the standard listening test conducted when creating the corpus (100% vowel recognition rate matching vowel intention). Intended *f*o for the voiced sounds related to 131 Hz for the man and 220 Hz for the woman, comparable to gender-specific average *f*o levels as given in formant statistics for adult speakers. As a result, for each speaker and each of the eight vowels, three natural reference sounds were compiled, resulting in a sample of 48 reference sounds (24 sounds per speaker).

***F*-pattern estimation:** Acoustic analysis accorded to the standard procedure of the Zurich Corpus. In a visual examination of the spectrum, the spectrogram, the formant tracks and the LPC filter curve of single sounds, *F*-patterns for subsequent Klatt synthesis were estimated based on rules as described in the literature (see Hillenbrand et al., 1995, for voiced and creaky sounds; see Sharifzadeh et al., 2012, for whispered sounds): If the average values of the LPC analysis and the corresponding formant tracks based on standard parameters (12 poles for men, 10 poles for women, frequency range of 0–5.5 kHz) were interpreted as approximately corresponding to the spectral envelope and the spectrogram of a sound, these values were set as Klatt filter parameters. Otherwise, average values of LPC analysis based on non-standard parameter settings (10 poles for men, 12 poles for women) were set as

Klatt filter parameters. If additional corrections were needed to approximate the estimated filter curve for synthesis with the spectral envelope of the natural reference sound, these corrections were added manually.

As a result, for each speaker and each of the eight vowels, three *F*-patterns related to the three natural reference sounds were assessed, resulting in a sample of 48 *F*-patterns (24 patterns per speaker; see Table 1 in the chapter appendix).

Again, the estimation of *F*-patterns often proved to be difficult, above all for whispered vowel sounds, and manual corrections were then needed in order to approximate the estimated filter pattern for synthesis to the spectral envelope of the natural reference sound. Accordingly, methodological substantiation of the estimates again was in question. Because of this, and because source characteristics, including phonation types, were also altered in the experiment, the sound production in the experiment was termed synthesis and not resynthesis.

Source characteristics used in synthesis: For vowel synthesis, every single *F*-pattern was combined with six different source characteristics, noise and five levels of f_o of 65–131–220–262–393 Hz (*F*-patterns related to the sounds of the man) and 65–165–220–262–440 Hz (*F*-patterns related to the sounds of the woman), respectively. An f_o of 65 Hz was investigated in terms of imitating creaky sounds and f_o of 131 Hz (man) and 220 Hz (woman) were investigated referring to f_o of the natural voiced reference sounds and to gender-specific average f_o levels as given in formant statistics for adult speakers. As a result, for each speaker, each of the eight vowels and each of the three phonation-related *F*-patterns, six combinations of *F*-pattern and source characteristics were created, resulting in a sample of 288 combinations (144 combinations per speaker; see Tables 2 and 3 in the chapter appendix).

Klatt synthesis: 288 sounds were produced related to the above 288 combinations of *F*-patterns and source characteristics using the KlattSyn synthesiser (cascade mode). Non-default parameters of the synthesiser were: Sound duration = 1 sec., fading = 0.05 sec., flutter = 0, aspiration = -50 dB. (Note that the default for breathiness is -25.) Whispered phonation was imitated with white noise, and f_o of voiced phonation was set as given above.

Listening test: Vowel recognition of the synthesised sounds was tested in two speaker-blocked subtests according to the standard procedure of the Zurich Corpus and involving the five standard listeners. However, the sound presentation was experiment-specific: One test item contained two sounds (separated by a 1 sec. pause), the first sound being the natural reference sound and the second sound being one of the six synthesised versions (*F*-pattern related to the natural reference sound, source characteristics set as one of the six options). The listeners were then asked to label the second sound only.

Results

Table 1 in the chapter appendix lists the natural reference sounds and their assessed *F*-patterns used for synthesis. Table 2 shows the detailed vowel recognition results, and Table 3 summarises the recognition results according to the three phonation types of the natural reference sounds. The results are discussed in general terms for recognition rates ≥ 60%. Note that (i) the vowel qualities /ɛ/ and /ə/ were not interpreted as marked quality differences, (ii) a labelling majority for a vowel boundary is indicated with two vowels given without a space, and (iii) labelling without a majority but mostly involving two adjacent vowel qualities is indicated with the two vowels in question separated by a hyphen. In Table 3 (see Columns 4 to 9), the results for the synthesis with source variation are given in the following order of source characteristics: creaky-like (f_o = 65 Hz), voiced-like up to f_o = 220 Hz, whispered-like (noise), voiced-like from f_o = 262 Hz upwards. This order accords with the interpretation given below that the best vowel recognition matches were found for whispered-related *F*-patterns with either a whispered-like or a voiced-like source at an f_o of 262 Hz of vowel synthesis.

Validation of *F*-patterns used for synthesis (Tables 3a to 3c, see Columns "rep"): All synthesised sounds for which both source and filter characteristics corresponded to the natural reference sounds were recognised according to vowel intention and vowel recognition of the natural references (compare Columns "nat" and "rep"). In these terms, the estimated *F*-patterns used for synthesis were validated. Because of the cascade mode used for the Klatt synthesiser, the spectral slope of the synthesised sounds with noise as the source differed from that of the natural whispered sounds. The possible effect of this difference on vowel recognition was not investigated. However, since there was no vowel confusion for synthesised sounds that replicated natural whispered sounds, the effect was assumed to be marginal.

Exchanging source characteristics for the synthesis of creaky-related *F*-patterns and the effect on vowel recognition (Table 3a): Concerning synthesis based on creaky-related *F*-patterns,

- a noise source affected only two sounds of the man in terms of a vowel quality recognition shift in an open–close direction (see Table 3a, Columns “65” and “noise”, Series 4 and 5); however, only one of these shifts exceeded a vowel boundary shift;
- a voiced source at f_o of 131 Hz (man) and 220 Hz (woman) showed no effect on vowel recognition (see Table 3a, Columns “65” and “131” or “220”, respectively);
- a voiced source at f_o of 262–393 Hz (man) and 262–440 Hz (woman) had an effect on vowel recognition for all sounds of both speakers for close-mid vowels and an effect for the sound of /ɛ/ of the man in terms of a shift in an open–close direction (see Table 3a, Columns “65” and “262–393” or “262–440”, respectively, Series 4–7 and 12–14).

Furthermore, a single and inconstant intermediate vowel recognition shift occurred for the sound of /a/ at f_o of 262 Hz in Series 8.

Exchanging source characteristics for the synthesis of voiced-related *F*-patterns and the effect on vowel recognition (Table 3b): Concerning synthesis based on voiced-related *F*-patterns (for statistically average f_o levels of 131 Hz for men and 220 Hz for women),

- a low level of f_o of 65 Hz imitating creaky phonation affected three sounds of the woman in terms of a shift in an open–close direction for increasing levels of f_o from low to high (see Table 3b, Columns “65” and “220”, Series 9, 10, 14); however, only the shifts for the two close vowels exceeded a vowel boundary; note also that the f_o distance between the level of the voiced reference sounds and 65 Hz is much more pronounced for sounds of the woman (220 - 65 Hz = 155 Hz) than for sounds of the man (131 - 65 Hz = 66 Hz);
- a noise source affected two sounds of the man in terms of a shift in an open–close direction (see Table 3b, Columns “131” and “noise”, Series 6 and 8); however, these shifts only concerned vowel boundaries;
- the levels of f_o of 220–393 Hz (man) and 262–440 Hz (woman) had an effect on vowel recognition for all sounds of close-mid vowels and of /ɛ/ of the man and for the sounds of the close-mid vowels of the woman in terms of a shift in an open–close direction (see Table 3b, Columns “131” and “220–262–393”, Series 4–7, and Columns “220” and “262–440”, Series 12–14).

As was the case for the sounds of creaky-related *F*-patterns, inconstant intermediate vowel recognition shifts (here only in terms of boundary shifts) with increasing f_o occurred for the sounds of /a/ in Series 8.

Exchanging source characteristics for synthesis of whispered-related *F*-patterns and the effect on vowel recognition (Table 3c): Concerning synthesis based on whispered-related *F*-patterns,

- a voiced source and a low level of f_o of 65 Hz imitating creaky phonation affected the sounds of all close-mid vowels and of /u/ for both speakers and, in addition, had an effect on the sounds of /y/ of the woman in terms of a shift in an open–close direction, the shift direction viewed from 65 Hz to noise (see Columns "65" to "noise", Series 3–6 and 10–14);
- a voiced source and levels of f_o of 131–220 Hz (man) and 165–220 Hz (woman) affected the sounds of /u, e, ø/ of the man and the sounds of /y, u, ɔ/ of the woman in terms of a shift in an open–close direction, the shift direction viewed from 131 and 220 Hz to noise (see Columns "131", "220" and "noise", Series 3–5 and 10, 11 and 14); however, only the shifts for the sounds of /u/ of both speakers and /o/ of the woman exceeded a vowel-boundary shift;
- a voiced source at an f_o of 262 Hz (both speakers) had no effect on the sounds except for the reference sound of /u/ of the man; for this reference sound, only the synthesis at an f_o level of 393 Hz produced a recognition of /u/;
- a voiced source and higher levels of f_o of 393 Hz (man) and 440 Hz (woman) affected the sounds of /e, o/ of the man and the sounds of /e, ø, o/ of the woman in terms of a shift in an open–close direction with increasing f_o levels (see Columns "noise" and "393" and "262–440", respectively, Series 4 and 6, and 12–14, respectively).

Thus, the best correspondence for vowel recognition of synthesised sounds based on whispered-related *F*-patterns but applying a voiced source was found for f_o of 262 Hz for both speakers (see Columns "noise" and "262", all series): Except for one sound only, synthesised sounds applying noise or voiced source characteristics were recognised similarly, and the remaining synthesised voiced sound was recognised similarly to the whispered sound at an f_o of 393 Hz (Series 3).

Discussion

The primary general indication given by the results concerned synthesised sounds related to *F*-patterns of natural whispered vowel sounds: If, in synthesis, these *F*-patterns were combined with either a creaky-like source (here voiced-like source with low f_o = 65 Hz) or a voiced-like source with gender-specific f_o levels as given in formant statistics for relaxed speech in adults (here f_o = 131 Hz for the man and 220 Hz for the woman), vowel recognition proved to be impaired for the sounds of some vowels, that is, vowel confusions in terms of recognised vowel qualities deviating from vowel intention occurred. But if these *F*-patterns were either combined with noise or with a voiced-like source at an intermediate f_o level, higher than the one given in formant statistics for relaxed speech – here f_o = 262 Hz for the patterns of both speakers, except for one sound with f_o = 393 Hz –, vowel quality was maintained for all sounds investigated independent of the source characteristic being noise or voiced (compare Table 3c, Columns 7 and 8, the latter highlighted with a red arrow). Notably, this intermediate level of f_o with the best match for vowel recognition was found for the sounds of both the man and the woman. (For the links to the tested sounds, see Chapter 5.3.)

This indication of vowel synthesis and vowel recognition paralleled the indication of the spectral comparison of natural whispered and voiced sounds reported in the previous chapter, that the lower part of the spectral envelope of whispered vowel sounds tended to correspond to the envelope of voiced sounds produced at intermediate levels of f_o of a speaker's vocal range. (Note in this context that, for *F*-patterns related to voiced vowel sounds, the replacement of voiced source characteristics with white noise was reported to affect the sounds produced by men more strongly than the sounds produced by children; see the above-cited study of Katz and Assmann, 2001. Note also the gender- and age-related differences in the results of the previous experiment discussed in Chapter M5.2.)

But how should these findings be understood?

Although whispered vowel sounds are commonly understood to have no periodicity, according to the literature, they are often perceived as having a pitch (for references, see Chapter M6.1). Thus, one possible explanation of the above findings is that pitch recognition of whispered sounds is comparable to pitch recognition of voiced sounds, and that the equivalent pitch level for the sounds investigated here was indicated as lying near 262 Hz, this level understood as a rough approximation. (Note that, according to the literature, recognised pitches of whispered sounds can vary to some extent. Here, the frequency estimate of the pitches of the whispered sounds is given only for the sounds investigated.) If this is the case, for perception and acoustics, the link between whispered and voiced vowel sounds is not measured f_o but perceived pitch (or if not pitch, then a comparable perceived sound characteristic; for this differentiation, see the following sixth main chapter). Also, if this is the case, f_o and pitch must be held apart when investigating vowel sounds, a matter that is addressed in the excursus on fundamental frequency and pitch (see Part II).

As mentioned in the introduction to Chapter M5.1, phonation-related spectral differences for whispered and voiced vowel sounds are commonly explained to be a result of different vocal tract resonance patterns due to physiological differences at the glottis level for the two phonation types. However, without taking into account the relation of the lower vowel spectrum to f_o – and possibly an indicated pitch-relation of the vowel spectrum – such a generalised statement is called into question and has to be clarified in future research.

Besides whispering, exchanging source characteristics for creaky- and voiced-related *F*-patterns had no substantial effect on the vowel recognition for sounds of the man, if f_o levels of the synthesised voiced sounds corresponded to the levels as given in formant statistics (see Table 3a, Columns "rep" and "131", and Table 3b, Columns "65" and "rep"), but it affected the sounds of one close-mid and two close vowels of the woman in terms of an open–close vowel quality shift with increasing f_o (and pitch; see Table 3a, Columns "rep" and "220" with no effect, and Table 3b, Columns "65" and "rep" with an effect on the three sounds of Series 9, 10 and 14). More pronounced open–close shifts occurred for the voiced source condition with f_o levels of 262–393 Hz (sounds produced by the man) and 440 Hz (sounds produced by the woman). In general, this finding supports the notion that the (lower) vowel spectrum is related to f_o (and/or pitch): The observed open–close vowel quality shifts that occurred for increasing f_o levels from low to high were expected on the basis of the previous experiments and results reported in this treatise, both in terms of the general direction as well as the nonuniform character of the shifts, above all depending on the vowel qualities and the levels and ranges of f_o examined.

The present experiment was based on the sounds of two speakers only. Possible speaker-related variations in phonation and articulation, as well as vocal effort variation, were not investigated. These aspects have to be taken into account when interpreting the results and their discussion, and they have to be investigated in more detail in future examinations.

However, in terms of summary and repetition: If the sounds investigated here were set in a pitch level order of 65–131–220–noise/262–393–440 Hz, the pitch at 65 Hz considered to be creaky-like and the pitch of the sounds with noise as a source presumed to be perceived near 262 Hz, then the recognition results for the synthesised sounds of all *F*-patterns obtained accorded to the rule for vowel recognition of either being maintained or shifting (in a nonuniform manner) in an open–close direction with increasing pitch from low to high levels.

Chapter Appendix

Note that, as was the case in the text, a pitch-related order of presentation is applied in the tables: creaky-like (f_o = 65 Hz), voiced-like up to f_o = 220 Hz, whispered-like (noise), voiced-like from f_o = 262 Hz upwards.

Table 1. Creaky, voiced and whispered natural reference vowel sounds and related estimated *F*-patterns for Klatt synthesis. First part of the table = sounds of the man; second part of the table = sounds of the woman. Columns 1–3 = natural reference sounds (V = intended and recognised vowel quality; P = phonation type, where c = creaky, v = voiced, w = whispered; Ref = number of the reference sound in the Zurich Corpus). Column 4 = LPC filter parameters used for *F*-pattern calculation (Par; P6 = 12 poles, P5 = 10 poles, frequency range = 5.5 kHz). Column 5 = indication of manual correction (MC; “x” = manual correction applied). Column 6 = resulting Klatt filter parameters used for synthesis (formant frequencies, bandwidths and levels; levels not used in cascade mode of Klatt synthesis). For reproduction in the Zurich Corpus, use the KlattSyn tool. For creaky-like and voiced-like source, combine the following link header with one of the Klatt synthesis parameters indicated in the table:
https://www.phones-and-phonemes.org/tools/klattSyn/#f0=<<add f_o value>>&flutterLevel=0&cascadeAspirationDb=-50&<<add Klatt synthesis parameters>>
For whispered-like source, combine the following link header with one of the Klatt synthesis parameters indicated in the table (please ignore the default parameters if you only access directly via the header):
https://www.phones-and-phonemes.org/tools/klattSyn/#glottalSourceType=noise&flutterLevel=0&cascadeAspirationDb=-50&<<add Klatt synthesis parameters>>
[M-05-03-T01]

Table 2. Synthesised vowel sounds related to *F*-patterns of natural creaky, voiced and whispered reference sounds, with a variation of source characteristics: Vowel recognition results (details). Columns 1–5 = natural reference sounds (S = series number; V = vowel intended; Ref = number of the reference sound in the Zurich Corpus; P/fo = phonation type imitated, where c = creaky, v = voiced with f_o = 131 Hz for the sounds of the man and 220 Hz for the sounds of the woman, w = whispered; VR = recognised vowel quality). Columns 6–11 = recognised vowel qualities (vowel recognition rate ≥ 60%) for the synthesised sounds (f_o given in Hz). Online table only: Columns 12–18 = details of the vowel recognition results in terms of the five labellings of the five listeners, given in phonetic order. Colour code: Dark red = vowel quality shifts in an open–close direction with increasing f_o and/or pitch level; light red = vowel boundary shifts in an open–close direction with increasing f_o and/or pitch level; purple = inverted vowel quality shifts in a close–open direction after an occurring open–close shift.
[M-05-03-T02]

Table 3. Synthesised vowel sounds related to *F*-patterns of natural creaky, voiced and whispered reference sounds, with a variation of source characteristics: Vowel recognition results (summary). Results are given separately according to the *F*-patterns of the three phonation types of the natural reference sounds: Creaky phonation (Table 3a), voiced phonation (Table 3b) and whispered phonation (Table 3c). Columns 1–3 = natural reference sounds (SP = speaker and speaker group; S = series number; V = intended and recognised vowel quality). Columns 4–9 = vowel recognition of the synthesised sounds with varied source characteristics. Colour code: Dark green = vowel recognition of the synthesised sound matching vowel intention and recognition of the natural reference sound; dark red = vowel quality shifts in an open–close direction with increasing f_o and/or pitch level for the synthesised sound when compared with the natural reference sound; light red = vowel boundary shifts in an open–close direction with increasing f_o and/or pitch level for the synthesised sound; purple = inverted vowel quality shift in a close–open direction for the synthesised sound succeeding an occurring previous open–close shift. Marks: Rep = source parameters of synthesis correspond to source characteristics of the natural reference sounds; black arrow = source characteristics correspond to the alternative two source characteristics of the natural reference sounds of investigation; red arrow = see text.
[M-05-03-T03]

Table 1. Creaky, voiced and whispered natural reference vowel sounds and related estimated F-patterns for Klatt synthesis. [M-05-03-T01]

Natural sounds			F-pattern		Klatt filter parameters applied
V	P	Ref	Par	MC	(sounds of the man; fo = 131 Hz)
i	c	135904	P6		f1=238/82/0&f2=2095/87/-21&f3=3086/123/-15&f4=3598/207/-17&f5=3964/178/-20
	v	135207	P6		f1=262/44/0&f2=2043/112/-33&f3=2854/146/-26&f4=3204/299/-31&f5=4224/1301/-48
	w	173175	P6	x	f1=280/150/-9&f2=2450/100/-5&f3=3250/120/-3&f4=3600/90/0&f5=4900/700/-9
y	c	135898	P6	x	f1=219/54/0&f2=1667/110/-19&f3=1872/100/-20&f4=3066/289/-31&f5=3828/203/-46&f6=4706/1287/-53
	v	135243	P6	x	f1=264/70/0&f2=1530/50/-21&f3=1827/90/-27&f4=2824/296/-40&f5=3247/207/-43
	w	173183	P6	x	f1=280/150/-3&f2=2000/160/-2&f3=2252/121/0&f4=3241/131/-9&f5=3903/984/-23&f6=5164/1125/-35
u	c	135900	P6		f1=236/96/0&f2=631/132/-16&f3=2075/586/-50&f4=3014/199/-44&f5=3666/96/-48&f6=5020/569/-67
	v	135159	P6	x	f1=288/70/0&f2=681/54/-13&f3=2293/702/-50&f4=2929/162/-44&f5=3408/170/-51&f6=4870/1293/-76
	w	135837	P6	x	f1=414/90/0&f2=799/90/-4&f3=2174/224/-26&f4=3191/265/-28&f5=3719/326/-26&f6=5201/410/-39
e	c	173184	P6		f1=346/88/0&f2=2390/102/-12&f3=2862/252/-18&f4=3837/248/-14&f5=4085/301/-15
	v	135195	P6	x	f1=375/52/0&f2=2060/100/-17&f3=2531/100/-22&f4=3393/247/-23&f5=3622/336/-28
	w	135840	P5	x	f1=487/104/0&f2=2428/163/-5&f3=2958/450/-9&f4=3677/280/-11&f5=4671/950/-24
ö	c	135906	P6		f1=314/48/0&f2=1334/64/-14&f3=1899/144/-25&f4=2954/232/-36&f5=3822/199/-47&f6=4364/1158/-49
	v	135231	P6	x	f1=331/60/-8&f2=1423/102/0&f3=2062/150/-13&f4=2846/728/-25&f5=3320/113/-22
	w	135843	P6	x	f1=500/85/-6&f2=1692/55/0&f3=2108/154/-8&f4=3080/125/-18&f5=4319/296/-32
o	c	173187	P6		f1=409/44/0&f2=665/92/-13&f3=2703/253/-45&f4=3406/166/-42&f5=4015/1035/-48&f6=4943/1268/-60
	v	135171	P6	x	f1=350/50/0&f2=593/64/-10&f3=2338/153/-17&f4=2971/338/-18&f5=3267/307/-27&f6=5185/91/-32
	w	173181	P6	x	f1=450/88/0&f2=800/300/-17&f3=2803/152/-23&f4=3282/138/-23&f5=4604/940/-42&f6=5381/185/-23
ä	c	135896	P6	x	f1=598/91/0&f2=1624/117/-12&f3=2484/313/-32&f4=3457/397/-31&f5=4206/237/-38&f6=4337/299/-37
	v	135219	P6		f1=580/73/-7&f2=1552/109/0&f3=2536/176/-11&f4=3436/345/-17&f5=3710/436/-20
	w	135860	P6		f1=736/109/0&f2=1767/172/-2&f3=2087/1337/-2&f4=2724/191/-8&f5=3641/417/-23&f6=4717/609/-34
a	c	135902	P6		f1=646/190/0&f2=1102/139/-3&f3=2637/303/-27&f4=3568/219/-27&f5=4385/247/-37
	v	135183	P6		f1=598/111/0&f2=1097/77/-4&f3=2489/197/-27&f4=3360/503/-37&f5=3726/225/-36
	w	135839	P6		f1=981/144/0&f2=1383/200/-7&f3=2628/288/-13&f4=3182/321/-22&f5=3958/231/-30&f6=5188/495/-30

Table 1 (continuation). [M-05-03-T01]

Natural sounds			F-pattern		Klatt filter parameters applied
V	P	Ref	Par	MC	(sounds of the woman; fo = 220 Hz)
i	c	141369	P5	x	f1=250/50/0&f2=2676/78/-13&f3=3586/439/-19&f4=4077/229/-19&f5=4306/2206/-18
	v	141497	P6	x	f1=320/50/0&f2=2430/100/-16&f3=3250/200/-19&f4=4100/120/-17&f5=4299/120/-17
	w	141324	P5	x	f1=270/200/-20&f2=2942/82/0&f3=3764/185/-3&f4=4270/250/-4&f5=4903/816/-12
y	c	160353	P5		f1=285/44/0&f2=1760/168/-21&f3=2217/155/-20&f4=3481/135/-34&f5=5039/285/-45
	v	141542	P6	x	f1=330/50/-17&f2=1925/100/0&f3=2500/150/0&f4=3500/100/-5&f5=4200/100/-11&f6=4800/50
	w	160347	P6	x	f1=380/30/-7&f2=2181/101/0&f3=2500/100/-7&f4=3565/880/-10&f5=3979/282/-10&f6=4828/245/-12
u	c	160356	P5	x	f1=250/70/0&f2=550/70/-11&f3=2381/845/-57&f4=3785/551/-59&f5=5081/267/-50
	v	141437	P5		f1=328/132/-9&f2=639/348/0&f3=2497/459/-32&f4=3783/635/-39&f5=4770/119/-27
	w	141320	P5	x	f1=400/50/0&f2=790/40/0&f3=2704/711/-31&f4=3792/584/-30&f5=4704/191/-24
e	c	141368	P5		f1=430/33/0&f2=2525/106/-18&f3=3104/153/-24&f4=4346/239/-28
	v	141482	P5	x	f1=437/14/0&f2=2444/80/-16&f3=3007/70/-21&f4=4350/100/-20&f5=4500/100/-20
	w	141323	P5	x	f1=500/200/-19&f2=2791/118/0&f3=3502/200/-2&f4=4161/300/-7
ö	c	141362	P5		f1=418/39/0&f2=1683/35/-12&f3=2628/49/-18&f4=3800/36/-22&f5=4957/50/-26
	v	141527	P5		f1=442/20/0&f2=1474/118/-47&f3=2627/125/-37&f4=3797/70/-30&f5=4467/215/-44
	w	141326	P5	x	f1=480/100/-5&f2=1970/100/-2&f3=2974/100/0&f4=3688/270/-10&f5=5013/1321/-24
o	c	141366	P6	x	f1=400/40/0&f2=650/50/-9&f3=2750/100/-49&f4=3500/100/-54&f5=4100/100/-57&f6=4800/200/-57
	v	160533	P6	x	f1=434/32/0&f2=665/48/-8&f3=2890/70/-41&f4=3572/60/-40&f5=4100/200/-51&f6=4250/120
	w	141321	P5	x	f1=500/90/-1&f2=850/80/0&f3=2929/128/-12&f4=3608/359/-17&f5=4352/160/-17
ä	c	141361	P5		f1=812/91/0&f2=2253/117/-4&f3=3026/217/-11&f4=4385/280/-23&f5=4807/324/-19
	v	160502	P5		f1=844/99/0&f2=2087/285/-13&f3=2889/478/-25&f4=4106/424/-35&f5=4435/405/-36
	w	160341	P5		f1=1041/73/0&f2=2574/334/-9&f3=3259/421/-13&f4=4019/1003/-17&f5=5251/1058/-22
a	c	141358	P5		f1=853/70/0&f2=1303/80/-7&f3=3226/343/-30&f4=4201/1129/-36&f5=4966/880/-36
	v	141467	P6		f1=896/129/0&f2=1249/1759/0&f3=1406/210/-4&f4=3290/384/-31&f5=4227/440/-23&f6=4559/201/-27
	w	160340	P5	x	f1=1150/100/0&f2=1350/100/-1&f3=3237/392/-14&f4=3950/629/-16&f5=5119/1099/-21

Table 2: Synthesised vowel sounds related to F-patterns of natural creaky, voiced and whispered reference sounds, with a variation of source characteristics: Vowel recognition results (details). [M-05-03-T02]. Extended online table: ☐

Natural reference sounds of the man					Vowel recognition ≥ 60% of the synthesised sounds					
S	V	Ref	P/fo	VR	c	v	v	w	v	v
			Hz		65	131	220	–	262	393
		135904	c	i	i	i	i	i	i	i
1	i	135207	131	i	i	i	i	i	i	i
		173175	w	i	i	i	i	i	i	i
		135898	c	y	y	y	y	y	y	y
2	y	135243	131	y	y	y	y	y	y	y
		173183	w	y	y	y	y	y	y	y
		135900	c	u	u	u	u	u	u	u
3	u	135159	131	u	u	u	u	u	u	u
		135837	w	u	o	o	o	u	o	u
		173184	c	e	e	e	e	i	i	i
4	e	135195	131	e	e	e	e	e	e	y
		135840	w	e	(ɛ–e)	(ɛ–e)	e	e	e	(e–i)
		135906	c	ø	ø	ø	(ø–y)	(ø–y)	y	(ø–y)
5	ø	135231	131	ø	ø	ø	øy	ø	(ø–y)	y
		135843	w	ø	ɛ	(ɛ–ø)	ø	ø	ø	ø
		173187	c	o	o	o	o	o	o	u
6	o	135171	131	o	o	o	o	(o–u)	(o–u)	u
		173181	w	o	ɔ	o	o	o	o	u
		135896	c	ɛ	ɛ	ɛ	ɛ	(ɛ–ə)	ø	ø
7	ɛ	135219	131	ɛ	ɛ	ɛ	ə	ə	ø	ø
		135860	w	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ
		135902	c	a	a	a	a	a	ɔ	a
8	a	135183	131	a	a	a	aɔ	(a–ɔ)	(ɔ–o)	a
		135839	w	a	a	a	a	a	a	a

Natural reference sounds of the woman					Vowel recognition ≥ 60% of the synthesised sounds					
S	V	Ref	P/fo	VR	c	v	v	w	v	v
			Hz		65	165	220	–	262	440
		141369	c	i	i	i	i	i	i	i
9	i	141497	220	i	e	e	i	i	i	i
		141324	w	i	i	i	i	i	i	i
		160353	c	y	y	y	y	y	y	y
10	y	141542	220	y	ø	ø	y	y	y	y
		160347	w	y	e	(ø–y)	–	y	y	y
		160356	c	u	u	u	u	u	u	u
11	u	141437	220	u	u	u	u	u	u	u
		141320	w	u	o	o	o	u	u	u
		141368	c	e	e	e	e	e	e	i
12	e	141482	220	e	e	e	e	e	e	i
		141323	w	e	(ɛ–e)	e	e	e	e	i
		141362	c	ø	ø	ø	ø	ø	ø	y
13	ø	141527	220	ø	ø	ø	ø	ø	ø	y
		141326	w	ø	(ɛ–ø)	ø	ø	ø	ø	y
		141366	c	o	o	o	o	o	o	u
14	o	160533	220	o	(ɔ–o)	o	o	o	o	u
		141321	w	o	ɔ	ɔ	o	o	o	(o–u)
		141361	c	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ
15	ɛ	160502	220	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ
		160341	w	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ
		141358	c	a	a	a	a	a	a	a
16	a	141467	220	a	a	a	a	a	a	a
		160340	w	a	a	a	a	a	a	a

Table 3. Synthesised vowel sounds related to F-patterns of natural creaky, voiced and whispered reference sounds, with a variation of source characteristics: Vowel recognition results (summary). [M-05-03-T03]

3a: F-patterns related to reference creaky sounds								
Natural			Resynthesis (fo in Hz)					
SP	S	V	rep	▼		▼		
			65	131	220	noise	262	393
man	1	i	i	i	i	i	i	i
	2	y	y	y	y	y	y	y
	3	u	u	u	u	u	u	u
	4	e	e	e	e	i	i	i
	5	ø	ø	ø	(ø–y)	(ø–y)	y	(ø–y)
	6	o	o	o	o	o	o	u
	7	ɛ	ɛ	ɛ	ɛ	(ɛ–ə)	ø	ø
	8	a	a	a	a	a	ɔ	a
SP	S	nat	rep		▼	▼		
			65	165	220	noise	262	440
woman	9	i	i	i	i	i	i	i
	10	y	y	y	y	y	y	y
	11	u	u	u	u	u	u	u
	12	e	e	e	e	e	e	i
	13	ø	ø	ø	ø	ø	ø	y
	14	o	o	o	o	o	o	u
	15	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ
	16	a	a	a	a	a	a	a

3b: F-patterns related to reference voiced sounds								
Natural			Resynthesis (fo in Hz)					
SP	S	V	▼	rep		▼		
			65	131	220	noise	262	393
man	1	i	i	i	i	i	i	i
	2	y	y	y	y	y	y	y
	3	u	u	u	u	u	u	u
	4	e	e	e	e	e	e	y
	5	ø	ø	ø	øy	ø	(ø–y)	y
	6	o	o	o	o	(o–u)	(o–u)	u
	7	ɛ	ɛ	ɛ	ə	ə	ø	ø
	8	a	a	a	aɔ	(a–ɔ)	(ɔ–o)	a
SP	S	nat	▼		rep	▼		
			65	165	220	noise	262	440
woman	9	i	e	e	i	i	i	i
	10	y	ø	ø	y	y	y	y
	11	u	u	u	u	u	u	u
	12	e	e	e	e	e	e	i
	13	ø	ø	ø	ø	ø	ø	y
	14	o	(ɔ–o)	o	o	o	o	u
	15	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ
	16	a	a	a	a	a	a	a

3c: F-patterns related to reference whispered sounds								
Natural			Resynthesis (fo in Hz)					
SP	S	V	▼	▼		rep	▼	
			65	131	220	noise	262	393
man	1	i	i	i	i	i	i	i
	2	y	y	y	y	y	y	y
	3	u	o	o	o	u	o	u
	4	e	(ɛ–e)	(ɛ–e)	e	e	e	(e–i)
	5	ø	ɛ	(ɛ–ø)	ø	ø	ø	ø
	6	o	ɔ	o	o	o	o	u
	7	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ
	8	a	a	a	a	a	a	a
SP	S	nat	▼		▼	rep	▼	
			65	165	220	noise	262	440
woman	9	i	i	i	i	i	i	i
	10	y	e	(ø–y)	–	y	y	y
	11	u	o	o	o	u	u	u
	12	e	(ɛ–e)	e	e	e	e	i
	13	ø	(ɛ–ø)	ø	ø	ø	ø	y
	14	o	ɔ	ɔ	o	o	o	(o–u)
	15	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ
	16	a	a	a	a	a	a	a

M5.4 Vowel Spectrum and Vocal Effort

Introduction

The question of the acoustic differences of vowel sounds produced with different vocal efforts, including shouting, is a matter of debate (for a comprehensive overview and evaluation of studies and results, see Koenig and Fuchs, 2019, in particular, Table 1; see also Liénard and Di Benedetto, 1999; Traunmüller and Eriksson, 2000; Raitio et al., 2013). Among the main acoustic features discussed in the literature as being related to an increase in the vocal effort are increased sound pressure level (SPL), increased f_o and F_1 (in some studies also F_2), and decreased spectral slope (emphasis of higher frequencies; however, note the limitation of this decrease as indicated in the study of Ternström et al., 2006) combined with increased levels of the higher formants. These acoustic effects of vocal effort variation are generally understood as a consequence of a simultaneous change in respiratory, laryngeal and supralaryngeal behaviour (variation of subglottal pressure and vocal fold tension, adaptation of articulation). Furthermore, when compared to relaxed speech, speech with a raised voice is often reported as more intelligible, whereas shouted speech is reported as less intelligible.

However, inconsistent results have been reported concerning the effect of increased vocal effort on f_o and F-patterns. Moreover, the experimental settings varied strongly among different studies, and results were indicated to depend not only on the experimental tasks but also on the vowel qualities and the speaker's reactions to the tasks. Finally, the methodological problem of formant estimation always hampers the investigation of F-patterns (on this matter and for the present context, see Traunmüller and Eriksson, 2000; see also Birkholz et al., 2019, discussing source–filter interaction with vocal effort variation and related effects on formant measurement).

In this treatise, the question is addressed in an observational manner only to extend the documentation and discussion of phonation- and articulation-related spectral variation of vowel sounds: In a further experiment, vocal effort-related spectral differences of vowel sounds produced by men and women were investigated for two sound samples, the first sample involving sounds produced with low and high vocal efforts at f_o levels in the lower vocal range of the speakers, and the second sample involving sounds produced with low vocal effort and as shouted sounds at f_o levels in the middle vocal range of the speakers.

The study was conducted in the context of an Interspeech Show and Tell conference presentation (Maurer et al., 2019, Chapter 6.2 in the online documentation) and was then transferred to the present treatise, including an extended description of the method and a new discussion.

Note that differences or similarities of P_1/F_1 and P_2/F_2 related to a methodological substantiation of either both characteristics or of only one characteristic. For better readability, any of these estimates are again abbreviated as P_1/F_1 and P_2/F_2, respectively. Spectral comparisons and related differences that occurred additionally are discussed below.

Experiment

Selection of speakers and sounds produced at f_o levels in the lower vocal range of a speaker: On the basis of the Zurich Corpus, for each of the eight long Standard German vowels, each of the two speaker groups of men and women and each of the two production parameters, low vocal effort and high vocal effort (systematically investigated in the corpus), two sounds of different speakers produced in nonstyle mode and V context at intended f_o levels in the ranges of 131–165 Hz (men) and 220–262 Hz (women) were selected. As a result, for each speaker group, for each of the eight vowels and f_o levels in the lower vocal range of the speakers, two sounds produced with low vocal effort were compared with two sounds produced with high vocal effort, resulting in a sample of 64 sounds (32 sounds per speaker group). (For details of the investigation of vocal effort variation, including shouting, see the handbook of the corpus.)

Selection of speakers and sounds produced at f_o levels in the middle vocal range of a speaker, including shouting: Likewise, for the same vowels, the same two speaker groups and each of the production parameters, low vocal effort and shouted, two sounds of different speakers produced in nonstyle mode and V context at intended f_o levels of 330 Hz (men) and 440 Hz (women), respectively, were selected. As a result, a second sample of 64 sounds (32 sounds per speaker group) produced at f_o levels in the middle vocal range of the speakers was created.

Additional selection criteria: The selection of all sounds was based on two additional criteria, full vowel recognition in the standard listening test conducted when creating the corpus (100% recognition rate matching vowel intention) and, if observable, marked differences in the lower part of the spectrum < c. 1 kHz related to vocal effort variation.

Acoustic features investigated, assessment of spectral differences: The acoustic analysis of the sounds accorded to the standard procedure of the Zurich Corpus. Sound comparisons were based on both the visual examination of the harmonic spectra, spectrograms, formant tracks and LPC filter curves and the calculated values for SPL (in dB), f_o (in Hz), F-patterns (in Hz) and alpha ratio (in dB; alpha ratio = level difference between the average SPL of the 1–5.5 kHz frequency region and the average SPL of the 0.05–1 kHz region). For each vowel, each speaker group and each of the related comparisons of two sounds with low vocal effort and two sounds with high vocal effort, the following spectral differences were assessed:

- Increase in dB SPL; if the SPL values of both sounds with high vocal effort were above the values of both sounds with low vocal effort, SPL was assigned as increased for the sounds with high vocal effort.
- Increase of P_1/F_1; if P_1/F_1 of both sounds with high vocal effort were above P_1/F_1 of both sounds with low vocal effort, and if the maximum frequency difference related to vocal effort variation exceeded 100 Hz, P_1/F_1 was assigned as increased for the sounds with high vocal effort.
- Increase of P_2/F_2; the same estimation was made for P_2/F_2.
- Increase in L_{P2}/L_{F2}; if either L_{P2}/L_{F2} or the spectral energy in the frequency region of a commonly assumed second spectral peak of both sounds with high vocal effort were above L_{P2}/L_{F2} of both sounds with a low vocal effort, L_{P2}/L_{F2} was assigned as increased for the sounds with high vocal effort.
- General increase in higher spectral energy; if the values of the alpha ratio of both sounds with high vocal effort were above the alpha ratio of both sounds with low vocal effort, the higher spectral energy was assigned as increased for the sounds with high vocal effort.

No specification was made if the assessment and comparison of a feature was considered critical. In addition, for the vowels /o, u/, the spectra of some sounds produced with low vocal effort showed weak or absent P_2/F_2, while P_2/F_2 for the sounds produced with high vocal effort were marked. This difference was separately noted.

A note on the notation of f_o frequency ranges: On the basis of the f_o range of recognisable vowel sounds documented in this treatise, we consider sounds produced by men at f_o of c. 131–165 Hz and by women at f_o of c. 220 Hz to be in a lower vocal range, and sounds produced by men at f_o of c. 330 Hz and by women at f_o of c. 440 Hz to be in the middle of their entire vocal range. It also has to be considered that the shouted sounds produced in the middle vocal range relate to mixed voice (and not to falsetto or head voice).

Results

Table 1 in the chapter appendix shows the results of the estimated vocal effort-related spectral differences for the comparison of sounds produced at f_o levels in the lower vocal range of the speakers. Table 2 shows the corresponding results for the sounds produced at f_o levels in the middle vocal range of the speakers. Sound links are included in the tables. Table 3 shows exemplary illustrations of the finding concerning the effect of vocal effort variation on F_1.

Vocal effort variation and dB SPL (see Column 5 in Tables 1 and 2): For all sounds investigated, SPL increased with increased vocal effort. No further analysis of numerical details was conducted because the purpose of the study was to provide exemplary documentation only.

Vocal effort variation and P_1/F_1 (see Column 6 in Tables 1 and 2): For sounds with f_o levels in the lower vocal range of the speakers, maximal P_1/F_1 difference related to vocal effort variation exceeded 100 Hz for 14 out of the 16 sound comparisons. Exceptions were found for the sounds of /ø/ produced by men (see Table 1, Series 5; P_1/F_1 difficult to estimate) and for the sounds of /u/ (no P_1/F_1 difference) produced by women (see Table 1, Series 11). In contrast, for sounds with f_o levels in the middle vocal range of the speakers, maximal P_1/F_1 difference related to vocal effort variation exceeded 100 Hz for the sounds of close-mid, open-mid (if assessable) and open vowels only. The effect of vocal effort variation on P_1/F_1 was thus indicated to relate to f_o and vowel quality (vowel openness).

Vocal effort variation and P_2/F_2 (see Column 7 in Tables 1 and 2): The indication of a correlation between increased vocal effort and P_2/F_2 was either weak and inconsistent among the sound comparisons or could not be assessed based on the sound spectra. Above all, the occurring estimation problems related to a “lacking” second spectral peak < 1 kHz for sounds of back vowels produced with low vocal effort, to flat or sloping spectral envelopes for some sounds of /a/ produced with low vocal effort and single spectral peaks for sounds of that vowel produced with high vocal effort, and to flat or sloping spectral envelopes for some sounds of /ø/ and /ɛ/ produced with low vocal effort.

Vocal effort variation and spectral slope (see Columns 8 and 9 in Tables 1 and 2): For all sound comparisons, an increase in the vocal effort resulted in an increase in the alpha ratio of the sound spectrum. Moreover, for almost all sound comparisons for which L_{P2}/L_{F2} could be estimated, L_{P2}/L_{F2} was also increased for sounds with high vocal effort. The only exception was the comparison of the sounds of /ɛ/ in Table 1, Series 7. Again, for the same reasons as for the increase in db SPL, no further analysis of numerical details was conducted.

Discussion

As explained in the introduction to this chapter, the conducted study only aimed at extending the documentation and discussion of occurring phonation- and articulation-related spectral variation of vowel sounds, and the results are given here in terms of observational findings. Mostly in line with the indications given in the literature, an increase in vocal effort from low to high or shouted resulted in an increase of SPL in general and an increase of spectral energy in the higher frequency range > 1 kHz in particular (alpha ratio), the higher frequency range commonly assumed to be related to *F*2 of sounds of front vowels and /a/ and to the higher formants of sounds of all vowels. As a consequence, for almost all sound comparisons for which L_{P2}/L_{F2} could be estimated, L_{P2}/L_{F2} also increased with increasing vocal effort. Further, with only two exceptions, P_1/F_1 (or the spectral centre of gravity of the frequency range generally assumed to be related to P_1/F_1) in its turn increased markedly with vocal effort for the sounds of all eight vowels produced at f_o in the lower vocal range of the speakers. However, this only held true for the sounds of close-mid, open-mid and open vowels produced at f_o in the middle vocal range of the speakers (except for one comparison for which no estimate could be generated). For the sounds of close vowels, no such increase was observed. Thus, the increase of P_1/F_1 proved to be dependent on f_o levels and vowel qualities (see also Koenig and Fuchs, 2019, for similar findings regarding the correlation difference of *F*1 and SPL for sounds of high tense and low lax vowels). For exemplary illustration of the effect of vocal effort variation on F_1, see Table 3.

In the context of the present treatise, the sometimes striking spectral differences in the frequency region commonly assumed to be related to P_1/F_1 deserve special attention. For sounds of a vowel produced at comparable f_o levels that show vocal effort-related differences in the lower vowel spectrum, these differences were very pronounced: Estimated P_1/F_1 of most of the sounds produced with a high vocal effort surpassed P_1/F_1 of the sounds produced with a low vocal effort by 100 Hz or more, a frequency difference that approximates or equals F_1 differences of two adjacent vowel qualities as given in formant statistics. (For P_1/F_1 differences due to a combination of vocal effort and f_o variation that can equal F_1 differences of two non-adjacent vowel qualities, see Chapter M7.8.) Likewise, the finding of P_1/F_1 variation being dependent on vowel openness and f_o levels is worth noting, too. It once again pointed to the nonuniform character of vowel-related spectral characteristics.

However, some relativisations have to be made, above all when reflecting on the P_1/F_1 differences. In part, the general methodological problem of numerical estimation of spectral peak frequencies somewhat limited the reliability of P_1/F_1 assessment despite the lower f_o levels of the sounds, and resynthesising the sounds based on calculated F-patterns, using the Klatt synthesiser, did not always confirm the estimated numerical values (author's estimate; for verification, see the sound links in Table 1 and use the Klatt synthesiser in the online archive; for exemplary illustration, see Table 3, Series 9–11). Also, some vowel timbre variations occurred for the natural sounds compared. Thus, the estimated differences in P_1/F_1 seemed to be only partly due to possible spectral variation for sounds of a given vowel owing to vocal effort variation. However, some of the observed vocal effort-related differences in P_1/F_1 were still impressive in that the estimated P_1/F_1 often surpassed a frequency difference that approximates or equals F_1 differences of two adjacent vowel qualities as given in formant statistics, with vowel quality of the resynthesised sounds based on the estimated F_1-patterns and calculated f_o levels maintained (author's estimate; for exemplary illustration, see Table 3, Series 1–8).

Besides, no clear indication was found for vocal effort-related variation in P_2/F_2 in terms of increased second peak frequencies as a general result of increasing vocal effort. This finding could be related to the dichotomy of the vowel spectrum (see Preliminaries, p. 67 and p. 238; see also the Afterword).

Because of the methodological problem of spectral peak and formant frequency estimation, the sounds of children were not included in the present study. However, a corresponding documentation of children's sounds produced at f_o in the lower and middle vocal range with low and high vocal effort is given in Maurer et al. (2019, Chapter 6.2.5 in the online documentation). Roughly speaking and relating to a visual comparison of the sound spectra only, the effects of vocal effort variation for these sounds were comparable to the indications obtained in the present study.

Finally, in contrast to some interpretations given in the literature, increasing f_o is not an aspect directly linked to increasing vocal effort: Vocal effort can be varied independently of f_o variation, and f_o levels of sounds with low vocal effort can be much higher than f_o levels of sounds with high vocal effort.

Chapter Appendix

Table 1. Spectral comparison of natural vowel sounds produced with low and high vocal effort at f_o levels in the lower vocal range of the speakers. Columns 1–4 = sound production (S/L = sound series and sound links; V = intended and recognised vowel quality; fo = range of intended f_o, in Hz; VE = vocal effort). Columns 5–9 = spectral comparison (SPL = minima and maxima of SPL for a sound pair related to a vocal effort, in dB; P1/F1 and P2/F2 = differences or similarities of P_1/F_1 and P_2/F_2; LP2/LF2 = differences or similarities in the level of the second spectral peak or estimated formant or the spectral energy in the region related to a commonly assumed second spectral peak; AR = minima and maxima of the alpha ratio for a sound pair related to a vocal effort, in dB. Colour code: Red = increased values related to increased vocal effort; blue = no pronounced spectral difference estimated. Abbreviations: incr = increased; ns = not specified (no assessment made, see text); w/a = spectral peak is weak or absent; m = spectral peak is marked.
[M-05-04-T01]

Table 2. Spectral comparison of natural vowel sounds produced with low and high vocal effort at f_o levels in the middle vocal range of the speakers. Columns are as given in Table 1, with sh = shouted vowel sounds.
[M-05-04-T02]

Table 3. Exemplary illustrations for F_1 differences of natural vowel sounds related to vocal effort. Columns 1–6 = pairs of sounds compared (S/L = sound series and sound links; Ref = reference series in Table 1 (T1) or 2 (T2) from which the sounds were selected; V = intended and recognised vowel quality; SG = speaker group, where m = men, w = women; fo = range of intended f_o levels, in Hz; VE = vocal effort). Column 7 = estimated F_1, in Hz. Column 8 = vowel recognition of Klatt resynthesis (author's estimate). Series 1–8 = vocal effort-related F_1 due to possible spectral variation for sounds of a given vowel as a result of vocal effort variation, with vowel quality of the resynthesised sounds based on the estimated F-patterns and calculated f_o levels of the natural reference sounds maintained. Series 9–11 = vocal effort-related F_1 differences possibly due to F-pattern measurement problems, with vowel quality of some of the resynthesised sounds not maintained. Colour code: Red = increased estimated F_1 related to increased vocal effort; purple = changes in vowel quality for Klatt resynthesis.
[M-05-04-T03]

Table 1. Spectral comparison of natural vowel sounds produced with low and high vocal effort at fo levels in the lower vocal range of the speakers. [M-05-04-T01]

Sounds of men, lower range of fo								
Production				Spectral comparison				
S/L	V	fo (Hz)	VE	SPL (dB) min/max	P1/F1	P2/F2	LP2/ LF2	AR (dB) min/max
1 ↗	i	147–165	low	59 / 65				-37 / -24
			high	77 / 86	incr.	–	incr.	-11 / -3
2 ↗	y	131–165	low	64 / 68				-26 / -25
			high	79 / 80	incr.	incr.	incr.	-9 / -3
3 ↗	u	131–165	low	62 / 64		ns	ns	-48 / -46
			high	78 / 86	incr.	ns	ns	-34 / -32
4 ↗	e	147–165	low	66 / 71				-24 / - 20
			high	86 / 87	incr.	–	incr.	-11 / +2
5 ↗	ø	131–165	low	67 / 67	ns			-25 / - 20
			high	81 / 86	ns	–	incr.	-9 / -6
6 ↗	o	147–165	low	64 / 64		ns	ns	-44 / -43
			high	81 / 84	incr.	ns	ns	-18 / -16
7 ↗	ε	131–165	low	68 / 68				-12 / -10
			high	87 / 88	incr.	–	–	-7 / -1
8 ↗	a	131–147	low	61 / 64				-20 / -16
			high	84 / 93	incr.	–	incr.	-7 / -1

Sounds of women, lower range of fo								
Production				Spectral comparison				
S/L	V	fo (Hz)	VE	SPL (dB) min/max	P1/F1	P2/F2	LP2/ LF2	AR (dB) min/max
9 ↗	i	220–262	low	57 / 70				-38 / -34
			high	80 / 86	incr.	–	incr.	-7 / +9
10 ↗	y	220–262	low	66 / 68				-45 / -40
			high	76 / 78	incr.	incr.	incr.	-11 / +5
11 ↗	u	247–262	low	59 / 69		w/a		-42 / -39
			high	79 / 82	–	m	incr.	-32 / -26
12 ↗	e	220–262	low	66 / 70				-31 / -18
			high	82 / 87	incr.	–	incr.	-11 / -2
13 ↗	ø	220–262	low	60 / 65				-40 / -24
			high	81 / 84	incr.	incr.	incr.	- 11 / -8
14 ↗	o	220–262	low	62 / 73		w/a		-42 / -42
			high	86 / 88	incr.	m	incr.	-13 / -5
15 ↗	ε	220–262	low	54 / 72				-32 / -16
			high	81 / 90	incr.	–	incr.	-5 / +4
16 ↗	a	220–262	low	59 / 63		ns		-30 / -19
			high	89 / 93	incr.	ns	incr.	-4 / +2

Table 2. Spectral comparison of natural vowel sounds produced with low and high vocal effort at fo levels in the middle vocal range of the speakers. [M-05-04-T02]

Sounds of men, middle range of fo								
Production				Spectral comparison				
S/L	V	fo (Hz)	VE	SPL (dB) min/max	P1/F1	P2/F2	LP2/ LF2	AR (dB) min/max
1 ↗	i	330	low	63 / 77				-27 / -26
			sh	92 / 96	–	–	incr.	-5 / +1
2 ↗	y	330	low	69 / 83				-37 / -34
			sh	92 / 92	–	–	incr.	-10 / -5
3 ↗	u	330	low	72 / 75		w/a		-43 / -34
			sh	83 / 92	–	m	incr.	-15 / -9
4 ↗	e	330	low	68 / 77				-20 / -19
			sh	84 / 88	incr.	–	incr.	-3 / +1
5 ↗	ø	330	low	63 / 76				-27 / -18
			sh	86 / 100	incr.	–	incr.	-13 / -5
6 ↗	o	330	low	63 / 70		ns	ns	-36 / -30
			sh	99 / 99	incr.	ns	ns	-18 / -4
7 ↗	ε	330	low	72 / 73	ns	ns	ns	-15 / -10
			sh	85 / 99	ns	ns	ns	-3 / +5
8 ↗	a	330	low	67 / 74				-12 / -12
			sh	98 / 101	incr.	incr.	incr.	+7 / +13

Sounds of women, middle range of fo								
Production				Spectral comparison				
S/L	V	fo (Hz)	VE	SPL (dB) min/max	P1/F1	P2/F2	LP2/ LF2	AR (dB) min/max
9 ↗	i	440	low	64 / 71				-33 / -26
			sh	87 / 87	–	–	incr.	-7 / 0
10 ↗	y	440	low	65 / 69				-36 / -31
			sh	85 / 90	–	–	incr.	-14 / -11
11 ↗	u	440	low	77 / 77		ns		-40 / -37
			sh	88 / 98	–	ns	incr.	-17 / -13
12 ↗	e	440	low	68 / 71				-27 / -26
			sh	80 / 86	incr.	–	incr.	-1 / +5
13 ↗	ø	440	low	75 / 77		ns		-31 / -22
			sh	89 / 95	incr.	ns	incr.	0 / +2
14 ↗	o	440	low	72 / 75		ns	ns	-28 / -27
			sh	91 / 94	incr.	ns	ns	-14 / -8
15 ↗	ε	440	low	68 / 74		ns		-19 / -17
			sh	95 / 96	incr.	ns	incr.	+6 / +11
16 ↗	a	440	low	65 / 67		ns	ns	-5 / -4
			sh	98 / 100	incr.	ns	ns	+16 / +19

Table 3. Exemplary illustrations for F1 differences of natural vowel sounds related to vocal effort. [M-05-04-T03]

Sounds						Spectrum		Resynthesis
S/L	**Ref**	**V**	**SG**	**fo (Hz)**	**VE**	**F1 (Hz)**	**Par**	**V**
Marked differences in P1/F1 for sounds produced with different vocal effort, vowel quality maintained in resynthesis.								
1 ↗	1 (T1)	i	m	147–165	low	210	P6	i
					high	315	P6	i
2 ↗	4 (T1)	e	m	147–165	low	316	P5	e
					high	436	P5	e
3 ↗	8 (T1)	a	m	131–147	low	482	P6	a
					high	834	P6	a
4 ↗	6 (T1)	o	m	147–165	low	245	P6	o
					high	463	P6	o
5 ↗	12 (T1)	e	w	247–262	low	399	P5	e
					high	509	P5	e
6 ↗	13 (T1)	ø	w	220–262	low	392	P5	ø
					high	541	P5	ø
7 ↗	15 (T1)	ɛ	w	220–262	low	667	P5	ɛ
					high	833	P5	ɛ
8 ↗	16 (T1)	a	w	220–262	low	580	P5	a
					high	937	P5	a
Marked differences in P1/F1 for sounds produced with different vocal effort, vowel quality in part not maintained in resynthesis.								
9 ↗	2 (T1)	y	m	131–165	low	151	P6	y
					high	369	P6	ø
10 ↗	14 (T2)	o	w	440	low	478	P5	u
					sh	832	P5	ɔ
11 ↗	15 (T1)	ɛ	w	220–262	low	315	P5	e
					high	793	P5	ɛ

M6 Vowel Sound, Vowel Spectrum and Pitch

M6.1 Pitch Recognition Comparing Natural Whispered and Voiced Vowel Sounds for Utterances of Single Speakers

Introduction

Many listeners perceive whispered speech as having pitch and pitch variation, albeit only within a limited range. Bosker et al. (2010) summarise the corresponding research on the matter (referring to Tartter, 1989; Tartter and Braun, 1994; Higashikawa et al., 1996) as follows: Whisper can cue correct perception of tone, voicing in consonants, and emotion, although the related sound itself is commonly assumed to lack periodicity since the vocal folds do not vibrate in whisper mode. The same is true for prosody (see also Konno et al., 2016, for a review of the literature and a discussion on the matter of pitch and pitch variation in whispered speech). There are even whispered singing styles (for an example, see a traditional Burundi song in Various Artists, 2015, nr. 1, “Chant avec cithara”). Furthermore, Konnai et al. (2017) discuss different subtypes of whisper in terms of inter- and intra-speaker variation. According to their appraisal of the studies published, there is no simple and uniform whisper production type to be related to in general terms.

In the specialist literature, however, no robust references are given that define how pitch levels and pitch ranges in whispered phonation should be assessed or how they should be related to each other with regard to different subtypes of whisper. Also, in the literature, several sound characteristics are discussed as possibly producing a pitch percept, above all formant contours, spectral envelope contours, spectral centre of gravity, spectral slope, relative vowel duration and mean vowel intensity production, but no conclusive, general concept of the acoustic correlates of pitch is established (for an overview on the debate, see Heeren, 2015; Konno et al., 2016). (In this context, note also the long tradition of the investigation of whisper and pitch, from the 17th century onwards, documented by Stumpf, 1926; in his book, he also reported his own extensive investigation on the matter.)

Here, only the general notion of whisper as associated with a pitch percept is retained from the literature, but no details of the various studies on the matter and their somewhat disparate results are discussed, and no reference is made to attempts at reconstruction of continuous voiced speech from whispers (for an overview, see McLoughlin et al., 2015;

Konno et al., 2016). Instead, the indications of the studies reported in the previous main chapter will be seized in order to introduce experimental attempts within the line of argument of this treatise. However, when discussing the results of the experiments reported below, attention is paid to the indications of possible production differences in whisper (whisper subtypes) and possibly related pitch differences. Corresponding relativisations are made when it comes to general estimates.

As a segue to the investigation of the pitch of whispered vowel sounds, and from here to the investigation of the vowel–pitch relation in general, the main reason for this shift in research focus will be repeated, summarising the exposed reflections in Chapters 5.3 and 5.5 and the excursus on fundamental frequency and pitch (see Part II). In the previous main chapter, it was demonstrated that the lower part of the spectral envelope for natural whispered vowel sounds tended to correspond to the lower part of the spectral envelope of voiced sounds if – and only if – the voiced sounds were produced at intermediate levels of f_o of an adult speaker's vocal range. This finding was confirmed in vowel resynthesis since vowel quality was maintained for almost all sounds resynthesised based on *F*-patterns of natural whispered sounds and applying a voiced-like source with f_o of 262 Hz, but vowel confusions occurred for both lower and higher f_o levels applied. Against this background, and in the general course of questioning the empirical finding that the vowel-related (lower) spectrum of voiced sounds is dependent on f_o, we have presumed that perception and acoustic characteristics of whispered vowel sounds – and indeed of all vowel sounds – relate to pitch (or to a comparable perceived sound characteristic). In consequence, the present main chapter addresses this vowel–pitch relation thesis.

At this stage of investigation, some of our conclusions are only speculative and briefly sketched, and they must be considered as such for future research. Moreover, as stated, appropriate experimental exploration is not trivial. In view of this, and to introduce the experiments presented in this main chapter, a few further and summarising indications are made regarding our conjectures, experimental designs and experiences in pre-studies and during the preparation of the experiments.

With regard to the conjectures: Pitch is a phenomenon of a quasi-constant sound quality over time. In the case of voiced sounds, pitch perception and recognition refer to a quasi-periodic repetition of a vibration pattern in the sound wave. From here arises the question of a particular kind of "periodicity" of whispered sounds, to which the perceptual process also refers – a particular kind of sound pattern repetition over time – however approximate this "periodicity" may be and however difficult its measurement or estimation based on an actual sound wave may be. (Note in this context the general term of pattern matching in modern theories of pitch recognition; see de Cheveigné, 2005.) We pose the question even though it seems unpromising, since recognisable vowel sounds can be synthesised using white noise as a source (e.g. in Klatt synthesis; see Chapter M5.3). We will return to this question.

Vowel quality is, in turn, a phenomenon of a quasi-constant sound quality over time. (In Part III, we will discuss in more detail why we consider the vowel a sound quality, not an aspect of sound timbre.) To further elaborate, perception in recognising a vowel may again have to refer to some kind of sound pattern repetition over time. If this holds true, why not consider the repeating pattern of a vowel sound as always consisting of two perceptual qualities, including whispered sounds: Pitch quality as the recognition of the entire pattern being repetitive, and the quality of a single vowel as the recognition of a very specific repetitive pattern in contrast to another pattern of its kind having the same temporal extent and the same repetition rate over time. We will return to this question, too.

With regard to different experimental attempts to investigate the vowel–pitch relation: As explained in the excursus on fundamental frequency and pitch (see Part II), when directly addressing the question of a vowel–pitch relation (or its alternative), we first created experimental designs to demonstrate interrelated recognised vowel qualities and pitch levels for whispered sounds, lacking f_o. Secondly, we tried to create experimental designs to demonstrate sounds for which the recognised vowel quality and the recognised pitch related to each other, but the pitch did not correlate in a simple and direct way with all three acoustic features of measured f_o, $H1$ and HCF. In these experiments and for the corresponding sounds investigated, we expected that pitch and its relation to vowel quality could be shown as standing in contrast to acoustic features generally assumed to be features of fundamental frequency. Thirdly, during the creation and conduction of these experiments, vowel quality and/or pitch level proved to be ambiguous for some sounds examined in that, for single sounds, two qualities and/or two (or even more) pitch levels could be identified by the listeners in the listening tests. Therefore, we also addressed the matter of double-vowel and/or double-pitch recognition by not only testing dominant or prominent vowel qualities and pitch levels but also asking the listeners to label secondary qualities and/or secondary pitch levels. Fourthly, further developing this experimental approach, we created experimental designs to demonstrate parallel shifts from one vowel quality and one pitch level to another quality and another level, including a transitional phase with occurring double-vowel and/or double-pitch recognition. The experiments of the first type and their results are presented in Chapters M6.1 to M6.3, the experiments of the second type in Chapters M6.4 to M6.7, the experiment of the third type in Chapter M6.8 and the experiments of the fourth type in Chapters M6.9 and M6.10.

With regard to the experimental exploration of the vowel–pitch relation comparing voiced and whispered sounds: In the attempt to design experiments investigating the spectral characteristics of voiced and whispered sounds and their respective pitch levels, we were experiencing some basic impediments.

In the beginning, when conducting pre-studies and creating experimental designs and related listening tests, we encountered very different reactions from the standard listeners of the Zurich Corpus when asking them to identify the pitch level of a whispered sound, although the listeners were professionally trained speakers and singers and were very experienced in vowel recognition tasks. Initially, one listener even refused to assign pitch levels to whispered sounds, explaining that, in her opinion, these sounds have no pitch. On the contrary, the author of this treatise was able to recognise a marked difference in pitch levels in numerous (though not all) comparisons of two whispered sounds or of a whispered and a voiced sound of the Zurich Corpus, and in many of these cases, he was even able to recognise an associated musical interval.

Furthermore, the whispered sounds used as stimuli in the pilot studies manifested different inter- and intra-speaker sound timbres (estimate of the author and some listeners), which may generally interfere with pitch recognition (see above, Konnai et al., 2017).

Finally, finding the right experimental design to study the pitch of whispered sounds posed several methodological questions. Above all, with respect to the stimuli, we had to decide whether to investigate natural, resynthesised or synthesised sounds and whether to compare two whispered vowel sounds, or a whispered and a voiced vowel sound, or a whispered vowel sound and a sinusoid, or a whispered vowel sound and a piano sound and so forth. With respect to pitch level identification, we had to decide whether listeners should be asked to identify pitch level differences between two sounds (to identify whether one sound has a lower/higher/equal pitch compared to another sound) or whether they should be asked to identify a specific pitch level (to assign the frequency level of a sinusoid or a musical note according to a musical scale).

Since there is no methodological standard for investigating the pitch of whispered vowel sounds given in the literature, we have developed our experimental designs during preliminary listening test trials with the standard listeners of the Zurich Corpus. On this basis, we have decided to conduct three pitch recognition experiments addressing identifiable pitch level differences in (i) an intra-speaker comparison of natural whispered and voiced sounds of a given vowel based on the sound sample examined in Chapter M5.2, (ii) an intra-speaker comparison of resynthesised whispered- and voiced-like sounds of a given vowel based on the sound sample examined in Chapter M5.3, and (iii) an inter-speaker comparison (including different genders and ages of speakers) of natural whispered sounds of a given vowel and, in parallel, an inter-speaker comparison of resynthesised whispered-like sounds of a given vowel again relating to the sound samples examined in Chapters M5.2 and M5.3.

This first chapter on the matter of whisper and pitch describes and discusses the first experiment, the intra-speaker comparison of natural whispered and voiced sounds. The other two experiments are the subjects of the following chapters.

Experiment

Sound sample investigated: Based on the sample of whispered and voiced sounds of a man, a woman and a child investigated in Chapter M5.2, for each of the three speakers and each single vowel quality, three sounds were selected: The whispered sound, the voiced sound produced at low intended f_o level (131 Hz for the man, 220 Hz for the woman and 262 Hz for the child) and the voiced sound produced with medium vocal effort at the highest intended f_o level (see Table 1 in Chapter M5.2, sounds 1, 2 and 4 of a speaker and a vowel).

Sound comparison (preparation for the listening test): For each speaker and each vowel, the whispered sound was compared with either the voiced sound at low f_o or the voiced sound at high f_o, in AB and BA order. Note that no sound of /i/ of the man produced with a medium vocal effort at the highest intended f_o was included because the original sample did not contain a voiced sound at high f_o (for details, see Chapter M5.2). As a result, 94 whispered–voiced or voiced–whispered sound pairs (30 pairs for the subsample of the man and 32 pairs each for the subsamples of the woman and the child) were created for the listening test (see Table 1 in the chapter appendix).

Listening test: Pitch recognition of the sounds was investigated in a listening test according to the standard procedure of the Zurich Corpus and involving the five standard listeners of the corpus, with the following experiment-specific adaptation: Each single test item contained one whispered and one voiced sound (low or high f_o level) of the same vowel and produced by the same speaker (separated by an approximately 1 sec. pause, somewhat depending on the onset and offset of the sounds), in AB (whispered–voiced) and BA (voiced–whispered) order. Three speaker-blocked subtests were conducted, in which the listeners were asked to identify the pitch level difference between the first and the second sound as falling, flat (no marked level difference) or rising, referring to dominant or prominent pitches.

A note on the adaptation of the recognition task: As mentioned in the excursus on fundamental frequency and pitch (see Part II), during the creation and in the process of the experiments investigating synthesised vowel sounds, the listeners reported that they could sometimes recognise two (or even more) pitch levels and/or two vowel qualities. Because of this, we adapted the recognition tasks for the present and the subsequent experiments: Unless otherwise specified, the listeners were asked only to label the dominant or prominent vowel quality and/or pitch level (forced choice).

Results

Table 1 in the chapter appendix shows the pitch recognition results for the three speaker-specific subsamples. Pitch level differences are given for a voiced sound being recognised as lower or higher than or equal to the whispered reference sound of comparison. In the following paragraphs, f_o levels are given in terms of intended f_o, and pitch recognition results are given according to the labelling majority.

Comparison of the sounds of the man: For all sound comparisons of the man, the pitch level of the whispered sounds was recognised above the level of the voiced sounds produced at a lower f_o of 131 Hz (see Table 1, Columns 6–8, results marked in blue). Inversely, the pitch level of the whispered sounds was recognised as below the level of the voiced sounds produced at higher f_o, with a frequency range of these higher levels of 294–494 Hz, depending on vowel quality (see Table 1, Columns 6–8, results marked in red).

Comparison of the sounds of the woman: For the sound comparisons of the woman, the pitch level of the whispered sounds was recognised as either above the level of the voiced sounds produced at the lower f_o of 220 Hz or as equal to this level. Inversely, the pitch level of all whispered sounds was recognised as below the level of the voiced sounds produced at higher f_o, with a frequency range of these higher levels of 330–784 Hz.

Comparison of the sounds of the child: For the sound comparisons of the child, the pitch level of the whispered sounds was again recognised as either above the level of the voiced sounds produced at a lower f_o of 262 Hz or equal to this f_o level. Inversely, with two exceptions, the pitch level of the whispered sounds was recognised as below the level of the voiced sounds produced at higher f_o, with a frequency range of these higher levels of 440–659 Hz. The two remaining comparisons concerned whispered and voiced sounds of /e, ø/, with f_o levels of the voiced sounds of approximately 330 Hz, surpassing 262 Hz only by four semitones. For these two comparisons, a weak labelling majority regarding equal pitch levels and a labelling minority regarding higher pitch levels for voiced than for whispered sounds were found.

Important additional aspects to note: For all sounds of all three speakers, the pitch level of the whispered sounds was never consistently recognised below the pitch level of the voiced sounds produced at the low level of comparison (only two related inconsistent labellings occurred for the sounds of /y/ and /e/ of the man), and it was never recognised above the pitch level of the voiced sounds produced at the high level. Furthermore, with respect to comparisons of whispered sounds with voiced sounds produced at the lower f_o levels of a speaker's vocal range, the results for the woman and the child were comparable and differed markedly from those of the man, for whom pitch differences were significantly more pronounced. In contrast, no marked speaker-related differences were found for the comparisons of whispered sounds with voiced sounds produced at higher f_o levels > 400 Hz. Finally, the pitch recognition rate for consistent labelling independent of the presentation order was 69–84% (consistent pitch level difference identification given by a listener for AB and BA order; see Table 1, online version, Columns 9–18, and bottom rows, AB/BA labelling consistency). Thus, numerous inconsistent pitch level differences occurred (either low or equal levels, or equal or high levels, and vice versa; see results marked in green in the table). However, no opposite low–high or high–low identifications were found except for listener L2 in Series 2 and 3. Besides, labelling inconsistency was somewhat scattered among sound comparisons and listeners.

Discussion

In sum, in the present experiment, the pitch level of a voiced sound produced in the lower vocal range of a speaker was recognised as either lower (majority of cases) or equal to the level of the whispered sound of comparison, and the pitch level of a voiced sound produced in the middle and higher vocal range of a speaker was recognised as either equal to or higher than (majority of cases) the level of the whispered sound. With respect to the voiced sounds produced in the lower vocal range of a speaker, the tendency for pitch levels to be identified as lower than the levels of the whispered sounds of comparison was more pronounced for the sounds of the man than of the woman and the child. Considering the fact that the average f_o level of these voiced sounds of the man was nine semitones below the corresponding level of the woman and one octave below the corresponding level of the child, the results thus indicated that the pitch levels of the whispered sounds of the man were recognised as closer to the levels of the whispered sounds of the woman and the child than to the average f_o level of 131 Hz of his voiced sounds. With respect to the voiced sounds produced in the higher vocal range of the speakers, that is, at f_o levels above 400 Hz, the tendency for their pitch level to be identified as higher than the level of the whispered sound of comparison was pronounced for all sounds of all speakers.

In terms of a first rough estimation, we conclude that the pitch levels of the whispered vowel sounds investigated here fell somewhere in the frequency range of 200–400 Hz. Such an estimate is, at least for a substantial part, in line with the findings of the spectral comparison of natural whispered and voiced vowel sounds reported in Chapter M5.2 and the resynthesis of natural whispered vowel sounds reported in Chapter M5.3. (Notably, these results were found to be unrelated to vowel quality despite the very different vowel-related *P*-patterns.)

However, some relativisations have to be made for the indications obtained and estimations made. Above all, the experiment, the results and our evaluation did not account for different whisper subtypes and for a possibly related pitch variation. Further, although we found pronounced recognition patterns for pitch level differences between voiced and whispered sounds, there were also considerable listener-specific differences in pitch level recognition and labelling consistency. Finally, the higher f_o levels of the voiced sounds of a speaker were not uniform but varied markedly because the sounds were selected according to specific spectral characteristics (see Chapter M5.2). Our conclusions should be understood accordingly, and the aforementioned aspects need to be researched further to allow for a more precise and generalised assessment and formulation of the pitch levels of whispered vowel sounds and their specific production parameters.

Chapter appendix

Table 1. Comparison of natural whispered and voiced vowel sounds of single speakers: Recognised pitch level differences. Columns 1–4 = sound production (S/L = series number and sound links; V = intended and recognised vowel quality; P = phonation type, where v = voiced and w = whispered; fo = f_o intended, in Hz). Column 5 = comparison of P_1 of a whispered sound (reference) and a voiced sound; differences are given for P_1 of a voiced sound as being lower or higher than or comparable to P_1 of the whispered reference sound (values transferred from Chapter M5.2, Table 1). Columns 6–8 = recognised pitch level differences (summarised results of the listening test, 10 identifications per sound; results are given for a voiced sound recognised as l = lower than, e = equal to, or h = higher than the whispered counterpart). Extended online table: Columns 9–18 = listener- and presentation-specific details (L(i) = listeners; AB = whispered–voiced presentation order, BA = voiced–whispered presentation order). Colour code (including the extended online table): Dark blue = labelling majority for a lower pitch level for a voiced sound compared with a whispered sound; light blue = lower or equal pitch level for a voiced sound compared with a whispered sound (without labelling majority for a single option or with a labelling majority for equal pitch); dark red = labelling majority for a higher pitch level for a voiced sound compared with a whispered sound; light red = higher or equal pitch level for a voiced sound compared with a whispered sound (with a weak labelling majority for equal pitch); dark green = contradictory results given by a listener for AB and BA sound presentation order (lower and higher pitch levels or vice versa were identified for AB and BA order of presentation); light green = inconsistent results given by a listener for AB and BA sound presentation order (equal and lower or equal and higher pitch levels or vice versa were identified for AB and BA order of presentation). Bottom rows (extended online table): AB/BA labelling consistency, analysis of listener-specific details. The number of sound pairs with consistent, inconsistent or contradicting identifications for AB and BA presentation order is given per listener (L1–L5) and for all listeners (Sum, including values given in %; the total number of sound pairs identified per speaker sample by all five listeners was 75 or 80, respectively).
[M-06-01-T01]

Table 1. Comparison of natural whispered and voiced vowel sounds of single speakers: Recognised pitch level differences. [M-06-01-T01] Extended online table: ↗

Sounds (man)					Recognition		
Production				Spectrum	Pitch		
S/L	V	P	fo (Hz)	P1	l	eq	h
		v	131	(P1 lacking)	10		
1	i	w	–	reference		reference	
↗		–	–	–		–	
		v	131	lower	8	1	1
2	y	w	–	reference		reference	
↗		v	494	higher			10
		v	131	lower	7	2	1
3	e	w	–	reference		reference	
↗		v	494	comparable		1	9
		v	131	lower	7	3	
4	ø	w	–	reference		reference	
↗		v	294	higher			10
		v	131	lower	6	4	
5	ɛ	w	–	reference		reference	
↗		v	392	higher		1	9
		v	131	lower	9	1	
6	a	w	–	reference		reference	
↗		v	392	higher		2	8
		v	131	lower	9	1	
7	o	w	–	reference		reference	
↗		v	392	comparable		1	9
		v	131	lower	8	2	
8	u	w	–	reference		reference	
↗		v	494	higher			10

Sounds (woman)					Recognition		
Production				Spectrum	Pitch		
S/L	V	P	fo (Hz)	P1	l	eq	h
		v	220	lower	9	1	
9	i	w	–	reference		reference	
↗		v	587	higher		2	8
		v	220	lower	8	2	
10	y	w	–	reference		reference	
↗		v	659	higher			10
		v	220	comparable	4	6	
11	e	w	–	reference		reference	
↗		v	523	higher		2	8
		v	220	comparable	5	5	
12	ø	w	–	reference		reference	
↗		v	587	higher		3	7
		v	220	lower	5	5	
13	ɛ	w	–	reference		reference	
↗		v	523	comparable		2	8
		v	220	lower	5	5	
14	a	w	–	reference		reference	
↗		v	523	comparable		2	8
		v	220	comparable	5	5	
15	o	w	–	(reference)		reference	
↗		v	330	(comparable)			10
		v	220	lower	3	7	
16	u	w	–	reference		reference	
↗		v	784	higher			10

Sounds (child)					Recognition		
Production				Spectrum	Pitch		
S/L	V	P	fo (Hz)	P1	l	eq	h
		v	262	lower	7	3	
17	i	w	–	reference		reference	
↗		v	659	higher		2	8
		v	262	comparable	5	5	
18	y	w	–	reference		reference	
↗		v	659	higher		1	9
		v	262	comparable	5	5	
19	e	w	–	reference		reference	
↗		v	330	higher		6	4
		v	262	comparable	2	8	
20	ø	w	–	reference		reference	
↗		v	330	comparable		6	4
		v	262	lower	5	5	
21	ɛ	w	–	reference		reference	
↗		v	523	comparable		1	9
		v	262	lower	6	4	
22	a	w	–	reference		reference	
↗		v	587	comparable		1	9
		v	262	comparable		10	
23	o	w	–	reference		reference	
↗		v	440	comparable		1	9
		v	262	lower	5	5	
24	u	w	–	reference		reference	
↗		v	587	higher			10

M6.2 Pitch Recognition Comparing Synthesised Whispered-Like and Voiced-Like Vowel Sounds Related to Natural Whispered Utterances of Single Speakers

Introduction

In a second experiment, pitch level differences for synthesised whispered- and voiced-like vowel sounds were investigated based on an extract of the sound sample described in Chapter M5.3. (See also this chapter for the use of the term synthesis here.)

Experiment

Sound sample investigated: From the sound sample described in Chapter 5.3, for each of the eight long Standard German vowels and both speakers (man and woman), four synthesised replicas (synthesis based on an estimated *F*-pattern of a natural whispered reference sound) were selected, related to four source characteristics in synthesis: Noise as a whispered-like source and a voiced-like source with f_o of 131–262–393 Hz (sounds of the man) or 165–262–440 Hz (sounds of the woman; note the selection of the lowest f_o level of 165 Hz investigated in Chapter 5.3 in order to reduce the low f_o level frequency difference of the voiced-like sounds of the woman and the man). For details of synthesis, see Chapter M5.3.

Sound comparison (preparation of the listening test): For each speaker and each vowel, the synthesised sound with noise as the source was compared with each of the three sounds with a voiced-like source at a lower, medium and higher f_o level, in AB and BA order. According to this procedure, two speaker-related subsamples of 48 whispered-like–voiced-like or voiced-like–whispered-like sound pairs were created for the listening test, resulting in a total of 96 test items.

Listening test: The pitch level difference between each sound pair was investigated in a listening test according to the procedure described in the previous chapter. Four subtests were conducted, separating the sounds of the speakers and the presentation order.

Results

Table 1 in the chapter appendix shows the pitch recognition results for the two speaker-related subsamples. Again, pitch level differences are given for a voiced-like sound being recognised as lower or higher than or equal to the whispered-like reference sound of comparison; f_o levels are given in terms of intended f_o; pitch recognition results are given according to the labelling majority.

For the sounds of the man, the pitch levels of the whispered-like sounds were recognised as higher than the levels of the voiced-like sounds synthesised at a low f_o of 131 Hz for six sound pairs (see Table 1, Columns 7–9, results marked in dark blue) and as higher or equal for the remaining two pairs (no labelling majority for a single option; see Table 1, Columns 7–9, results marked in light blue). Conversely, for all sound pairs where the voiced-like sounds were synthesised at a high f_o of 392 Hz, the pitch levels of the voiced-like sounds were recognised as higher than those of the whispered-like sounds (see Table 1, Columns 7–9, results marked in red). Finally, for the sound pairs where the voiced-like sounds were synthesised at a middle f_o of 262 Hz, the recognition results were somewhat mixed, and numerous contradicting identifications occurred; however, the results indicated a tendency towards either equal pitch levels or somewhat balanced contradicting identifications for voiced-like and whispered-like sounds (see Table 1, Columns 7–9, results marked in purple).

Similar results were obtained for the sound comparisons of the woman.

As was the case for the previous experiment, the pitch recognition results for the AB versus the BA order were somewhat inconsistent (see Table 1 online, bottom rows, AB/BA labelling consistency): Consistent identifications were found for only 67–70% of the sound pairs. For the remaining pairs, inconsistent identifications occurred in terms of equal and lower pitch levels or equal and higher pitch levels of a voiced-like sound in comparison to a whispered-like sound for AB and BA order, or vice versa. Exceptions were rare cases of contradicting level assessments (5 of 240 sound pairs labelled in an AB and BA order with contradicting assessments, 4 pairs thereof labelled by one listener and 1 pair by a second listener). Given this general tendency, identification consistency proved to be both sound-related and listener-specific.

Discussion

The results of this second experiment comparing synthesised replicas based on estimated *F*-patterns of natural whispered reference sounds with noise-like and voiced-like source characteristics were in line with the results of the previous experiment comparing natural sounds: As a general tendency, the pitch levels for whispered-like sounds were perceived as higher in comparison to voiced-like sounds at $f_o \leq 165$ Hz and lower in comparison to voiced-like sounds at $f_o \geq 393$ Hz. Furthermore, when comparing whispered-like sounds with voiced-like sounds at $f_o = 262$ Hz, no general tendency of marked and consistent pitch level differences was found, and numerous cases of contradicting identifications occurred. (Notably, no such contradictions occurred for the comparisons with voiced-like replicas at $f_o \leq 165$ Hz and ≥ 393 Hz.) Thus, our above estimate that the pitch level of the investigated natural whispered vowel sounds may be assessed as very often lying above c. 200 Hz and below c. 400 Hz was again supported on the bases of the investigated synthesised replicas. However, the same relativisations as mentioned in the previous chapter apply for the indications and conclusions made here.

Chapter appendix

Table 1. Comparison of synthesised whispered-like and voiced-like vowel sounds, synthesis based on *F*-patterns of natural whispered reference utterances of single speakers: Recognised pitch level differences. Columns 1–5 = sounds (SP = speaker and speaker group; S/L = series number and sound links; V = intended and recognised vowel quality of the natural reference sounds; P = phonation type in terms of source characteristics of Klatt synthesis, where w = whispered-like, v = voiced-like; fo = f_o of synthesis, in Hz; indications transferred from Chapter M5.3, Table 3). Column 6 = vowel recognition results of the synthesised sounds (V; transferred from Chapter M5.3, Table 3). Columns 7–9 = recognised pitch level differences (summarised results of the listening test, 10 identifications per sound; results are given for a voiced sound recognised as l = lower than, e = equal to, or h = higher than the whispered counterpart). Extended online table: Columns 10–19 = listener- and presentation-specific details (L(i) = listeners; AB = whispered–voiced presentation order, BA = voiced–whispered presentation order). Colour code (including the extended online table): Blue and red, see Table 1 in the previous chapter; purple = results of the comparison of whispered-like sounds with voiced-like sounds synthesised at an f_o of 262 Hz. Bottom rows (extended online table): AB/BA labelling consistency, analysis of listener-specific details (see also the legend of Table 1 in the previous chapter; the total number of sound pairs identified per speaker sample by all five listeners was 120).
[M-06-02-T01]

Table 1. Comparison of synthesised whispered-like and voiced-like vowel sounds, synthesis based on F-patterns of natural whispered reference utterances of single speakers: Recognised pitch level differences. [M06-02-T01]
Extended online table: ↗

Sounds					Recognition			
SP	S/L	V	P	fo	V	Pitch		
				Hz		l	eq	h
man	1 ↗	i	v	131	i	8	2	
			wh	–	i	reference		
			v	262	i	4	6	
			v	392	i		3	7
	2 ↗	y	v	131	y	5	5	
			wh	–	y	reference		
			v	262	y	1	8	1
			v	392	y		4	6
	3 ↗	e	v	131	(ɛ–e)	5	5	
			wh	–	e	reference		
			v	262	e	2	2	6
			v	392	(e–i)			10
	4 ↗	ø	v	131	(ɛ–ø)	7	3	
			wh	–	ø	reference		
			v	262	ø	2	7	1
			v	392	ø		3	7
	5 ↗	ɛ	v	131	ɛ	6	4	
			wh	–	ɛ	reference		
			v	262	ɛ	1	7	2
			v	392	ɛ		2	8
	6 ↗	a	v	131	a	6	4	
			wh	–	a	reference		
			v	262	a	3	1	6
			v	392	a			10
	7 ↗	o	v	131	o	7	3	
			wh	–	o	reference		
			v	262	o	1	7	2
			v	392	u		1	9
	8 ↗	u	v	131	o	8	2	
			wh	–	u	reference		
			v	262	o	2	6	2
			v	392	u			10
woman	9 ↗	i	v	165	i	8	2	
			wh	–	i	reference		
			v	262	i	5	5	
			v	440	i		2	8
	10 ↗	y	v	165	(ø–y)	8	2	
			wh	–	y	reference		
			v	262	y	4	3	3
			v	440	y			10
	11 ↗	e	v	165	e	5	5	
			wh	–	e	reference		
			v	262	e	6		4
			v	440	i			10
	12 ↗	ø	v	165	ø	7	3	
			wh	–	ø	reference		
			v	262	ø	4	6	
			v	440	y		4	6
	13 ↗	ɛ	v	165	ɛ	6	4	
			wh	–	ɛ	reference		
			v	262	ɛ	4	3	3
			v	440	ɛ		2	8
	14 ↗	a	v	165	a	4	6	
			wh	–	a	reference		
			v	262	a	1	4	5
			v	440	a		1	9
	15 ↗	o	v	165	ɔ	8	2	
			wh	–	o	reference		
			v	262	o	4	6	
			v	440	(o–u)		4	6
	16 ↗	u	v	165	o	7	3	
			wh	–	u	reference		
			v	262	u	5	5	
			v	440	u		4	6

M6.3 Pitch Recognition Comparing Either Natural Whispered or Synthesised Whispered-Like Vowel Sounds Related to Utterances of Speakers Different in Age and Gender

Introduction

In the previous two experiments, the pitch of whispered and whispered-like vowel sounds was investigated by means of a comparison with voiced and voiced-like vowel sounds, all comparisons being related to utterances of single speakers. In two further experiments, using the same two sound samples of the previous chapters, pitch level differences were investigated by means of a comparison of whispered or whispered-like sounds (either natural or synthesised replicas) produced by different speakers of different ages or gender. The initial goal was to investigate whether pitch assessment is subject to age- and gender-related differences.

Experiment 1

Sound sample investigated: All 24 natural whispered sounds of the eight long Standard German vowels produced by the man, the woman and the child investigated in Chapter M6.1 were selected.

Sound comparison (preparation of the listening test): For each vowel, each of the three sounds were compared with one of its opposing sounds, in AB and BA order, resulting in six test items as six sound pairs per vowel and 48 test items in total (see Table 1 in the chapter appendix).

Listening test: Pitch recognition of the sounds was investigated in a listening test according to the procedure described in Chapter M6.1. Each test item contained two whispered-like sounds of the same vowel produced by two different speakers (separated by an approximately 0.5 sec. pause). All sound pairs were tested in one test run. The listeners were asked to identify the pitch level difference between the first and the second sound as falling, flat or rising, referring to dominant or prominent pitches.

Results 1

Table 1 in the chapter appendix shows the pitch recognition results. Pitch level differences are given for the sounds of the man when compared to those of the woman or the child and for the sounds of the woman when compared to those of the child. As a general tendency, the pitch level of the sounds of the man was identified as either lower than or equal to the pitch of the sounds of the women and the child, and the same held true for the comparison of the sounds of the women and the child. However, the results were not uniform among all vowel qualities, and labelling consistency among vowel qualities, order of presentation and listeners was somewhat limited (overall labelling consistency = 74%).

Experiment 2

Sound sample investigated: All 16 synthesised whispered-like replicas of the sounds of the eight long Standard German vowels produced by the man and the woman investigated in Chapter M6.2 were selected.

Sound comparison (preparation of the listening test): For each vowel, the two sounds produced by the two speakers were compared with each other, in AB and BA order, resulting in two test items as two sound pairs per vowel and in 16 test items in total (see Table 2 in the chapter appendix).

Listening test: The pitch recognition of the sounds was investigated in a listening test according to the procedure described in the previous experiment.

Results 2

Table 2 in the chapter appendix shows the pitch recognition results. Pitch level differences are given for the sounds of the man compared with those of the woman. For the synthesised sounds of front vowels and of /a/ related to the natural reference sounds produced by the man, the pitch level was unanimously identified as lower than the level of the sounds of the woman. In contrast, for the synthesised replicas related to the natural reference sounds of /o/ produced by the man, the pitch level was either identified as lower than or equal to the level of the sounds of the woman, with a single occurring labelling inconsistency (see listener L4). For the synthesised replicas of /u/, no inter-speaker differences were identified for the pitch levels.

Discussion

As a tendency, the recognition results for natural whispered sounds indicated lower or equal pitch levels for the sounds of the man compared to the sounds of the woman and the child and lower or equal levels for the sounds of the woman compared to the sounds of the child. Hence, for the investigated sound sample, a somewhat limited and inconsistent tendency towards age- and gender-related pitch level recognition was found. Correspondingly, labelling consistency did not surpass 74%. (Besides, one of the listeners claimed to recognise the age and gender of the speakers when listening to the sounds and felt that discerning pitch differences between men and women was easier than between adults and children. Also, in preparation for the listening test, the author had the impression that the individual sound timbre may have impacted the pitch recognition of whispered vowel sounds.)

The results of the listening test for synthesised sounds of front vowels and /a/ revealed much more pronounced gender-related differences, with a 100% pitch recognition rate for all vowels and with a corresponding labelling consistency: The pitch level of the sounds produced by the man was unanimously identified as below the level of the sounds of the woman. Concerning the two back vowels, the pitch level of the sounds of the man was identified as either lower than or equal to the sounds of the woman.

In these terms, for the sounds investigated, the tendency towards age- and gender-related pitch recognition proved to be dependent on sound types and vowel qualities.

The finding that pitch differences were easier to discern for synthesised whispered-like sounds than for natural whispered sounds (note the pronounced difference in the labelling consistency) was unexpected and needs to be addressed in future research. (One listener indicated explicitly that he perceived the pitch of natural whispered sounds to be different from that of synthesised replicas and that this perceptual difference had an impact on whether or not he could assign a pitch level difference with certainty.) However, the fact that, here, the second experiment did not relate to the same natural reference sounds as examined in the first experiment has to be taken into consideration for the results obtained.

These results represent only preliminary indications based on two small samples. A more detailed investigation would again have to address several additional methodological aspects: The size of the sound sample and the number of speakers, subtypes of whispered phonation, intra- and inter-vowel comparisons, method of spectral peak estimation for the subsequent synthesis, comparison of natural utterances and resynthesised replicas or purely synthesised sounds not related to natural utterances, different types of recognition tasks, number of listeners and listener background. (Note that the training of listeners may prove to be of primary importance for tests of this kind.) Further, an analysis of the results would have to relate the identifications of pitch level differences (or assigned pitch levels) to the recognition consistency of individual listeners.

Here, the experimentation only served as an addition to the preceding investigation in which a first attempt was made to estimate the pitch frequency range of whispered vowel sounds. In this respect, three indications can be derived from the above results, which are important for future research on whispered speech and pitch: (i) The pitch of whispered vowel sounds may to some extent relate to age and gender; (ii) age- and gender-related pitch differences for the whispered and whispered-like sounds in the present study were somewhat limited in their extent when compared with age- and gender-related f_o differences (and related pitch level differences) for voiced sounds as generally given in formant statistics; (iii) pitch recognition of natural whispered vowel sounds and (re-)synthesised replicas may prove not to be directly comparable in general.

Chapter appendix

Table 1. Comparison of natural whispered vowel sounds produced by a man, a woman and a child: Recognised pitch level differences. Columns 1–3 = sounds and sound comparison (S/L = series number and sound links; V = intended and recognised vowel quality of the natural reference sounds; Sound pair = sounds compared, in the order of m, w, c for man, woman and child). Columns 4–6 = pitch recognition results (comparison of pitch levels, summarised results, where l = lower than, eq = equal to, h = higher than; results are given according to Column 3). Columns 7–16 = listener- and presentation-specific pitch recognition results (L(i) = listeners; AB or BA = presentation order, where AB = sounds in man–woman, man–child and woman–child order, BA = vice versa). Colour code: Dark blue = labelling majority for a lower pitch level for a sound comparison; light blue = mostly lower or equal pitch level recognition for a sound comparison (with or without a labelling majority for equal pitch); purple = occurring additional identifications of a higher pitch level for the sounds of the man than for the woman or the child, and also for the sounds of the woman than for the child; dark green = contradicting results given by a listener for AB and BA sound presentation order (lower and higher pitch levels or vice versa were identified for AB and BA order); light green = inconsistent results given by a listener for AB and BA order (equal and lower or equal and higher pitch levels or vice versa were identified for AB and BA order). Bottom rows: AB/BA labelling consistency, analysis of listener-specific details (see the legend of Table 1 in Chapter M6.1; total number of sound pairs identified = 24 per listener and 120 in total).
[M-06-03-T01]

Table 2. Comparison of synthesised whispered-like vowel sounds related to natural whispered reference sounds produced by a man and a woman: Recognised pitch level differences. For columns and colour code, see Table 1.
[M-06-03-T02]

Table 1. Comparison of natural whispered vowel sounds produced by a man, a woman and a child: Recognised pitch level differences. [M-06-03-T01]

Sounds			Recognition			Recognition (details)									
Comparison			Pitch			AB order					BA order				
S/L	V	Sound pairs	l	eq	h	L1	L2	L3	L4	L5	L1	L2	L3	L4	L5
1 ↗	i	m is … than w		9	1	eq	eq	h	eq	eq	eq	eq	eq	eq	eq
		m is … than c	2	8		eq	eq	l	eq	eq	eq	l	eq	eq	eq
		w is … than c	2	6	2	eq	eq	h	eq	eq	eq	l	l	eq	h
2 ↗	y	m is … than w	6	4		eq	l	l	eq	l	eq	l	l	eq	l
		m is … than c	8	2		l	l	l	eq	l	l	l	l	eq	l
		w is … than c	6	4		l	l	l	eq	eq	eq	l	l	eq	l
3 ↗	e	m is … than w	5	5		eq	l	l	eq	eq	eq	l	l	eq	l
		m is … than c	6	4		eq	l	l	eq	eq	l	l	l	eq	l
		w is … than c	5	5		eq	l	l	eq	eq	eq	l	l	eq	l
4 ↗	ø	m is … than w	4	5	1	eq	l	l	eq	eq	eq	l	h	eq	l
		m is … than c	8	2		l	l	l	eq	l	l	l	l	eq	l
		w is … than c	7	3		eq	l	l	eq	l	l	l	l	eq	l
5 ↗	ɛ	m is … than w	8	2		l	l	l	eq	l	l	l	l	eq	l
		m is … than c	9	1		l	l	l	eq	l	l	l	l	l	l
		w is … than c	3	4	3	eq	h	l	eq	eq	l	l	h	eq	h
6 ↗	a	m is … than w		8	2	eq	eq	h	eq	eq	eq	eq	h	eq	eq
		m is … than c	3	7		eq	l	l	eq	eq	eq	l	eq	eq	eq
		w is … than c	7	2	1	l	l	l	eq	l	l	l	h	eq	l
7 ↗	o	m is … than w	6	3	1	l	l	h	eq	eq	l	l	l	eq	l
		m is … than c	3	6	1	eq	l	eq	eq	eq	l	l	h	eq	eq
		w is … than c	2	7	1	eq	l	eq	eq	eq	eq	l	h	eq	eq
8 ↗	u	m is … than w	8	2		eq	l	l	eq	l	l	l	l	l	l
		m is … than c	2	8		eq	l	eq	eq	eq	eq	l	eq	eq	eq
		w is … than c	1	5	4	h	h	h	eq	eq	eq	l	h	eq	eq
Total		m is … than w	37	38	5										
		m is … than c	41	38	1										
		w is … than c	33	36	11										

	Identifications	L1	L2	L3	L4	L5	Sum
AB/BA labelling consistency (24 sound pairs investigated)	consistent	17	20	14	22	16	= 89 of 120 (74%)
	inconsistent	7	2	5	2	8	= 24 of 120 (20%)
	opposite	0	2	5	0	0	= 7 of 120 (6%)

Table 2. Comparison of synthesised whispered-like vowel sounds related to natural whispered reference sounds produced by a man and a woman: Recognised pitch level differences. [M-06-03-T02]

Sounds			Recognition			Recognition (details)									
Comparison			Pitch			AB order					BA order				
S/L	V	Sound pair	l	eq	h	L1	L2	L3	L4	L5	L1	L2	L3	L4	L5
1 ↗	i	m is … than w	10			l	l	l	l	l	l	l	l	l	l
2 ↗	y	m is … than w	10			l	l	l	l	l	l	l	l	l	l
3 ↗	e	m is … than w	10			l	l	l	l	l	l	l	l	l	l
4 ↗	ø	m is … than w	10			l	l	l	l	l	l	l	l	l	l
5 ↗	ε	m is … than w	10			l	l	l	l	l	l	l	l	l	l
6 ↗	a	m is … than w	10			l	l	l	l	l	l	l	l	l	l
7 ↗	o	m is … than w	5	5		eq	eq	l	eq	l	eq	eq	l	l	l
8 ↗	u	m is … than w		10		eq	eq	eq	eq	eq	eq	eq	eq	eq	eq
Total		**m is … than w**	65	15											
AB/BA labelling consistency (8 sound pairs investigated)			**Identifications**			**L1**	**L2**	**L3**	**L4**	**L5**	**Sum**				
			consistent			8	8	8	7	8	= 39 of 40 (97.5%)				
			inconsistent			0	0	0	1	0	= 1 of 40 (2.5%)				
			opposite			0	0	0	0	0					

M6.4 Natural Vowel Sounds With a Suppressed Fundamental

Introduction

The pitch of a periodic sound can be perceived independently of whether or not the first harmonic, *H*1, commonly termed the fundamental, is present in the spectrum (“missing fundamental” phenomenon; see the excursus on fundamental frequency and pitch in Part II; for an overview, see also Houtsma, 1995; Yost, 2009). For example, a sound with a series of harmonics as integer multiples of 200 Hz remains at a 200 Hz pitch level even if the first harmonic is (or also some of the lower harmonics are) removed. Similarly, speech remains intelligible even if frequencies below c. 300 Hz are filtered, including the original pitch contour (consider e.g. fixed telephone line transmission with a band-pass filter of c. 300–3400 Hz).

In an early study investigating natural sounds of back, central and front vowels produced by a man as sustained isolated sounds at an f_o of c. 140 Hz, Lehiste and Peterson (1959) found that HP filtering of the sounds with CF set to 550 Hz did not cause a substantial decrease of accurate vowel recognition. Therefore, they concluded that “[…] the fundamental can be eliminated without disturbing the vowel recognition significantly”. In a more recent study, Fahey and Diehl (1996) investigated the effect of f_o variation on vowel quality recognition in vowel synthesis for unfiltered and HP-filtered sounds, with HP filtering deleting *H*1 or *H*1–*H*2. Vowel sounds were synthesised (Klatt synthesis) based on seven *F*-patterns for the /ɪ/–/ɛ/ range, with endpoint formant frequencies equal to mean adult male values of the first three formants reported by Peterson and Barney (1952) for the corresponding corner vowel categories. Five f_o levels in the range of 100–200 Hz were applied. Testing vowel recognition for the unfiltered and the HP-filtered sounds showed that increasing the f_o level resulted in a vowel boundary shift in an open–close direction for all conditions of synthesis, independent of whether *H*1 or *H*1–*H*2 were deleted. Thus, as Fahey and Diehl (1996) concluded, vowel quality-specific spectral characteristics of voiced sounds cannot be related to *H*1 in a systematic manner, rejecting the claim of Traunmüller (1981) that the tonotopic distance F_1–f_o in terms of F_1–H_1 is the cue for vowel height. (However, note in this context that only synthesised sounds of front vowels and an f_o variation limited to the frequency range of 100–200 Hz were investigated in this study.)

In the previous chapters, vowel sounds having a pitch but lacking measurable f_o were demonstrated. Vowel sounds with a “missing fundamental” represent a second phenomenon of sounds for which f_o-related acoustic characteristics, vowel quality and pitch level are at issue. Therefore, a corresponding experiment was conducted.

Experiment

Selection of speakers and sounds: Based on the Zurich Corpus, for each of the eight long Standard German vowels and each of the speaker groups of men, women and children, three sounds produced by three speakers in nonstyle mode, V context and with a medium vocal effort at intended f_o of 131 Hz (men), 220 Hz (women) and 262 Hz (children) were selected. According to the results of the standard listening test conducted when creating the corpus, all sounds were fully recognised (100% vowel recognition rate matching vowel intention). As a result, a sample of 72 natural reference sounds produced by different speakers was investigated.

f_o measurement: Acoustic analysis accorded to the standard procedure of the Zurich Corpus.

Suppression of *H*1 and *H*1–*H*2: Based on the 72 unfiltered natural reference sounds, a second sample of 72 filtered sounds with suppressed *H*1 (HP filtering the reference sounds with CF = 2×calculated f_o) and a third sample of 72 filtered sounds with suppressed *H*1–*H*2 (HP filtering the reference sounds with CF = 3×calculated f_o) was created. HP filtering was conducted using the Hann filter in Praat (default parameters).

Listening tests: For each of the two samples with suppressed *H*1 or *H*1–*H*2 separately (two subtests), vowel recognition was tested according to the standard procedure of the Zurich Corpus and involving the five standard listeners.

Results

Table 1 in the chapter appendix shows the sound samples and the vowel recognition results, including sound links. According to the labelling majority of the listening test, with one exception, all sounds with suppressed *H*1 were recognised as either matching vowel intention of the speakers or an adjacent vowel quality or as a vowel boundary or area of intended and adjacent qualities. The same held true for 57 of the 72 sounds with suppressed *H*1–*H*2. For the remaining sounds, vowel confusion involving more than one adjacent vowel occurred.

With two exceptions for sounds produced by men, according to the labelling majority, vowel quality shifts or vowel confusions triggered by suppressed *H*1 only occurred for sounds of close vowels produced by women and children. Vowel quality shifts or vowel confusions triggered by suppressed *H*1–*H*2 occurred for sounds of close vowels produced by men and sounds of close and close-mid vowels produced by women and children. The (initial) shift direction from the vowel quality intended by the speaker to an adjacent or non-adjacent vowel quality was found to generally be close–open, and only two subsequent reverted shifts occurred in an open–close direction back to the vowel quality of the natural reference sound (see Series 10 and 18, results marked in purple).

Discussion

In the present experiment, when suppressing *H*1 or *H*1–*H*2, vowel recognition did not, in general, relate to *H*1. If vowel quality was affected by the suppression, either (initial) close–open shifts or vowel confusions occurred relating to the vowel openness of the sounds since only reference sounds of close and close-mid vowels were affected. Further, the occurring differences between the sounds of the men and the sounds of the women and the children may have resulted from different f_o levels of sound production, different spectral energy distribution < 1 kHz and different CFs of HP filtering. Further, the occurring differences for sounds of the same vowel produced by speakers of the same speaker group indicated an additional effect of individual sound production. Also, the lack of within-speaker variation of f_o and vocal effort must be considered when evaluating the findings. We will return to some of these aspects in Chapter M8.2.

Pitch recognition was not investigated in this experiment. However, it is assumed here that no pitch variation resulted from the HP filtering applied. (Note in this context that, with few exceptions, measured f_o for the sounds with suppressed *H*1 or *H*1–*H*2 corresponded to measured f_o of the unfiltered reference sounds. Note also that pitch equivalence can be examined by listening to the sounds investigated; see the sound links in Table 1.) Accordingly, if vowel quality shifts occurred, they concerned either (initial) shifts in a close–open direction or vowel confusions, in contrast to the shifts found for increasing f_o in sound synthesis, keeping the spectral envelope unchanged, which resulted in shifts in an open–close direction. Thus, if suppression of *H*1 or *H*1–*H*2 affected vowel quality recognition, these shifts are assumed here as unrelated to pitch. This interpretation is in line with the conclusion of Fahey and Diehl (1996) that vowel recognition does not rely on *H*1, and sounds with suppressed *H*1 or *H*1–*H*2 represent cases for which fundamental frequency and HCF cannot be equated with $H1$. However, here, the indication that the suppression of $H1$ or $H1$–$H2$ did not affect pitch only concerned sounds produced by speakers in their lower vocal range, and the question of whether or not this indication can be confirmed for all recognisable sounds independent of their f_o level of production is left open.

Based on the background exposed in the introduction of this chapter, the above results were to be expected. However, the main aim of the present experiment was to create a documentation of sounds and sound spectra needed for a general discussion of the role of f_o, $H1$, HCF, periodicity and pitch for vowel recognition.

The occurring shifts in the present experiment may be understood as comparable to the shifts found in the resynthesis of F-patterns at f_o surpassing statistical average F_1 of close and close-mid vowels (see Chapter M3.1), which we have interpreted as a consequence of an alteration of the relation between spectral energy and energy maxima below and above 1 kHz for sounds of front vowels, and this may also hold true for sounds of back vowels in terms of an alteration of the relation between spectral energy and energy maxima below and above c. 0.3–0.5 kHz (depending on f_o). In Chapter M8.2, in a more extensive HP filtering experiment, we will document and discuss further sound examples associated with vowel quality shifts of this type.

Finally, there were a few cases of erroneous f_o measurements for sounds with suppressed $H1$ or $H1$–$H2$. Examples are documented in Figure 1 in the chapter appendix. These cases are important to consider because they exemplify the possibility of manipulated natural voiced sounds for which a direct parallelism of calculated f_o and pitch is lacking.

Chapter appendix

Table 1. Vowel sounds produced by men, women and children, with suppressed *H*1 and *H*1–*H*2: Vowel recognition results. Columns 1 and 2 = sounds (S/L = series number and sound links; V = intended and recognised vowel quality of the unfiltered natural reference sounds). Columns 3–6 = recognition results for the sounds with suppressed *H*1 and *H*1–*H*2 (V = vowel quality recognised; Maj = labelling majority of the five listeners; vowel boundaries are given as two characters without a space in between; results given in parenthesis indicate vowel recognition involving two or more vowel qualities). Extended online table: Columns 7–16 = listener-specific details of the vowel recognition (L(i) = listeners). Colour code: Dark red = close–open vowel quality shifts (comparison of intended and recognised vowel qualities); light red = close–open vowel boundary shifts; purple = reverted open–close vowel quality shifts subsequent to an initial close–open shift; grey = other vowel confusions.
[M-06-04-T01]

Figure 1. Vowel sounds produced by men, women and children, with suppressed *H*1 and *H*1–*H*2: Examples of occurring erroneous f_o measurements. Sounds 1 and 2 = two sounds of /y/ and /e/ produced by a child at an intended f_o of 262 Hz with suppressed *H*1–*H*2; measured f_o = 187 Hz and 126 Hz. Sound 3 = a sound of /ø/ produced by a woman with suppressed *H*1–*H*2; measured f_o = 109 Hz. Sounds 4 and 5 = a sound of /u/ produced by a child with suppressed *H*1 and *H*1–*H*2; measured f_o = 492 Hz and 994 Hz. For all sounds, the calculated f_o level differed markedly from the intended f_o and pitch level. (See also the sounds online, Layout L.)
[M-06-04-F01] ↗

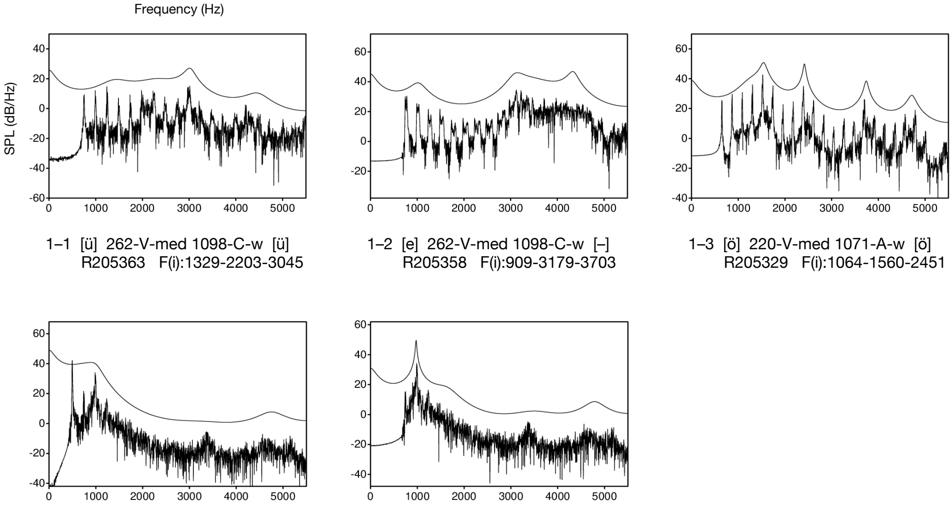
Table 1. Vowel sounds produced by men, women and children, with suppressed H1 and H1–H2: Vowel recognition results. [M-06-04-T01] Extended online table: ↗

Sounds		Recognition			
Men		H1		H1/H2	
S/L	V	V	Maj	V	Maj
1 ↗	i	i	4/5	(ε–ei)	–
		i	4/5	e	3/5
		i	5/5	e	4/5
2 ↗	y	(ø–ey–y)	–	e	3/5
		y	5/5	(ø–e–y)	–
		y	4/5	e	3/5
3 ↗	u	u	5/5	o	3/5
		u	3/5	(u–o)	–
		(u–o)	–	o	4/5
4 ↗	e	e	5/5	e	5/5
		e	5/5	e	5/5
		e	5/5	e	5/5
5 ↗	ø	ø	5/5	ø	4/5
		ø	5/5	ø	5/5
		ø	5/5	ø	4/5
6 ↗	o	o	4/5	o	4/5
		o	4/5	o	5/5
		o	4/5	o	4/5
7 ↗	ε	ε	5/5	ε	5/5
		ε	5/5	ε	4/5
		ε	5/5	ε	5/5
8 ↗	a	a	5/5	a	5/5
		a	5/5	a	5/5
		a	5/5	a	5/5

Sounds		Recognition			
Women		H1		H1/H2	
S/L	V	V	Maj	V	Maj
9 ↗	i	(e–i)	–	(ε–i)	–
		(e–i)	–	i	3/5
		e	3/5	(ε–i)	–
10 ↗	y	y	4/5	y	3/5
		ø	3/5	y	3/5
		y	3/5	y	3/5
11 ↗	u	u	4/5	u	4/5
		u	3/5	u	3/5
		o	3/5	ɔ	3/5
12 ↗	e	e	5/5	ε	3/5
		e	4/5	(εe–i)	–
		e	5/5	(ε–e)	–
13 ↗	ø	ø	5/5	(ε–y)	–
		ø	5/5	ø	3/5
		ø	4/5	ø	3/5
14 ↗	o	o	5/5	(uo–ɔ)	–
		o	5/5	ɔ	5/5
		o	4/5	ɔ	4/5
15 ↗	ε	ε	5/5	ε	5/5
		ε	5/5	ε	5/5
		ε	5/5	ε	5/5
16 ↗	a	a	5/5	a	5/5
		a	5/5	a	5/5
		a	5/5	a	5/5

Sounds		Recognition			
Children		H1		H1/H2	
S/L	V	V	Maj	V	Maj
17 ↗	i	e	3/5	(ε–i)	–
		i	3/5	(ε–i)	–
		e	3/5	(εe–i)	–
18 ↗	y	ø	3/5	ø	3/5
		ø	4/5	(ə–y)	–
		(ø–y)	–	y	3/5
19 ↗	u	u	4/5	u	4/5
		o	3/5	(ɔ–u)	–
		u	3/5	u	3/5
20 ↗	e	e	5/5	εe	3/5
		e	5/5	(ε–i)	–
		e	5/5	(ε–ei)	–
21 ↗	ø	ø	5/5	ø	3/5
		ø	5/5	(εø–ə–y)	–
		ø	5/5	ə	3/5
22 ↗	o	o	5/5	ɔ	4/5
		o	5/5	ɔ	5/5
		o	5/5	ɔ	3/5
23 ↗	ε	ε	5/5	ε	5/5
		ε	5/5	ε	5/5
		ε	5/5	ε	5/5
24 ↗	a	a	5/5	a	4/5
		a	5/5	a	5/5
		a	5/5	a	5/5

Figure 1. Vowel sounds produced by men, women and children, with suppressed *H*1 and *H*1–*H*2: Examples of occurring erroneous f_o measurements.
[M-06-04-F01]



1–1 [ü] 262-V-med 1098-C-w [ü]
R205363 F(i):1329-2203-3045

1–2 [e] 262-V-med 1098-C-w [–]
R205358 F(i):909-3179-3703

1–3 [ö] 220-V-med 1071-A-w [ö]
R205329 F(i):1064-1560-2451

1–4 [u] 262-V-med 1098-C-w [u]
R205290 F(i):619-1066

1–5 [u] 262-V-med 1098-C-w [u]
R205362 F(i):981-1578

M6.5 Sinewave Vowel Sounds I – Replicas Related to Statistical Formant Patterns

Introduction

There is a comprehensive amount of literature on the intelligibility of sinewave speech, that is, synthesised replicas of utterances based on time-varying sinusoidal patterns following the changing formant centre frequencies of the natural sounds (Remez et al., 1981). Of particular interest to the present context are single sinewave vowel sounds that are synthesised based on a small number of sinusoids related to statistical average *F*-patterns of natural sounds produced in V context or the context of minimal pairs: Sounds of this type can be synthesised with sinusoid frequency configurations that are directly related to these *F*-patterns but lacking both a fundamental *H*1 as well as HCF comparable to a harmonic spectrum of a natural vowel sound. Thus, the question of the relation between f_o measure, pitch level and vowel recognition is posed from a broader perspective.

In the literature, the recognition of sinewave vowel sounds replicating *F*-patterns of natural sounds produced in citation-form words (hVd syllables produced by men, women and children, with replicated *F*-patterns comparable to statistical average *F*-patterns) is reported as impaired when compared with the recognition of natural vowel sounds (Hillenbrand et al., 2011; see also their overview on the subject matter and earlier studies including sinewave speech). However, and most importantly, vowel confusion for synthesised sounds is also indicated to relate to vowel openness: In the study of Hillenbrand et al., for vowels comparable to Standard German (excluding sounds of ɚ) and prior to listener training, the highest recognition rate was found for the close vowels /i, u/ (76.7% and 75.9%, respectively) and the lowest recognition rate was found for the open-mid and open vowels /ɛ, ɑ/ (30.4% and 33.1%, respectively; see also Morton and Carpenter, 1962, for an early indication of a possible correlation of vowel openness and recognition rate for two-sinusoid vowel sounds, the recognition rate decreasing from /u/ to /a/). (As the results of Hillenbrand et al. show, the vowel recognition of these types of synthesised sounds can be significantly improved by listener training. However, this training effect is not investigated here.)

When considering the very pronounced recognition difference for sounds of different vowels in the Hillenbrand et al. study, attention should be given to the fact that S_1, the frequency level of the first sinusoid representing F_1 of the natural sounds, was substantially lower for *S*-patterns related to sounds of close vowels than for *S*-patterns related to the sounds of all other vowel qualities, with the highest S_1 for open-mid and open vowels. Studies of pitch recognition in sinusoidal sentences indicated that pitch approximately relates to the frequency of the first sinusoid (Remez and Rubin, 1984, 1993). However, this relation may not be independent of the frequency distances of the sinusoids and of their harmonicity. But for cases in which S_1 and recognised pitch level indeed relate to each other – assuming that vowel recognition in its turn relates to pitch –, vowel quality shifts or confusions are expected to occur with increasing differences between the pitch levels of the natural reference sounds and the pitch levels of the related sinewave replicas, at least for some of the vowel qualities, as is indicated by the previous findings presented in this treatise.

When we started designing a sinewave vowel experiment to further investigate f_o, *H*1, HCF, periodicity and pitch for vowel recognition, a major aspect of the formant pattern and spectral shape ambiguity phenomenon was taken into account: Given single spectral envelopes of sounds of close-mid vowels produced at f_o of 200–250 Hz, a one-octave increase in f_o in synthesis, keeping the spectral envelope unchanged, was shown to result in pronounced close-mid–close vowel recognition shifts (see Chapter M3), in contrast to sounds of these vowels produced at f_o of 100–125 Hz and a one-octave increase in f_o in synthesis. We, therefore, assumed that statistical average *F*-patterns of vowel sounds produced by women might represent a promising outset for the exploration of the matter: Statistical average *F*-patterns of women are generally reported for sounds produced at f_o of 200–250 Hz, and a one-octave f_o variation appertains to the everyday speech of women.

In view of the foregoing, vowel and pitch recognition of sinusoid S_1–S_2–S_3 sounds replicating statistical F_1–F_2–F_3 patterns of the eight Standard German vowels for women were investigated according to the following main idea and assumption: If vowel and pitch recognition interrelate, and if pitch recognition of sounds in three-sinusoid synthesis relates to S_1, then lower pitch levels and less vowel confusion can be expected for sounds replicating the *F*-patterns of close vowels, above all when compared with the levels of sounds replicating the *F*-patterns of close-mid vowels. However, because of the nonuniform relation of the vowel spectrum to f_o, no assumption was made for sounds of the open-mid and open vowels. (For a corresponding earlier experiment, see Maurer, Suter et al., 2018. The follow-up experiment presented here was conducted with a larger number of vowel qualities, with adjusted synthesis parameters and a new pitch recognition test design.)

Experiment

Selection of vowels and statistical average *F*-patterns: Statistical average F_1–F_2–F_3 patterns for sounds of the eight Standard German vowels /i, y, e, ø, ɛ, a, o, u/ of women as reported by Pätzold and Simpson (1997; see Chapter M3.1) were selected for investigation.

Sinewave synthesis: Static three-sinewave S_1–S_2–S_3 replicas of the F_1–F_2–F_3 patterns of the eight vowels were synthesised using the SinSyn tool (see Chapter 1.2). Sinewave levels L_{S1}–L_{S2}–L_{S3} were set to 100–80–80 dB for sounds of front vowels and to 100–90–70 dB for sounds of back vowels and /a/. (Sinewave level configurations were set in a generalised manner according to the author's estimate with regard to comparable occurring *F*-patterns of natural sounds produced by women in the f_o range of 200–250 Hz, as documented in the Zurich Corpus.) Sound duration was 1 sec. including a 0.05 sec. fade in/fade out. As a result, a sample of eight sinewave vowel sounds was created.

f_o measurement: Acoustic analysis of the synthesised sounds accorded to the standard procedure of the Zurich Corpus.

Listening tests: Vowel recognition of the synthesised sounds was investigated in a listening test according to the standard procedure of the Zurich Corpus and involving the five standard listeners. Two test specifications were adopted: The listeners were asked to label one of the eight long Standard German vowels or /ɔ/ or /ə/ (forced choice, excluding vowel boundaries) referring to dominant or prominent qualities, and each sound was presented twice (resulting in 10 identifications per sound).

Separately, pitch recognition of the synthesised sounds was investigated in two subtests involving the same listeners. In the first subtest, the sounds related to the *F*-patterns of close-mid and close and of close-mid, open-mid and open vowels were compared as follows: A close-mid versus a close vowel sound, a close-mid versus an open-mid vowel sound and a close-mid versus an open vowel sound, in AB and BA order. Thus, each test item consisted of two vowel sounds (separated by a 0.5 sec. pause). The listeners were asked to identify whether the pitch level of the second sound when compared with the pitch level of the first sound was falling, flat or rising, referring to dominant or prominent levels (see below). In the second subtest, pitch recognition was investigated by comparing all sinewave vowel sounds with two single sinusoids of 349 Hz and 440 Hz separately: One test item consisted of either the 349 Hz or the 440 Hz sinusoid followed by a sinewave vowel sound (separated by a 0.5 sec. pause). Listeners were again asked to identify the pitch level difference.

Before performing the recognition test, according to the standard procedure of the Zurich Corpus, the listeners were asked to listen to the sounds to become familiar with the sound timbre and the listening task.

Results

Table 1 in the chapter appendix shows the sound sample and the vowel and pitch recognition results, including sound links. In the table, measured f_o values are also given.

With respect to vowel recognition, according to the labelling majority, all three sounds with *S*-patterns related to *F*-patterns of close vowels were recognised according to vowel intention, with a recognition rate of ≥ 80%. On the contrary, all three sounds related to *F*-patterns of close-mid vowels were confused and were recognised as close vowels, again with a recognition rate of ≥ 80%. The sound related to the *F*-pattern of /ɛ/ was confused with /ø/, and the sound related to the *F*-pattern of /a/ was mostly identified as /a/ or /ɔ/.

With respect to pitch recognition, uniform results were obtained across listeners, tests and order of sound presentation: The pitch level of all sounds related to the *F*-patterns of close vowels was recognised as being lower than the pitch level of all sounds related to the *F*-patterns of close-mid vowels (first pitch recognition subtest), and it was identified as lower than or equal to the 349 Hz sinusoid (second pitch recognition subtest). The pitch level of all sounds related to the *F*-patterns of the close-mid vowels was recognised as being lower than the pitch level of all sounds related to the *F*-patterns of the open-mid and open vowels, and it was identified as being above the 349 Hz sinusoid and lower than or equal to the 440 Hz sinusoid. Finally, the pitch level of all sounds related to the *F*-patterns of the open-mid and the open vowels was recognised as being higher than the 440 Hz sinusoid.

Discussion

An experiment based on synthesised sounds with *S*-patterns corresponding to statistical average *F*-patterns of natural sounds allows for an investigation of the following three questions: Do *F*-patterns – represented by *S*-patterns and corresponding spectral peak patterns – *per se* represent the vowel qualities of natural reference vowel sounds, with no further spectral fine structure? Are sounds produced with this type of *S*-pattern perceived as having a pitch, although HCF comparable to a harmonic sound spectrum is lacking? If *F*-patterns represented by *S*-patterns do not per se represent the vowel qualities of natural reference vowel sounds, does the recognised vowel quality interact with the recognised pitch level?

According to the results of the experiment, the vowel qualities of the sounds related to *F*-patterns of close vowels were matched successfully to the intended quality (majority of labelling), with the pitch of the sounds being recognised at a lower level than the level of the sounds of the other vowels. In contrast, the vowel qualities of the sounds related to *F*-patterns of close-mid and open-mid vowels were confused, and the pitch of these sounds was recognised at middle or higher levels. Thus, spectral peak frequencies did not *per se* represent vowel qualities. They were related to pitch in the perceptual process of vowel recognition. Besides, the fact that the vowel quality of sounds of /a/ was recognised in the /a-ɔ/ range accorded with the observation of a weak or absent relation of the *F*-patterns or spectral envelopes to f_o for sounds of that vowel, as indicated in the earlier experiments presented in this treatise. (In this context, note the study of Rosen and Hui, 2015, comparing sinewave synthesis with noise-vocoded synthesis and reporting a possible effect of pitch for the recognition of sinewave vowels.) Thus, in sum and in response to the above questions, evidence is provided that (i) *F*-patterns – represented by *S*-patterns and corresponding spectral peak patterns – do not *in general* represent vowel qualities of natural vowel sounds, (ii) synthesised sounds with *S*-patterns corresponding to statistical average *F*-patterns of natural sounds can be perceived as having a pitch, and (iii) for sounds of this type, vowel quality recognition does indeed interact with pitch.

Remarkably enough, the confusions above all for the sounds of close-mid vowels found in the present experiment paralleled vowel quality shifts for natural sounds of these vowels as demonstrated in the chapters on formant pattern and spectral shape ambiguity for an increase of f_o (and pitch) of one octave from c. 200–250 Hz to c. 400–500 Hz: Accordingly, in the present experiment, the sounds with *S*-patterns related to the *F*-patterns of /e/, /ø/ and /o/ were mostly recognised as /i/, /y/ and /u/, respectively.

The vowel recognition results of the present experiment were in line with the results of the preceding study by Maurer, Suter et al. (2018). However, some differences were found concerning pitch recognition: In the preceding study, the listeners were asked to label pitch levels using a virtual electronic piano. The results indicated a tendency towards lower pitch levels for sounds related to *F*-patterns of close vowels than for sounds of close-mid vowels, but exceptions occurred, possibly due to octave mismatches. In the present experiment, sounds related to

F-patterns of close, close-mid and open-mid/open vowels were directly compared with each other, and only the pitch difference was labelled, resulting in a uniform pitch recognition among vowel openness and listeners. Thus, pitch recognition results depended on the design of the listening test, and it seems that a direct comparison of vowel sound replicas and their pitch level differences provides more robust results than free ad-hoc pitch frequency assignment of the replicas according to musical notes.

To test whether the lack of harmonicity has a general effect on experiments of this type, we also investigated “harmonically corrected” S_1–S_2–S_3 configurations related to the *F*-patterns of Pätzold and Simpson (1997). The corresponding results showed no substantial effect of harmonicity for sounds of this type (see Maurer, Suter et al., 2018, experiment 2).

However, some limitations also have to be considered when interpreting the results. Above all, in the Klatt resynthesis of the *F*-patterns of Pätzold and Simpson, the sounds at f_o of 220 Hz related to *F*-patterns of close vowels were confused, and they were only recognised at f_o of 330 Hz (see Chapter M3.1). This finding lowers the reliability of the relation between the *F*-patterns and the vowel qualities investigated. Also, the frequency configuration of *S*-patterns (the interrelation of the frequencies) investigated may affect vowel and pitch recognition, not allowing for a simple generalisation of the recognition results related to any given set of *S*-patterns. Both of these aspects are addressed in the next two chapters.

Finally, note that for the three synthesised sounds investigated, measured f_o was not found to be related to S_1 or to pitch levels (see Table 1, Column 7, vowels /i, o, a/).

Chapter appendix

Table 1. Synthesised three-sinusoid sounds related to statistical *F*-patterns of sounds of the long Standard German vowels produced by women: *S*-patterns investigated, f_o measured and vowel and pitch recognition results. Columns 1–7 = sounds (S/L = sound series and sound links; VO = vowel openness; V = vowel qualities of the statistical reference *F*-patterns; *S*(i) = frequencies of the three sinusoids relating to the statistical *F*-patterns, in Hz; fo = calculated f_o, in Hz). Columns 8–16 = vowel recognition results (confusion matrix, summary of labelling). Columns 17–19 = results of the first pitch recognition subtest (summary of labelling; 1<2 = the pitch level of the sounds of the close vowels recognised as lower than the pitch level of the sounds of close-mid vowels, 3>2 = the pitch level of the sounds of the open-mid and open vowels recognised as higher than the pitch level of the sounds of close-mid vowels). Columns 20–22 = results of the second pitch recognition subtest. Columns 23–27 = listener-specific details of vowel recognition (L(i) = listeners). Colour code: Blue = majority of vowel recognition corresponded to vowel intention (vowel quality related to the reference *F*-patterns) associated with a lower recognised pitch level of the sounds; red = majority of vowel recognition indicated a vowel quality shift in an open–close direction when compared to vowel intention (note that the /a/–/ɔ/ difference was ignored) associated with middle or higher recognised pitch levels of the sounds.
[M-06-05-T01]

Table 1. Synthesised three-sinusoid sounds related to statistical F-patterns of sounds of the long Standard German vowels produced by women: S-patterns investigated, fo measured and vowel and pitch recognition results. [M-06-05-T01]

Sounds							Recognition (summary)															Recognition (details)				
Reference			Synthesis			fo	Vowel									Pitch (test 1)			Pitch (test 2)			Vowel				
S/L	VO	V	S-pattern				close			close-mid			open-mid open			low	middle	high	low	middle	high	L1	L2	L3	L4	L5
			S1	S2	S3											1<2	2>1 2<3	3>2	≤349Hz	>349Hz ≤440Hz	>440Hz					
			Hz	Hz	Hz	Hz	i	y	u	e	ø	o	ɜ	a	ɔ											
close																										
		i	329	2316	2796	165	8	2								10			10			i i	i i	y y	i i	i i
1	↗	y	342	1667	2585	338		8			2					10			10			y y	y y	ø ø	y y	y y
		u	350	1048	2760	350			10							10			10			u u	u u	u u	u u	u u
close-mid																										
		e	431	2241	2871	430	8	1			1						10			10		i i	i i	ø y	i i	i i
2	↗	ø	434	1646	2573	431		8			2						10			10		y y	y y	ø y	y y	ø y
		o	438	953	2835	73			8			2					10			10		u u	u u	o o	u u	u u
open-mid open																										
3	↗	ɜ	592	1944	2867	591		1		1	8							10			10	ø ø	ø ø	ø ø	ø ø	e y
		a	779	1347	2785	194						1		4	5			10			10	ɔ ɔ	ɔ ɔ	o ɔ	a a	a a

M6.6 Sinewave Vowel Sounds II – Replicas Related to Estimated Formant Patterns of Single Natural Vowel Sounds

Introduction

Having obtained the results of the first sinewave vowel sound experiment conducted, the experiment was replicated based on a sample of single natural sounds produced by women at an intended f_o of 220 Hz and the respective estimated *F*-patterns of these sounds. The aim of this second sinewave experiment was to create a basis for and documentation of vowel quality recognition in sinewave synthesis that allows for a direct relation between single natural reference sounds and synthesised replicas. This direct relation strengthens the reliability of the relation between *F*-patterns, *S*-patterns and vowel and pitch recognition for the sounds investigated.

Experiment

Selection of speakers and sounds: Based on sounds of the Zurich Corpus produced by women in nonstyle mode with medium vocal effort in V context, for each of the eight long Standard German vowels, a sound produced at intended f_o of 220 Hz was selected by the author. The vowel recognition rate of all sounds was 100% (matching vowel intention) according to the standard listening test conducted when creating the corpus. With two exceptions, the methodological substantiation of *F*-pattern estimation for these sounds was non-critical (see below), this aspect being the main criteria for the selection of the eight sounds. (In order to comply with this methodological condition, sounds of different speakers were chosen.)

***F*-pattern estimation:** F_1–F_2–F_3 estimation for the sounds accorded with the standard analysis of the Zurich Corpus: Automatically calculated values with a speaker-group default setting of LPC analysis were taken as formant frequency values, with two exceptions: For the sound of /u/, a third spectral maximum below 3 kHz was lacking and F_3/S_3 was added manually (F_3/S_3 = 2750 Hz, in an approximative reference to statistical F_3 for /u/ given by Pätzold and Simpson, 1997); for the sound of /ɛ/, in the crosscheck of calculated F_2–F_3 based on the spectrum and the spectrogram, the calculated values had to be corrected manually (note that, according to the author's estimate, resynthesis of the corrected *F*-pattern with the Klatt synthesiser confirms the vowel quality /ɛ/; for verification, use the KlattSyn link in the online corpus). For the resulting values, see Table 1 in the chapter appendix. In the table, the added or corrected values are given in parentheses.

Sinewave synthesis: Static three-sinewave S_1–S_2–S_3 replicas of the F_1–F_2–F_3 patterns of the eight natural reference vowel sounds were synthesised using the SinSyn tool. Sinewave levels were set according to the experiment described in Chapter M6.5. The sound duration was 1 sec., including a 0.05 sec. fade in/fade out. As a result, a sample of eight sinewave vowel sounds was created.

***f_o* measurement:** Acoustic analysis of the natural and the synthesised sounds accorded to the standard procedure of the Zurich Corpus. Concerning natural sounds, the range of calculated f_o was 217–228 Hz. (This difference between intention and production of f_o not exceeding a semitone is neglectable here.)

Listening tests: The vowel and pitch recognition tests accorded to the procedures described in the previous chapter, with the second pitch recognition subtest related to sinusoid frequencies of 330 Hz (above F_1/S_1 of the investigated sounds of close vowels) and 440 Hz (below F_1/S_1 of the investigated sounds of the open-mid and open vowels, and also corresponding to F_1/S_1 of the investigated sounds of the close-mid vowels).

Results

Table 1 in the chapter appendix shows the sound sample and the vowel and pitch recognition results, including measured f_o values for the synthesised sounds and sound links. The sound links present the pairs of natural reference sounds and their sinewave replicas.

With only marginal differences, the results were in line with the findings of the previous experiments. According to the vowel recognition results (labelling majority), all three sounds with *S*-patterns related to *F*-patterns of close vowels were recognised according to vowel intention, with a recognition rate of 100%. On the contrary, all three sounds related to *F*-patterns of close-mid vowels were confused and were identified as close vowels with a recognition rate of ≥ 70%. The sound related to the *F*-pattern of /ɛ/ was mostly confused with /e/, and the sound related to the *F*-pattern of /a/ was mostly identified as /a/ or /ɔ/.

Concerning pitch recognition, uniform results were again obtained across listeners, tests and sound presentation order: The pitch level of all sounds related to the *F*-patterns of close vowels was recognised as being lower than the pitch level of all sounds related to the *F*-patterns of close-mid vowels (first pitch recognition subtest) and as lower than the sinusoid of 330 Hz (second pitch recognition subtest). The pitch level of all sounds related to the *F*-patterns of close-mid vowels was recognised as being lower than the pitch level of all sounds related to the *F*-patterns of the open-mid and the open vowels, and it was identified as being above the 330 Hz sinusoid and lower than or equal to the 440 Hz sinusoid. Finally, the pitch level of all sounds related to the *F*-patterns of the open-mid and the open vowels was recognised as being higher than the 440 Hz sinusoid. Notably, again, pitch recognition accorded to the frequency ranges of S_1.

Discussion

The vowel and pitch recognition of the synthesised sounds of the present experiment with *S*-patterns relating to *F*-patterns of single natural sounds produced by women corresponded to the recognition of synthesised sounds with *S*-patterns relating to statistical average *F*-patterns. (Note that the range of f_o levels for the single natural sounds investigated here was comparable to the range of average levels generally given in formant statistics.) Above all, the results again strongly supported the two notions of recognised vowel quality being related to pitch and this relation being nonuniform: Firstly, vowel confusion occurred for sounds of close-mid and open-mid vowels with higher pitches than the sounds of close vowels, which were recognised according to vowel intention. Secondly, the effect of high pitch levels on vowel recognition was somewhat limited for the sounds of intended /a/, the most open vowel quality (note that the difference between /a/ and /ɔ/ is not a difference of two long vowel qualities in Standard German). As mentioned, this accorded with the observation of a weak or absent relation of the *F*-patterns or spectral envelopes to f_o for sounds of that vowel (see Chapters 2 and 3).

In this context, some further considerations are of importance. As Remez et al. (1981) stated, one would expect that replicas of speech sounds produced with three or four sinusoids with no harmonicity are perceived as three or four individual sinusoids, as single and simultaneous tones. However, they are often recognised as speech-like utterances, as was the case in the present experiment: Without exception, the replicas of natural vowel sounds investigated here were recognised as having a dominant or prominent vowel quality and a dominant or prominent pitch level. However, most importantly, the results of both sinewave experiments presented also showed that no general statement should be made about an overall vowel recognition of sinewave replicas but that the recognition relates to specific vowel qualities and interacts with pitch. Besides, the measured f_o for four of the sinusoid vowel sounds was not found to be related to S_1 or to pitch levels (see

Table 1, Column 7, intended vowels /i, y, u, ø/; see also the previous experiment discussed in Chapter 6.5).

The number of natural vowel sounds and related *F*- and *S*-patterns investigated here was very limited. Likely, a more extensive investigation, including different *F*- and *S*-patterns for one single vowel quality and a variation of sinewave level configurations, will yield more complex recognition results.

When exploring sinewave speech in terms of replicating read text, intelligibility is sometimes surprisingly high. However, exploring sinewave replicas of minimal pairs, we experienced vowel confusions similar to the two experiments reported here (investigation of minimal pairs of the Zurich Corpus by the author, unpublished). In these terms, concerning vowel recognition, we consider sinewave speech intelligibility as not being directly comparable to the recognition of isolated sinewave vowel sounds or sounds in the context of minimal pairs or syllables.

Chapter appendix

Table 1. Synthesised three-sinusoid sounds related to single natural vowel sounds produced by women: *S*-patterns investigated, f_o measured and vowel and pitch recognition results. Columns accord with Table 1 of the previous chapter. For some occurring (marginal) differences between estimated *F*-patterns when conducting the experiment and *F*- and *S*-patterns as given in the Zurich Corpus, see the Introduction (differences in the figures are < 5 Hz).
[M-06-06-T01]

Table 1. Synthesised three-sinusoid sounds related to single natural vowel sounds produced by women: S-patterns investigated, fo measured and vowel and pitch recognition results. [M-06-06-T01]

Sound							Recognition (summary)															Recognition (details)				
Reference			Synthesis			fo	Vowel and openness									Pitch (test 1)			Pitch (test 2)			Vowels				
S/L	VO	V	S-pattern				close			close-mid			open-mid			low	middle	high	low	middle	high	L1	L2	L3	L4	L5
			S1	S2	S3								open			1<2	2>1	3>2	<330Hz	>330Hz	>400Hz					
			Hz	Hz	Hz	Hz	i	y	u	e	ø	o	ɛ	a	ɔ		2<3			≤440Hz						
	close																									
		i	262	2457	3514	130	10									10			10			i i	i i	i i	i i	i i
1 ⧉		y	235	2019	2474	118	10									10			10			y y	y y	y y	y y	y y
		u	274	762	(2750)	69	10									10			10			u u	u u	u u	u u	u u
	close-mid																									
		e	440	2511	3065	437	8			2							10			10		i i	i i	e e	i i	i i
2 ⧉		ø	435	1753	2846	218		8			2						10			10		y y	y y	ø ø	y y	y y
		o	449	893	3185	449			7			3					10			10		u u	u u	o o	u o	u u
	open-mid open																									
		ɛ	596	(2350)	(3200)	599				9		1						10			10	e e	e e	e o	e e	e e
3 ⧉		a	661	1281	2932	655						4		2	4			10			10	ɔ ɔ	ɔ ɔ	o o	o o	a a

M6.7 Harmonic Synthesis I – Changing Vowel Quality by Changing Either the Lower Spectral Energy Maximum or the Highest Common Factor of Sinusoid Configurations

Introduction

As shown, natural voiced vowel sounds can be recognised with suppressed first or suppressed first and second harmonic(s). For these cases, the HCF of the HP-filtered spectrum is not affected and can be assumed to generally represent the sound periodicity to which perception and recognition relate. However, in contrast, synthesised vowel sounds with only a few partials in their spectrum can be recognised, too, independently of whether or not the partials are in a harmonic relation, and pitch recognition tests indicated that, for these sounds, the frequency of the lowest partial often represents the sound periodicity which perception and recognition relate to. In consequence, concerning vowel recognition of voiced-like sounds, fundamental frequency – and pitch – do not simply relate to the first harmonic and its multiples of the vowel spectrum, that is, *H*1 and HCF.

From this perspective, in two further sinewave vowel synthesis experiments, attempts were made to trigger a vowel quality shift by a change in either the HCF or S_1 (frequency of the lowest sinusoid used in synthesis). In the first experiment, based on sounds produced with three sinusoids in a harmonic relation, the HCF was altered in terms of changing either S_2–S_3 distance (sounds expected to be recognised as front vowels) or S_1–S_2 distance (sounds expected to be recognised as back vowels). In the second experiment, based on sounds with a single lower sinusoid < 1 kHz combined with equal-amplitude sinusoid series in frequency ranges > 1 kHz, all sinusoids in a harmonic relation, periodicity variation was caused by changing either the frequency of the low harmonic < 1 kHz (change of a relative spectral maximum) or the frequency distance of the higher harmonics > 1 kHz (change of HCF) with the frequency range of the higher harmonics kept unchanged. The first experiment was conducted in the context of sinewave synthesis related to statistical *F*-patterns (see Maurer, Suter et al., 2018), the second experiment was conducted in the context of synthesised sounds with flat vowel spectra or spectral parts (see Maurer and Suter, 2017a, b). The experiments are transferred to this treatise with additional testing of pitch recognition (see experiment 2).

Experiment 1

Experimental design and *S*-patterns investigated: Based on the previous experimental experience with sinewave vowel sounds, we aimed at a synthesis setting in which a change in harmonicity – that is, a change in HCF of all sinusoids used for synthesis – would cause a change in recognised vowel quality and pitch level. Following this approach, pairs of S_1–S_2–S_3 configurations were compiled with fixed S_1 and S_3 and varying S_2 only. All three sinusoids of both configurations of a pair were in a harmonic relation, with frequency levels of HCF being either $0.5 \times S_1$ (configuration a) or equal to S_1 (configuration b), that is, changing HCF by one octave. For an illustration of the experimental design, refer to the sound links in Table 1 in the chapter appendix.

For sounds expected to be recognised as front vowels, S_1 was set to 400, 410 or 420 Hz, and S_2 and S_3 were set to ≥ 1.2 kHz. For sounds expected to be recognised as back vowels, S_1 was set to 400 Hz, S_2 was set to ≤ 1.2 kHz and S_3 was set to 2.8 kHz. Because a smaller frequency change for lower harmonics is related to a larger change in the higher harmonics, which might affect vowel recognition, S_1–S_2–S_3 configurations related to one-octave HCF variations of 200–400 Hz, 210–420 Hz and 220–440 Hz were investigated for front vowels; however, only S_1–S_2–S_3 configurations related to an HCF variation of 200–400 Hz were investigated for back vowels. The range of HCF was chosen based on the previous experiences regarding sinewave experiments (see Chapters M6.5 and M6.6), formant pattern and spectral shape ambiguity for sounds of adjacent vowels (documented in Chapter M3) and the perceptual effect observed by the author when creating the experiment. As a result, a total of 15 pairs of S_1–S_2–S_3 configurations were analysed. (For the full sound sample, see Maurer, Suter et al., 2018, Materials).

Sinewave synthesis: Based on these *S*-patterns, static sounds of 1 sec. (including a 0.1 sec. fade in/out) were synthesised using the SinSyn tool, with the sinewave levels set to 100–80–80 dB for sounds of front vowels and 100–90–70 dB for sounds of back vowels. As a result, a sample of 30 sounds (15 sound pairs) was created.

Listening test – vowel recognition: Vowel recognition of the synthesised sounds was tested in a listening test according to the standard procedure of the Zurich Corpus and involving the five standard listeners. Three test specifications were adopted: Each sound was presented twice in the test (10 identifications per sound in total), /ə/ was included as a labelling option, and labelling was restricted to single vowel categories excluding vowel boundaries (forced choice).

Listening test – pitch recognition: Pitch recognition of the sounds was tested in an experiment-specific listening test, again involving the five standard listeners of the corpus. Single sounds were presented, and the listeners were asked to label the pitch level they recognised using a prepared paper form and an online electronic piano keyboard (assignment of the dominant or prominent pitch level only, forced choice; the listeners wrote down levels as musical notes within the C-major scale).

Selection of sounds for documentation and discussion: When analysing the results of the entire sound sample, vowel and pitch recognition results were found to vary markedly among the different *S*-patterns of the 15 sound pairs investigated. Above all, some pairs showed a marked vowel quality shift in an open–close direction related to both an increase of HCF and an indication of an increase in the recognised pitch level, while for other pairs, no indication of a vowel quality shift or only a weak one was observed (see Maurer, Suter et al., 2018, Materials). However, for an increase of HCF, neither inverse vowel quality shifts in a close–open direction nor inverse decreasing pitch levels occurred. On this basis, for this treatise, we decided to select four sound pairs related to four recognised close-mid–close vowel quality shifts /e–i/, /e–y/, /ø–y/ and /o–u/ in terms of "best cases" (i.e., highest recognition rates for vowel quality shifts and associated pitch level shifts in the above recognition tests) in order to demonstrate and document cases for which this type of a change of HCF could trigger a parallel vowel quality and pitch level shift.

Crosscheck of corresponding values for HCF and measured f_o: For all selected synthesised sounds, f_o was calculated according to the standard procedure of the Zurich Corpus, and the resulting frequency levels were compared with HCF.

Results 1

Table 1 in the chapter appendix shows the *S*-patterns of the four selected sound pairs and the vowel and pitch recognition results investigated in experiment 1, including sound links (as mentioned, see these links for an illustration of the spectral configurations examined). In addition, Table 2 shows the individual recognition profiles of single listeners and an analysis thereof.

Vowel recognition: For all four sound pairs presented, according to the labelling majority, increasing HCF resulted in a close-mid–close vowel quality shift (see Table 1, Columns 7–12). For the sound pairs recognised as front vowels, an increase in HCF and a related close-mid–close vowel quality shift resulted from a decrease in S_2. Notably, therefore, the sound of /i/ was associated with a lower S_2 than the sound of /e/ (see Series 1), in opposition to statistical F_2 generally reported as being higher for sounds of /i/ than of /e/. The same held true for the sounds of /y/ and /e/ (see Series 2) and /y/ and /ø/ (see Series 3). For the sound pair recognised as back vowels, an increase in HCF and a related close-mid–close vowel quality shift resulted from an increase in S_2. Notably again, there is no general indication given in the literature on formant statistics that a change of only F_2 is related to an /o/–/u/ change in recognised vowel quality. Besides these general results, between-listener differences and recognition inconsistencies (within-listener recognition differences for the two equal sounds presented in the test) were observed (see Table 2). Note also a front–back "confusion" for listener L3 in Series 3.

Pitch recognition: According to the labelling majority, a one-octave upward shift was indicated by the listening test results for three of the four sound pairs. However, the indication was only pronounced for the sounds recognised as back vowels, and marked between-listener differences occurred.

Details of the parallelism between vowel quality and pitch level shifts taking into consideration between- and within-listener differences: Because of the somewhat limited indication of parallelism of vowel quality and pitch level shifts and because of the marked listener-specific recognition differences, the results were further analysed concerning the individual listener recognition profiles and the possible combinations of vowel quality and pitch level shifts and shift directions. This analysis is shown in Table 2. In the upper part of the table, the individual listener profiles are shown. In the lower part of the table, different types or configurations of the relation of simultaneous vowel quality and pitch level recognition and their classification are shown. Note that the vowel quality shifts investigated only concerned vowel openness in terms of close-mid–close shifts. Unrounded–rounded differences were ignored.

According to the indications in the table, Listener L1 perceived consistent parallel close-mid–close vowel quality and upward pitch level shifts (see V–P shift relation Type 1, in short Type 1); listener L2 did not hear any consistent vowel quality shifts and only perceived low pitch levels (see Type 6); listener L3 recognised a consistent parallel vowel quality and upward pitch level shift for one sound pair (Type 1), an inconsistent vowel quality shift associated with an upward pitch shift for a second sound pair (see Type 3) and inconsistent vowel quality shifts associated with high pitch levels for the remaining two sound pairs (see Type 4/4a); listener L4 recognised consistent parallel vowel quality and upward pitch level shifts for two sound pairs (Type 1) and inconsistent vowel quality shifts associated with high pitch levels for the other two sound pairs (see Type 4/4a); listener L5 recognised consistent vowel quality shifts for all sound pairs, but only one of these shifts was associated with an upward pitch level shift (Type 1), the pitch of all other sounds being labelled on higher levels (Type 2/2a).

Further, we analysed whether there were sound pairs for which only a single lower pitch level was associated with a vowel quality shift (see Condition 2, Types 2b and 4b) and whether there were sound pairs for which only a single vowel quality (equal in openness) was associated with a pitch level shift or with only a higher level (see Condition 2, Type 5). Notably, none of these recognition patterns occurred: For sound pairs with consistent or inconsistent close-mid–close vowel quality shifts, for all sounds investigated, either upward pitch level shifts from lower to higher levels or higher levels for both sounds of a sound pair occurred. If no close-mid–close shifts occurred, lower pitch levels were recognised. (However, see the inconsistent vowel quality recognition of listener L2 and sound pair 3, associated with a lower pitch level.)

Correspondence of HCF and measured f_o: For all selected sounds, measured f_o corresponded to HCF.

Experiment 2

Experimental design and *S*-patterns investigated: On the basis of extensive acoustic analyses of natural front vowel sounds with flat spectral envelopes or flat envelope parts (vowel-related frequency ranges with consecutive harmonics equal in amplitude, see below, Chapter M7.3; see also the Preliminaries, Chapters 7.2 and M7.2) and on the basis of a broader investigation of synthesised sounds related to consecutive harmonics equal in amplitude (see below, Chapter 7.4), an attempt was made in the second experiment to trigger vowel quality and pitch level shifts by changing the frequencies of either *S*1 or HCF. The main idea was to create an experimental design in which two different spectral variations were directly opposed, possibly triggering the same change in the recognised quality of front vowels but with only one type of spectral variation affecting the sound periodicity. At the same time, a spectral peak structure of harmonics > 1 kHz was avoided to create equal bands of higher spectral energy for the configurations compared with each other. In consequence, no filter configuration corresponded to the higher spectrum > 1 kHz. For an illustration of the experimental design, refer to the sound links in Table 3 in the chapter appendix.

Eight series of *S*-patterns were compiled with configurations as detailed below (for numerical values, see Table 3 for Series 1–7 and Table 4 for Series 8 in the chapter appendix).

Series 1–3: For each single series, three *S*-patterns in terms of three sinewave configurations a, b and c were created for a comparison of the related synthesised sounds. The first two *S*-patterns of a series (a and b) consisted of identical consecutive equal-amplitude harmonics as multiples of 220 Hz in a series-specific frequency range above 1 kHz and, therefore, the frequency ranges of the higher harmonics and HCF of the two sounds were identical. However, the patterns differed concerning the frequency level of an additional single low harmonic at either 220 Hz or 440 Hz. The second and third *S*-pattern of a series (b and c) consisted of an identical frequency level of the single low harmonic at 440 Hz and an identical range of consecutive equal-amplitude harmonics above 1 kHz, but the higher harmonics differed in their frequency distance, this distance being either 220 Hz or 440 Hz. Therefore, HCF differed for the second and the third sound, the HCF frequencies being either 220 Hz or 440 Hz.

Thus, *S*-patterns a and b differed concerning a one-octave shift of the low harmonic from 220 Hz to 440 Hz with an unchanged HCF of 220 Hz, and *S*-patterns b and c differed concerning a one-octave HCF shift from 220 Hz to 440 Hz with an unchanged low harmonic equal in frequency. In these terms, an attempt was made to produce sound pairs related to a close-mid–close vowel quality shift with either a decrease of low harmonic frequency level (b to a) or an increase of HCF (b to c). From the perspective of formant theory, vowel quality shifts related to frequency differences of low harmonics might be expected, as sounds of close vowels are predicted to manifest a lower F_1 than sounds of close-mid vowels. However, vowel quality shifts related to a change in HCF can barely be explained within the framework of the prevailing theory. Note that, within a series of *S*-patterns, *S*-pattern b represents the reference sinewave configuration for the two types of spectral variation (see corresponding arrows in Table 3).

Series 4–6: The *S*-patterns of Series 4–6 corresponded to the concept described for the first three series, except for the single low sinusoid frequency being either 300 Hz or 450 Hz and the frequency distance of the higher harmonics being either 150 Hz or 450 Hz. As a consequence, comparing configurations a and b, they differed in an approximately seven-semitone shift of the low harmonic from 300 Hz to 450 Hz with an unchanged HCF of 150 Hz, and configurations b and c differed in an approximately 1.5-octave HCF shift from 150 Hz to 450 Hz with an unchanged lower harmonic. These series were created to increase upward shifts in HCF and examine whether the associated vowel quality shifts would prove to be more pronounced. In addition, two series for the vowel sounds /e/ and /i/ were investigated to examine a reduction of the upper frequency limit of the harmonics from 3520 Hz to 3150 Hz, the latter being more comparable to the vowel-related spectral energy observed for natural sounds of adults.

Series 7: The *S*-patterns of Series 7 also accorded to the concept described for the first three series, except for the single low sinusoid frequency being either 450 Hz or 600 Hz and the frequency distances of the higher harmonics being either 150 Hz or 300 Hz. This series was created to decrease the frequency distance of the higher harmonics for both lower and higher HCF of the sounds.

Series 8: In the last series, six *S*-patterns (a–f) were created for a comparison of the related synthesised sounds. These *S*-patterns again accorded to the same concept as applied in the previous series, except for testing out three different single low harmonic frequency levels of 300–450–600 Hz, and three different frequency distances of the higher harmonics and HCF, 150–300–600 Hz, respectively. When comparing configurations a, b and d, they differed in two subsequent limited shifts of the low harmonic of approximately seven and five semitones (one octave in total), with unchanged HCF of 150 Hz, and configurations d, e and f differed in two subsequent octave shifts of HCF from 150 Hz to 300 Hz and then from 300 Hz to 600 Hz, with an unchanged low harmonic. These *S*-patterns of Series 8 were created to increase HCF shifts further and attempt to produce synthesised sounds related to open-mid–close-mid and subsequent close-mid–close vowel quality shifts. Note that, in this series, configuration d represents the reference configuration for the two types of spectral variation. Configuration c was added as an intermediate configuration for comparison with configuration e.

Sinewave synthesis: Based on the above *S*-patterns, static sounds were synthesised using the SinSyn tool. All sinewave levels were set to 100 dB. Sound duration was 1.2 sec., including a 0.1 sec. fade in/fade out. As a result, a sample of 27 synthesised sounds was created.

Listening tests – vowel recognition: The vowel recognition of the synthesised sounds was tested in two listening tests according to the standard procedure of the Zurich Corpus and involving the five standard listeners, with the labelling restricted to long Standard German vowel qualities (forced choice, no vowel boundaries). In the first test, the test items consisted of single sounds, which the listeners were asked to assign to a dominant or prominent vowel quality. In the second test, each test item consisted of two different sounds to investigate vowel contrasts: For Series 1–7, sound b was presented with either sound a or sound c per test item (separated by a 1 sec. pause), in AB and BA order. For Series 8, likewise, sound d was presented with each of the other sounds per test item in AB and BA order. The listeners were asked to assign the dominant or prominent vowel quality of the second sound only. The sounds were tested in eight series-specific subtests separated by a break of a minimum of 15 minutes.

Listening test – pitch recognition: Sound presentation and test procedure of the pitch recognition test accorded with the second vowel recognition test. The listeners were asked to label whether the pitch level of the second sound when compared to the level of the first sound was falling, flat or rising, referring to dominant or prominent levels.

Crosscheck of corresponding values for HCF and measured f_o: For all synthesised sounds, f_o was calculated according to the standard procedure of the Zurich Corpus, and the resulting frequency levels were compared with HCF.

Results 2

Tables 3 (Series 1–7) and 4 (Series 8) in the chapter appendix show the *S*-patterns and the vowel and pitch recognition results for the synthesised sounds investigated in this second experiment, including sound links (as mentioned, see these links for an illustration of the spectral configurations examined). The results for the reference sounds b (Series 1–7) and d (Series 8) are marked in blue, the results for the synthesis condition with lowered S_1 are marked in green, and the results for the synthesis condition with increased HCF are marked in red.

Vowel recognition for sounds related to the variation of *S*1 frequency only (HCF equal in frequency): Comparing sounds b and a (Series 1–7) or d, b and a (Series 8), decreasing the frequency level of one single low harmonic but keeping HCF of all harmonics unchanged (HCF equal in frequency) resulted in a pronounced change of vowel quality in an open–close direction. For the sounds of all series, if unrounded–rounded variants are disregarded, the recognition rates for the open-mid–close-mid, open-mid–close and close-mid–close shifts were ≥ 92%.

Vowel recognition for sounds related to HCF variation only (*S*1 equal in frequency): Comparing sounds b and c (Series 1–7) or d, e and f (Series 8), increasing the frequency of HCF but keeping the low harmonics unchanged (*S*1 of synthesis equal in frequency) in turn resulted in a pronounced change of vowel quality in an open–close direction. If unrounded–rounded variants are again disregarded, the recognition rates for the close-mid–close shifts (all Series) and the open-mid–close-mid shift (Series 8) were ≥ 84%. (Note also the /e–y/ shift for sounds b and c in Series 8.)

Pitch recognition for sounds related to HCF variation only (*S*1 equal in frequency): In parallel to the above vowel quality shifts associated with an increase of HCF, with the exception of three single labellings of equal pitch levels (see Series 4 and 7), all listeners identified upward pitch level shifts for all sounds compared.

Additional pitch recognition results: As expected, upward pitch level shifts for sounds a and c in Series 1–7 were recognised by all listeners for all sounds compared (see Table 3, Columns 18–20), this shift being unassociated with a shift in vowel quality but associated with spectral variation of *S*1 and HCF. Regarding the comparison of sounds b and a with equal HCF, one would expect the listening test to reveal uniform equal-pitch labelling. However, some pitch recognition differences were indicated by the results for Series 1–7 in that some pitch identifications were assigned to a lower pitch level for close vowels than for close-mid vowels (see Table 3, Columns 21–23). This finding may point to sound timbre affecting pitch recognition for sounds of this kind (see also below; corresponding results were obtained for sound comparisons a, b and d of Series 8, not given in Table 4).

Correspondence of HCF and measured f_o: For all sounds, measured f_o corresponded to HCF.

Discussion

In two further synthesis experiments based on sinusoids, as explained in the introduction to this chapter, attempts were made to trigger a vowel quality shift by changing either HCF or *S*1 frequency. In the first experiment, investigating three-sinusoid vowel sounds with the sinusoids in a harmonic relation, HCF was altered in terms of changing either S_2–S_3 distance (front vowels) or S_1–S_2 distance (back vowels) while S_1–S_3 was kept unchanged. In the second experiment, investigating sounds with a single low sinusoid < 1 kHz combined with equal-amplitude sinusoid series in frequency ranges > 1 kHz and with all sinusoids in a harmonic relation, spectral variation concerned either a change in the low harmonic < 1 kHz (change of a spectral maximum) or a change in the frequency distance of the higher harmonics > 1 kHz (change of HCF), the frequency range of the higher harmonics being kept unchanged.

The results of the first experiment showed that, due to the variation of a single intermediate sinusoid frequency of a three-sinusoid sound, an HCF increase by one octave could trigger both a vowel quality shift in an open–close direction and a parallel one-octave upward pitch level shift, even if this indication was somewhat limited by marked between- and within-speaker recognition differences and inconsistencies. A detailed analysis of the listeners' recognition profiles showed, however, that vowel quality shifts were associated with either pitch shifts from lower to higher levels or constant higher pitch levels for both sounds of a sound pair, but no close-mid–close vowel quality shift associated with a downward pitch level shift or lower pitch levels for both sounds of a sound pair occurred. This finding can be understood as supporting the indication of associated vowel quality and pitch level shifts and shift directions, above all when considering the "borderline" character of this type of sound in general and taking into account the very unnatural and sharp sound timbre of synthesised sounds used for this experiment in particular, possibly affecting the results of the vowel and pitch recognition tasks: For these sounds, vowel quality, pitch and timbre are very difficult to separate from each other, and results may depend strongly on the listeners, even if they are professionally trained singers or speakers and even if they are experienced in recognition tasks. Further, in this first experiment, the pitch recognition test was not extended to include level comparisons of two sounds presented as a single test item, as was the case in the second experiment. However, the above results of the first experiment give reason to consider different possible effects for the investigation and understanding of the relation of vowel quality to pitch: (i) Occurring variation in recognition results may be due to a specific listener recognition profile and/or to an interaction of sound timbre, vowel quality and pitch and/or to the design of the listening test; (ii) pitch shifts may precede vowel quality shifts; (iii) vowel quality recognition may relate to a perceptual referencing to a sound pattern repetition over time, a process which may not always be clearly revealed in conducted pitch recognition tasks. Finally, the parallelism of vowel quality and pitch level shifts for the selected sound pairs depended on specific configurations of *S*-patterns (specific frequency levels and distances of the sinusoids; for details, see the results section of experiment 1).

The results of the second experiment showed that a vowel quality shift in an open–close direction could be triggered by a downward shift of a single low harmonic only, with the remaining spectral configuration kept unchanged. But, in line with the results of the first experiment, the same shift could also be triggered by changing HCF in terms of changing the frequency distance of the higher harmonics, both the low harmonic and the higher frequency range of prominent spectral energy kept unchanged in synthesis. Thereby, parallel shifts of vowel quality and pitch levels for sounds with HCF variation were very pronounced for sounds of this type.

Comparing vowel quality and pitch level shifts for the sounds of Series 1–3 (one-octave pitch level shift, see Table 3) and Series 4–6 (c. 1.5-octave pitch level shift), comparable results were found. Comparing the results of Series 4 and 5, limiting the upper frequency range of the sinusoids to 3150 Hz resulted in an increase in the occurrence of /y/ instead of /i/, but still with the majority of labelling being /i/. Concerning the sounds of Series 8 with a further increase in the pitch level shift up to two octaves, notably, a vowel quality shift exceeding two adjacent qualities could be demonstrated.

In conclusion, vowel quality shifts could be triggered by either an energy maximum change in the lower vowel spectrum, with HCF and recognised pitch level kept unchanged, or by a change in HCF and recognised pitch level, with the lower prominent spectral energy as well as the frequency range of the higher prominent spectral energy in the vowel spectrum kept unchanged. Thereby, consistent shift directions were found: open–close vowel quality shifts associated with an increase in the pitch level.

Concerning the second experiment, it may be objected that keeping the frequency range of higher equal-amplitude harmonics unchanged but halving the harmonic number (in order to double HCF) represents a spectral change beyond HCF which may well affect vowel recognition. This argument cannot be ruled out entirely. However, notably, the findings of experiment 1 and experiment 2 corresponded concerning HCF variation, and the S_2 variation for the sounds of front vowels in experiment 1 was found to be in opposition to the commonly assumed F_2 variation. Also, the frequency distances of 200–210 Hz and 400–420 Hz of the higher harmonics in vowel synthesis corresponded to distances for which auditory spectral integration is discussed (and often assumed) in the literature (see e.g. Fox et al., 2011). Both aspects run counter to an interpretation of vowel quality shifts directly resulting from spectral variation beyond HCF variation.

Although not explicitly discussed for the synthesis experiments presented in the two preceding chapters as well as in the present chapter, the listeners involved in the recognition tasks of these experiments constantly reported cases of sounds for which, when giving very specific attention to the sound characteristics during the tests, they could recognise two vowel qualities and/or two (or even more) pitch levels (see also the excursus on fundamental frequency and pitch and the corresponding note in Chapter M6.1). The experiments discussed in the following chapters were designed and conducted to integrate this double-vowel and/or double-pitch phenomenon into the general investigation of the vowel–pitch relation (or its alternative).

Chapter appendix

Table 1. Synthesised sounds based on three sinusoids in a harmonic relation, with HCF variation due to changing S_2–S_3 or S_1–S_2 distance: *S*-patterns investigated and vowel and pitch recognition results. Columns 1–5 = sounds (S/L = sound pairs a–b and sound links; S1–S3 = sinusoid frequencies, in Hz; HCF = highest common factor, in Hz). Columns 6–12 = vowel recognition results (summary in terms of the confusion matrix; Maj = labelling majority). Columns 13–14 = pitch recognition results (summary). Colour code: Blue = labelling majority for a close-mid vowel quality, associated with a low HCF and, for three pairs, with a lower pitch level; red = labelling majority for a close vowel quality associated with high HCF and a higher pitch level; purple = occurring case of a labelling majority for a higher pitch level associated with a close-mid vowel quality and low HCF.
[M-06-07-T01]

Table 2. Synthesised sounds based on three sinusoids in a harmonic relation, with HCF variation due to changing S_2–S_3 or S_1–S_2 distance: Analysis of listener-specific recognition profiles. Upper part: Columns 1–5 = see Table 1 (excluding sound links). Columns 6–15 = details of vowel and pitch recognition, per listener (L(i) = listeners; V = vowel quality recognised; P = pitch level assigned, in Hz according to the musical C-major scale). Colour code: Dark blue and dark red = consistent close-mid–close vowel quality shifts, and/or (associated or not associated) a one-octave upward pitch level shift for a sound pair, the shifts associated with an increase in HCF; light blue and light red = inconsistent close-mid–close vowel quality shifts (inconsistent labelling for the two identical sounds presented in the listening test); no colour = no close-mid–close vowel quality shift and no pitch level shift for a sound pair. Lower part: Further analysis of the individual recognition profiles of the listeners (for details, see text).
[M-06-07-T02]

Table 3. Synthesised sounds based on a single lower sinusoid < 1 kHz combined with equal-amplitude sinusoid series > 1 kHz, with a variation of either the lower sinusoid or HCF: *S*-patterns investigated and vowel and pitch recognition results for the sound triplets investigated. Columns 1–8 = sounds (S/L = sound triplet and sound links; S1 and higher S(i) given as numbers = low harmonic and range of higher harmonics given as harmonic numbers H(i) in reference to HCF; S1 and higher S(i) given in Hz = frequency level of the low harmonic and frequency range of the higher harmonics; HCF = highest common factor, in Hz; ΔS1 = frequency difference of the low harmonics of the sounds a and b, in Hz; ΔHCF = frequency difference of HCF of the sounds b and c, in Hz). Columns 9–12 = summary of the recognition results (V = vowel recognition according to the labelling majority; P = pitch recognition of the five listeners for the comparison of sounds b versus c, where l = lower, eq = equal, h = higher). Columns 13 ff. = details of the vowel recognition results, and additional results for pitch recognition (V [matrix] = confusion matrix including the labelling of both vowel recognition subtests, where o-m = open-mid, c-m = close-mid, c = close; P [additions] = pitch recognition of the five listeners for the comparison of sounds a versus c and a versus b). Colour code: Green = S_1 and labelling majority for close or close-mid vowels for sound a; blue = S_1 and labelling majority for close-mid or open-mid vowels associated with a lower pitch level for sound b when compared with sound c; red = S_1 and labelling majority for close or close-mid vowels associated with a higher pitch level for sound c when compared with sound b. Arrows: Green = downward shift of S_1 associated with a vowel quality shift in an open–close direction; red = upward shift of HCF associated with a vowel quality shift in an open–close direction.
[M-06-07-T03]

Table 4. Synthesised sounds based on a single lower sinusoid < 1 kHz combined with equal-amplitude sinusoid series > 1 kHz, with variation of either the lower sinusoid or HCF: *S*-patterns investigated and vowel and pitch recognition results for the sound sextuple investigated. Columns 1–8 = see Table 3. Columns 9–18 = summary of recognition results (V = vowel recognition according to labelling majority; P = pitch recognition of the five listeners for the comparison of sounds d versus e, d versus f and e versus f). Columns 19–23 = details of vowel recognition results (V [matrix] = confusion matrix including the labelling of both vowel recognition subtests, where o-m = open-mid, c-m = close-mid, c = close). Colour code and arrows accord with the system as indicated in Table 3, separating the results related to S_1 or HCF variation.
[M-06-07-T04]

Table 1. Synthesised sounds based on three sinusoids in a harmonic relation, with HCF variation due to changing S2–S3 or S1–S2 distance: S-patterns investigated and vowel and pitch recognition results. [M-06-07-T01]

Sounds						Recognition								
						Vowel							Pitch	
S/L		S-patterns (Hz)				close-mid			close			Maj	Levels	
		S1	S2	S3	HCF	e	ø	o	i	y	u		low	high
1 ↗	a	420	**2730**	2940	210	6			4			e	2	3
	b	420	**2520**	2940	420				9	1		i	1	4
2 ↗	a	400	**2200**	2400	200	7				3		e	3	2
	b	400	**2000**	2400	400					10		y	1	4
3 ↗	a	400	**1400**	1600	200		8			1	1	ø	3	2
	b	400	**1200**	1600	400		1			7	2	y	1	4
4 ↗	a	400	**600**	2800	200			10				o	5	
	b	400	**800**	2800	400			3			7	u	1	4

Table 2. Synthesised sounds based on three sinusoids in a harmonic relation, with HCF variation due to changing S2–S3 or S1–S2 distance: Analysis of listener-specific recognition profiles. [M-06-07-T02]

Sounds						Details of vowel and pitch recognition									
						L1		L2		L3		L4		L5	
Sound pair		S-pattern / HCF (Hz)				V	P	V	P	V	P	V	P	V	P
		S1	S2	S3	HCF		Hz		Hz		Hz		Hz		Hz
1	a	420	**2730**	2940	210	e e	208	i i	208	i e	415	i e	415	e e	415
	b	420	**2520**	2940	420	i i	415	i y	208	i i	415	i i	415	i i	415
2	a	400	**2200**	2400	200	e e	196	y y	196	e e	196	e y	392	e e	392
	b	400	**2000**	2400	400	y y	392	y y	196	y y	392	y y	392	y y	392
3	a	400	**1400**	1600	200	ø ø	196	y ø	196	ø u*	392	ø ø	196	ø ø	392
	b	400	**1200**	1600	400	y y	392	y ø	196	u* u*	392	y y	392	y y	392
4	a	400	**600**	2800	200	o o	196	o o	196	o o	196	o o	196	o o	196
	b	400	**800**	2800	400	u u	392	o o	196	u o	392	u u	392	u u	392

Further analysis						L1	L2	L3	L4	L4
Types		V shift		P shift						
V–P shift types	1	c-m	c	low	high	4	0	1	2	1
	2	c-m	c	no shift		–	–	–	–	3
	3	(c-m)	(c)	low	high	–	–	1	–	–
	4	(c-m)	(c)	no shift		–	–	2	2	–
	5	no shift		low	high	–	–	–	–	–
	6	no shift		no shift		–	4	–	–	–
Cond. 1	1	c-m	c	low	high	4	4	4	4	4
	2a	c-m	c	high	high					
	3	(c-m)	(c)	low	high					
	4a	(c-m)	(c)	high	high					
	6a	no shift		low	low					
Cond. 2	2b	c-m	c	low	low	0	0	0	0	0
	4b	(c-m)	(c)	low	low					
	5	no shift		low	high					
	6b	no shift		high	high					

Table 3. Synthesised sounds based on a single lower sinusoid < 1 kHz combined with equal-amplitude sinusoid series > 1 kHz, with a variation of either the lower sinusoid or HCF: S-patterns investigated and vowel and pitch recognition results for the sound triplets investigated. [M-06-07-T03]

Sounds									Recognition				Recognition (details)										
									V	P			V (matrix)					P (additions)					
S/L		S1	Higher S(i)	S1	Higher S(i)	HCF	ΔS1	ΔHCF		b versus c			o-m	c-m		c		a versus c			a versus b		
		Numbers		Hz	Hz	Hz	Hz	Hz	Maj	l	eq	h	ε	e	ø	i	y	l	eq	h	l	eq	h
	a	H1	H10–H16	220	2200–3520	220	220 ↑		i					1		19	5	5			1	4	–
1	b	H2	H10–H16	440	2200–3520	220	1 oct	220	e	5				25					–			4	1
↗	c	H1	H5–H8	440	2200–3520	440		↓ 1 oct	i			5				23	2			5		–	
	a	H1	H6–H10	220	1320–2200	220	220 ↑		y								25	5			1	4	–
2	b	H2	H6–H10	440	1320–2200	220	1 oct	220	ø	5					25				–			4	1
↗	c	H1	H3–H5	440	1320–2200	440		↓ 1 oct	y			5			3		22			5		–	
	a	H2	H10–H16	440	2200–3520	220	220 ↑		e					25				5			1	4	–
3	b	H4	H10–H16	880	2200–3520	220	1 oct	220	ε	5			25						–			4	1
↗	c	H2	H5–H8	880	2200–3520	440		↓ 1 oct	e			5	4	21						5		–	
	a	H2	H15–H24	300	2250–3600	150	150 ↑		i					2		23		5			2	3	–
4	b	H3	H15–H24	450	2250–3600	150	7 st	300	e	5			2	23					–		–	4	1
↗	c	H1	H5–H8	450	2250–3600	450		↓ 1.5 oct	i		1	4				25				5		–	
	a	H2	H15–H21	300	2250–3150	150	150 ↑		i							17	8	5			2	3	–
5	b	H3	H15–H21	450	2250–3150	150	7 st	300	e	5				25					–		–	4	1
↗	c	H1	H5–H7	450	2250–3150	450		↓ 1.5 oct	i			5				20	5			5		–	
	a	H2	H9–H15	300	1350–2250	150	150 ↑		y								25	5			2	3	–
6	b	H3	H9–H15	450	1350–2250	150	7 st	300	ø	5				2	23				–		–	4	1
↗	c	H1	H3–H5	450	1350–2250	450		↓ 1.5 oct	y			5			1		24			5		–	
	a	H3	H12–H20	450	1800–3000	150	150 ↑		e					18	7			5			1	4	–
7	b	H4	H12–H20	600	1800–3000	150	5 st	150	ε	4	1		25						–			4	1
↗	c	H2	H6–H10	600	1800–3000	300		↓ 1 oct	e		1	4		22	3					5		–	

Table 4. Synthesised sounds based on a single lower sinusoid < 1 kHz combined with equal-amplitude sinusoid series > 1 kHz, with a variation of either the lower sinusoid or HCF: S-patterns investigated and vowel and pitch recognition results for the sound sixtuple investigated. [M-06-07-T04]

Sounds									Recognition										Recognition (details)				
									V	P									V (matrix)				
S/L		S1	Higher S(i)	S1	Higher S(i)	HCF	ΔS1	ΔHCF		d versus e			d versus f			e versus f			o-m	c-m		c	
			Numbers	Hz	Hz	Hz	Hz	Hz	Maj	l	eq	h	l	eq	h	l	eq	h	ε	e	ø	i	y
8	a	H2	H12–H24	300	1800–3600	150			y													13	42
	b	H3	H12–H24	450	1800–3600	150	300 1 oct		e										3	42	10		
	c	(H1)	(H6–H12)	(300)	(1800–3600)	(300)			(y)													20	35
	d	H4	H12–H24	600	1800–3600	150			ε	5			5						55				
	e	H2	H12–H24	600	1800–3600	300		450 2 oct	e			5				5				55			
	f	H1	H3–H6	600	1800–3600	600			y						5			5				26	29

M6.8 Harmonic Synthesis II – Sinewave-Like Replicas Related to Harmonics at or Near Spectral Peaks of Single Natural Vowel Sounds

Introduction

Monotonous sounds produced with sinewave synthesis for the above experiments were of very artificial sound quality, and vowel and pitch recognition may have been affected as a result. This experimental condition is a major drawback. However, as a first approach to sounds of this type of spectral characteristics, we wanted to exclude all dynamic characteristics.

The results of the first two experiments on sinewave vowel sounds described in Chapters M6.5 and M6.6 indicated that, for static three-sinusoid sound synthesis with *S*-patterns relating to estimated *F*-patterns of natural vowel sounds (statistical *F*-patterns and *F*-patterns related to single sounds), vowel recognition was related to pitch, although in a nonuniform manner. The results of the experiments described in the previous chapter supported this indication for synthesised vowel sounds based on sinusoids (with an HCF) which were only indirectly related to natural sounds. Furthermore, although only indicated and not explicitly discussed, the vowel–pitch relation (or its alternative) seemed to be associated with the phenomenon of sounds for which listeners could recognise two vowel qualities and/or two (or even more) pitch levels.

The sound samples investigated in the previous experiments on sinewave vowel sounds were small for the following reasons: The number of sounds should allow for an initial introductory exploration, including different recognition tests; the test results should be easily comprehensible and also verifiable by the reader when reproducing and listening to the sounds in question; the experimental setting should offer a paradigm for replication using the *F*-patterns investigated here or using *F*-patterns of other languages or other single sounds, possibly also including *F*-pattern variation for sounds of one vowel quality.

In order (i) to further evaluate vowel and pitch recognition for synthesised sinewave-like vowel sounds related to two or three harmonics (dominant harmonics in the spectrum of a natural reference sound), but using sounds with a more natural sound quality than was the case in the previous experiments, (ii) to address at the same time the question of different configurations of spectral peak frequencies for sounds of a given vowel and (iii) to extend the investigation in terms of including the question of double-vowel and/or double-pitch recognition, a further sinewave-like experiment on the matter of the vowel–pitch relation was conducted based on extracted harmonics (their dynamic course) of natural vowel sounds at or near the first three estimated peaks in their spectrum.

Experiment

Selection of speakers and sounds: Based on the Zurich Corpus, for each of the eight long Standard German vowels and each of the three age- and gender-related speaker groups of men, women and children and related intended levels of f_o of 131 Hz, 220 Hz and 262 Hz, respectively (levels usually associated with age- and gender-related statistical *F*-patterns), three natural vowel sounds produced with voiced phonation in nonstyle mode and V context were selected which manifested a spectral peak structure that allowed for the assignment of single harmonics as their representation. Vocal effort was disregarded. The vowel recognition rate was 100% (matching vowel intention) for all selected sounds according to the standard listening test conducted when creating the corpus. As a result, a sample of 72 natural reference sounds was created.

Spectral basis, sound selection, assignment of *D*-patterns: The sound evaluation was based on average spectra calculated according to the standard acoustic analysis of the corpus. Sounds of the back vowels and /a/ were selected for which two peaks were manifest in the spectrum below 2 kHz. However, a few examples of sounds with only one peak in this frequency range were also included. Sounds of the front vowels were selected for which three or more peaks were manifest in the entire spectrum.

For the selected sounds of back vowels and /a/, *D*1–*D*2 patterns in terms of the two dominant or prominent harmonics that corresponded to either the two lower spectral peaks or the estimated spectral envelope of a sound were assigned. For the selected sounds of front vowels, *D*1–*D*2–*D*3 patterns in terms of one dominant or one of two prominent harmonics that corresponded to one of the first three peaks of a sound spectrum were assigned. Note that the selected dominant or prominent harmonics are abbreviated here as *D*1–*D*2 or *D*1–*D*2–*D*3 (including their levels) or D_1–D_2 or D_1–D_2–D_3 (frequencies only), and their patterns are termed *D*-patterns in order not to confuse the number of a dominant or prominent harmonic and the number of any harmonic *H*(i) in the original spectrum of the natural sound (see the Introduction). As a result, a sample of 72 *D*-patterns in total was created.

In the selection process, the inclusion of some spectral variation for sounds of a single vowel was also attempted: If possible, for a given vowel and a given f_o level, sounds with one or two different spectral peak frequencies and/or peak levels resulting in different patterns of harmonic frequencies and/or levels of dominant harmonics used for synthesis were selected.

Extraction of harmonics and subsequent harmonic synthesis: The dynamic harmonic spectra of the entire natural reference sounds were analysed using the HarmSyn tool (default parameter setting). Subsequently, based on the analysis and the selected two or three harmonics assigned in a *D*-pattern, sounds were synthesised. As a result, a sample of 72 synthesised sounds was created.

Original reference sounds, list of *D*-patterns and synthesised replicas: Table 1 in the chapter appendix lists the selected natural reference sounds and the assigned *D*-patterns for sound synthesis. The sounds are accessible via sound links: For each vowel and in the order of sound listing in the table, a link refers to the pairs of natural reference sounds and their synthesised two- or three-harmonic replicas. The presentation of sounds and spectra serves the purposes of sound playback and illustration.

Listening test: Vowel and pitch recognition of the synthesised sounds was investigated in two experiment-specific tests involving the five standard listeners of the Zurich Corpus. In the first test, single synthesised sounds were presented in random order, and the listeners were asked to simultaneously label the recognised dominant or prominent vowel quality and dominant or prominent pitch level. Vowel quality labelling accorded with the standard procedure of the corpus, including vowel boundaries. For pitch level recognition, the listeners used an online electronic piano keyboard. They selected a note with a level (C-major scale) comparable to the level of the synthesised sound in question. In the second test, single synthesised sounds were again presented in random order, and the listeners were asked to assign whether they heard a second non-dominant or non-prominent vowel quality and/or a second non-dominant or non-prominent pitch level. If this was the case, they were asked to label the respective vowel quality and/or pitch level according to the procedure of the first test.

Sounds produced by men, women and children were tested separately (subtests of 24 sounds per speaker group). Before each subtest, the listeners listened to the sounds of a subsample in random order to become familiar with the timbre of the sounds. The hardware of the tests accorded to the standard procedure of the corpus. The listeners were given a prepared paper form and were asked to write down their answers on this form.

Results

As indicated, Table 1 in the chapter appendix shows the sound sample and assigned *D*-patterns and provides sound links. Table 2 shows the results of dominant or prominent vowel and pitch recognition, and Tables 3 and 4 show the results of double-vowel and/or double-pitch recognition. Below, general results for vowel and pitch recognition are discussed with respect to the corresponding labelling majority, and details are given concerning single identifications of single listeners.

Sound sample (see Table 1): As a consequence of the attempt to create different harmonic configurations for sounds of a vowel and to investigate the impact of this variation on vowel and pitch recognition, the sound sample included variations in *D*-patterns within and between speaker groups (and associated f_o levels). Notably, this variation caused some variation in HCF (see Table 1, Column 10) which is important to consider with regard to the pitch recognition results.

Recognition of dominant or prominent vowel quality and pitch level for synthesised sounds related to natural reference sounds of close vowels (see Table 2): All synthesised replicas of the natural close vowel sounds were recognised as close vowels and, for the most part, were successfully matched with the intended vowel quality of the natural reference sound. Aside from this general finding, single labelling for vowel boundary recognition, some single unrounded–rounded confusions, two single sounds with front–back confusions and one sound with a single labelling of a close-mid vowel occurred. Except for the second replica of /y/ produced by a man and a few single identifications, the listeners assigned pitch levels to the lower and narrow frequency band of 196–262 Hz; the recognised pitch level for the second replica of /y/ of men was assigned to a lower frequency band of 110–131 Hz.

Recognition of dominant or prominent vowel quality and pitch level for synthesised sounds related to natural reference sounds of close-mid vowels (see Table 2): For the synthesised replicas of the natural close-mid vowel sounds, the results were somewhat vowel-specific. Concerning /e/, all replicas were confused in terms of a close-mid–close shift when compared with vowel intention. With few exceptions, the recognised pitch levels were assigned to the middle frequency band of 392–440 Hz for the replicas of the adults and equal to or above 523 Hz for the replicas of the children, that is, approximately one octave or more above the replicas of close vowels. Concerning /ø/, seven of nine replicas were confused in terms of a close-mid–close shift when compared with vowel intention. In parallel, except for one sound, the recognised pitch levels mostly corresponded to the levels found for sounds of /e/. Concerning /o/, four of nine replicas were confused in terms of a close-mid–close shift when compared with vowel intention, four replicas were recognised according to vowel intention, and one replica was recognised in the vowel boundary of /o–u/. In parallel, recognised pitch levels were somewhat scattered. However, vowel confusion was markedly higher for sounds with higher pitch levels.

Recognition of dominant or prominent vowel quality and pitch level for synthesised sounds related to natural reference sounds of open-mid and open vowels (see Table 2): All synthesised replicas of the open-mid vowel were confused (open-mid–close-mid or open-mid–close shifts when compared with vowel intention). In parallel, the recognised pitch levels were somewhat scattered but with a marked tendency towards middle and higher levels up to 659 Hz and above. Concerning /a/, five replicas were mostly recognised according to vowel intention, and four replicas were recognised as /a/ or /ɔ/ or in the /a–ɔ/ boundary. The recognised pitch levels were again scattered, with a tendency towards the middle and higher levels up to 659 Hz and above.

Pitch level, D_1 and HCF: The relation of the frequency levels of recognised pitch, D_1 and HCF was investigated in a corresponding comparison, also given in Table 2. For the replicas of close, close-mid and open-mid vowel sounds, recognised pitch levels related either to D_1 only or to D_1 and HCF with their frequency levels being equal, with few exceptions. However, the relation was mixed for the sounds of /a/, with the pitch relating to either *D*1 and/or HCF or neither of them.

Recognition of secondary vowel quality and pitch level for synthesised sounds (see Tables 3 and 4): Testing double-vowel and/or double-pitch recognition, that is, a second vowel quality and/or a second pitch level which is recognised in addition to the dominant or prominent vowel quality and pitch level of a replica, provided several main results:

– Parallel double-vowel and double-pitch recognition of single listeners occurred for 35 of the 72 synthesised replicas and concerned 39 single identifications. Except for two single labellings, the secondary pitch level was recognised above the primary level and, in most cases, exceeded 523 Hz.

- Double-pitch-only recognition of single listeners occurred for 60 of the 72 synthesised replicas and concerned 114 single identifications.
- In these terms, double-pitch recognition without double-vowel recognition was far more frequent than parallel double-vowel and double-pitch recognition. Thus, double-vowel and double-pitch recognition were not in a strict relation. Furthermore, parallel double-vowel and double-pitch recognition were also highly listener-specific: In most cases of single sounds, such an associated double recognition concerned only one single listener, in contrast to double-pitch recognition (without double-vowel recognition) which often occurred for two or more listeners.
- Double-vowel-only recognition of single listeners occurred for 4 of the 72 synthesised replicas and only concerned 4 single identifications. Hence, it was very rare.
- Besides, double-vowel and double-pitch recognition also proved to be somewhat dependent on vowel qualities (compare e.g. the replicas of /e/ and of /u/), on the speaker group and/or on f_o of the reference sounds (compare e.g. the replicas of /y/ and of /ɛ/) and on the individual harmonic configurations for sounds of a given vowel (compare e.g. the replicas of /e/ and of /a/ of men).

Further examining cases of double-vowel and related pitch recognition (see Table 3, rows DV&DP and DV; for details, see Table 4): In addition to the above general findings, some details are worth noting regarding double-vowel recognition:

- Vowel qualities of double-vowel recognition depended on the qualities of the natural reference sounds. For replicas of the close front vowels, except for two cases, double-vowel recognition concerned only simultaneous close unrounded–rounded qualities associated with two recognised pitch levels. Exceptions were one simultaneous /e/ and /i/ and one simultaneous /u/ and /y/ recognition associated with an increase in recognised secondary pitch level (compared to the recognised dominant level). For replicas of the close back vowel, no double-vowel recognition occurred. For replicas of /e/ and /o/, with one exception, double-vowel recognition either concerned simultaneous close-mid and close qualities or simultaneous close unrounded and rounded qualities. For replicas with simultaneous close-mid and close qualities, the recognised close qualities were associated with a recognised increase in the secondary pitch level. Likewise, for replicas of /ɛ/, double-vowel recognition either concerned simultaneous close-mid and close qualities or simultaneous close-mid unrounded and rounded qualities associated with either an increase in recognised secondary pitch level or no secondary pitch level. Corresponding results were also found for seven of the eleven replicas of /ø/. However, for this vowel, three cases of increased secondary pitch levels and inverted close–close-mid shifts for primary and secondary vowel qualities also occurred. For the four replicas of /a/, double-vowel and double-pitch recognition included almost all of the above types and included a single identification of a secondary pitch level below the dominant level.

- For sounds associated with two vowel qualities and two pitch levels, with few exceptions, the secondary pitch level was equal to or above 698 Hz, that is, above the dominant or prominent level. For only four replicas, a single listener recognised only one pitch level but two vowel qualities. However, for three of these four cases, another listener recognised two vowel qualities and two pitch levels. In these terms, double-vowel recognition not associated with double-pitch recognition was almost lacking.
- Similarly to the other vowel synthesis experiments reported in this treatise, an occasional case of front–back confusion also occurred in this study (see the corresponding second sound of /i/ of children).

Discussion

Concerning the recognition of dominant vowel qualities and pitch levels, the results again strongly supported the thesis that vowel quality is related to pitch and that this relation is nonuniform: If unrounded–rounded confusions are ignored, marked vowel confusion in terms of marked vowel quality shifts in an open–close direction did not occur for synthesised replicas of the natural close vowel sounds and, in parallel, the recognised dominant or prominent pitch level of the synthesised replicas of these vowels did not surpass 262 Hz (labelling majority), in strong contrast to the replicas of the sounds of the other vowels. Conversely, vowel confusion generally occurred for replicas of the natural close-mid and open-mid vowel sounds, and, as said, the dominant or prominent pitch level of the replicas of these vowels was mostly recognised as largely above the corresponding levels of the replicas of natural close vowel sounds. It is noteworthy that the occurring vowel quality shifts in an open–close direction related to the shifts of the pitch levels were comparable to the findings reported for *F*-pattern and spectral shape ambiguity and the previous synthesis experiments. The same held true for the nonuniform character of this relation since the effect of a high pitch on vowel recognition was rather limited for sounds of /a/.

Besides these general recognition tendencies, some variation in recognised vowel qualities and pitch levels was also found among speaker groups and individual sounds. This finding indicated a possible impact of the individual harmonic configuration and related HCF a sound synthesis was based on.

For most sounds of the close, close-mid and open-mid vowels, the recognised dominant pitch level was related to *D*1 frequency (or coinciding frequencies of *D*1 and HCF). However, exceptions occurred, and no general tendency was manifest for sounds of the vowel /a/.

No further details were analysed in this study. However, when investigating sounds of this type, different aspects have to be considered for the relation of pitch levels and acoustic measures, e.g. possible octave mismatches of the listeners (see below), the impact of either equal or concurring frequencies of *D*1 and HCF (see the previous chapter), the frequency level of *D*1, and the possible impact of the frequency distance of *D*1 and *D*2 or *D*2 and *D*3. In addition, the sound timbre of synthesised sounds based on only two or three harmonics, as investigated here, may also influence recognition.

Concerning the recognition of secondary vowel qualities, this type of double recognition occurred for 50% of the sounds (36 of 72 sounds, 43 single identifications), in almost all cases associated with a secondary pitch level. Recognition of secondary pitch levels occurred in 83% of the sounds (60 of 72 sounds, 114 single identifications). However, double-vowel recognition and, to a lesser degree, double-pitch recognition strongly depended on the listeners. Thus, the perception and recognition of vowel quality and pitch level seem to relate to a referencing operation which – at least for vowel sounds of the type investigated here – is to some degree an individual operation of a listener: Depending on an individual listener’s attention span or listening focus, simultaneous recognitions of vowel qualities and/or of pitch levels may or may not occur. This indication leads to the question of how to design experiments that provide more uniform results among listeners for double-vowel and double-pitch recognition.

With few exceptions, higher secondary than dominant or prominent pitch levels were found for the replicas with double-vowel and double-pitch recognition. For the replicas of the close vowels, these higher levels were associated with secondary close unrounded or rounded vowel qualities, that is, with no change in openness. For the majority of the replicas of the close-mid vowels, the higher levels were associated with secondary close vowel qualities, that is, with a quality shift in an open–close direction when compared with vowel intention. Likewise, for the replicas of the open-mid vowel, the higher pitch levels were associated with secondary close-mid or close vowel qualities, that is, again, with a quality shift in an open–close direction when compared with vowel intention. In these terms, primary and secondary vowel and pitch recognition corresponded in their shift direction.

As an additional but important finding, according to the listeners' comments, the sound quality of many of the replicas synthesised with the HarmSyn tool was much more natural-like than the synthesised sounds of the previous experiments based on sinusoids. Indeed, the fact that highly recognisable vowel sounds with a natural-like quality can be synthesised based on only two or three extracted harmonics and their dynamic course, as was obtained for the present sample, is remarkable and represents a major gain for future experimental designs. (For sound quality examination, refer to the sound links in Table 1 in the chapter appendix). Notably, synthesised sounds of this type provide evidence that the spectral shape in terms of an estimated spectral envelope, including the fine structure of the vowel spectrum, is by no means a better acoustic representation of vowel quality than the respective *F*-pattern: There is no spectral fine structure in synthesised sounds based on two or three extracted harmonics of a natural reference sound, and no common spectral shape concept accounts for the documented synthesised sounds.

Three further experiences of the listeners performing the recognition tests are also worth noting. For some listeners and some sounds with double-vowel and/or double-pitch recognition, the recognisability and clarity of the two vowels and/or the two pitches were quasi-equal. Then, the selection of the primary and secondary quality or pitch level (forced choice in the test) was somewhat arbitrary. Also, for some sounds, besides an identifiable pitch level below 1 kHz, a higher level or a whistle was perceived, which could not be assigned to a level according to a musical scale. Finally, for some sounds, octave levels were difficult to assess. These aspects must be considered when interpreting the results of experiments of this type and designing future experiments.

The finding that secondary pitch levels were mostly assigned to high levels markedly surpassing primary pitch levels was not further analysed here, as is true for comparing pitch levels of replicas with and without double-vowel recognition. The question of the relation between secondary pitch levels and frequencies of the dominant harmonics used for synthesis (or of HCF), as well as the question of a specific sound timbre created by specific configurations of harmonics that may or may not trigger double-vowel and/or double-pitch recognition, is left open here.

Sounds for which two vowel qualities and two pitch levels are recognised offer paradigmatic cases for experimental exploration of the vowel–pitch relation thesis. In this context, experiments addressing a transition from one to another recognised vowel quality and, in parallel, from one to another pitch level may provide more experimental evidence on the matter. The experiments reported in the following two chapters address this question. (Note in this context that the recognition phenomenon discussed here is not an aspect of concurrent vowel identifications resulting from testing two vowel sounds simultaneously; for an overview of the corresponding experiments and literature, see Smith et al., 2018. Here, double-vowel and double-pitch recognition concern single sounds with specific harmonic spectra.)

Chapter appendix

Table 1. Synthesised sounds related to harmonics at or near spectral peaks of single natural vowel sounds: Sound sample and *D*-patterns investigated. Columns 1–3 = natural reference sounds (V/L = natural reference vowel sounds, their intended and recognised vowel quality, and sound links; for each vowel and in the order of sound listing in the table, a link refers to the pairs of natural reference sounds and their synthesised two- or three-harmonic replicas; SG = speaker group, where m = men, w = women and c = children; fo = f_o intended, in Hz). Columns 4–6 and 7–9 = numbers and frequencies (in Hz) of the selected dominant or prominent harmonics used in synthesis. Column 10 = HCF of the *D*-patterns (in Hz). Note that the intended f_o of the natural reference sounds is given according to the musical C-major scale and that the frequency levels of *D*1, *D*2 and *D*3 and HCF are given in reference to f_o (as multiples of f_o).
[M-06-08-T01]

Table 2. Synthesised sounds related to harmonics at or near spectral peaks of single natural vowel sounds: Results of dominant or prominent vowel quality and pitch level recognition. The table legend is given for the sounds of a given vowel. Abscissa (except for Rows 1 and 2) = three sounds per speaker group and related intended f_o levels of the natural reference sounds; ordinate (except for Rows 1 and 2) = recognised pitch levels of the replicas, as frequency levels or frequency ranges (in Hz; values are given according to the musical C-major scale). Labelling = single vowel identifications of single listeners are given as single characters, separated with a space; single identifications of vowel boundaries are given as combined characters without a space. Rows 1 and 2 = *D*1 and HCF frequencies (in Hz; values are given as multiples of f_o intended, see Table 1). Colour code: Blue = labelling majority of vowel recognition for a synthesised replica accorded to vowel openness of the natural reference sound (unrounded–rounded differences are ignored); red = labelling majority of vowel recognition for a synthesised replica accorded to a vowel quality shift in an open–close direction when compared with vowel intention; dark green = labelling majority of pitch level recognition for a synthesised replica accorded to the frequency level of either *D*1 or HCF or both, if equal; light green = labelling

minority of pitch recognition for a replica accorded to the frequency level of either *D*1 or HCF or both, if equal. Note “ns” for “not specified”.
[M-06-08-T02]

Table 3. Synthesised sounds related to harmonics at or near spectral peaks of single natural vowel sounds: Results of double-vowel (DV) and/or double-pitch (DP) recognition. The table legend is given for the sounds of a given vowel. Abscissa = three sounds per speaker group and related intended f_o levels of the natural reference sounds; ordinate = secondary vowel quality and/or secondary pitch level recognition results (DV&DP = secondary vowel quality and secondary pitch level recognised; DV = only secondary vowel quality recognised; DP = only secondary pitch level recognised). The number of single identifications for a secondary vowel quality and/or a secondary pitch level is given, associated with the related vowel qualities or quality boundaries. Colour code: Blue = secondary vowel quality recognition accorded to the openness of vowel intention of the natural reference sound; red = secondary vowel quality recognition accorded to an open–close vowel quality shift when compared with vowel intention of the natural reference sound; green = secondary pitch level recognition only.
[M-06-08-T03]

Table 4. Synthesised sounds related to harmonics at or near spectral peaks of single natural vowel sounds: Further examination of double-vowel and double-pitch recognition. Columns 1 and 2 = natural reference sounds (V = intended and recognised vowel quality; fo = f_o intended; note that f_o levels were speaker group-related, see Table 1). Column 3 = primary (dominant or prominent) and secondary pitch level and related vowel quality recognition. Column 4–6 = shifts from the primary to the secondary pitch level (1p = only one pitch level recognised, ris = rising, that is, the secondary pitch level is recognised as higher than the prominent or dominant level, fal = falling. Results are given for each identification of a single listener separately. If the results of two listeners are given for a single sound, accordingly, a single f_o level is given. Colour code: Blue = secondary vowel quality recognition accorded to the openness of vowel intention of the natural reference sound; red = secondary vowel quality recognition accorded to an open–close vowel quality shift when compared with vowel intention of the natural reference sound; purple = occurring cases for which secondary vowel quality recognition accorded to vowel intention of the natural reference sound in contrast to primary vowel recognition and despite a simultaneous rise of the secondary pitch level when compared to the primary level, that is, cases of rising pitch levels associated with vowel quality shifts in an close–open direction. ns = not specified.
[M-06-08-T04]

Table 1. Synthesised sounds related to harmonics at or near spectral peaks of single natural vowel sounds: Sound sample and *D*-patterns investigated. [M-06-08-T01]

Sounds (S)			D–patterns							S	D–patterns							S	D–patterns						
V/L	SG	fo	Number			Frequency				V/L	Number			Frequency				V/L	Number			Frequency			
		Hz	D1	D2	D3	Hz	Hz	Hz	HCF		D1	D2	D3	Hz	Hz	Hz	HCF		D1	D2	D3	Hz	Hz	Hz	HCF
		131	2	14	23	262	1834	3013	131		2	15	17	262	1965	2227	131		2	4	–	262	524	–	262
	m	131	2	14	21	262	1834	2751	131		1	11	13	131	1441	1703	131		2	6	–	262	786	–	262
		131	2	19	27	262	2489	3537	131		2	12	14	262	1572	1834	262		2	5	–	262	655	–	131
		220	1	14	18	220	3080	3960	220		1	9	12	220	1980	2640	220		1	4	–	220	880	–	220
i	w	220	1	11	16	220	2420	3520	220	**y**	1	8	10	220	1760	2200	220	**u**	1	4	–	220	880	–	220
↗		220	1	11	14	220	2420	3080	220	↗	1	9	11	220	1980	2420	220	↗	1	3	–	220	660	–	220
		262	1	12	16	262	3144	4192	262		1	8	11	262	2096	2882	262		1	3	–	262	786	–	262
	c	262	1	13	16	262	3406	4192	262		1	7	11	262	1834	2882	262		1	2	–	262	524	–	262
		262	1	12	17	262	3144	4454	262		1	9	12	262	2358	3144	262		1	4	–	262	1048	–	262
		131	3	16	18	393	2096	2358	131		2	12	16	262	1572	2096	262		2	3	–	262	393	–	131
	m	131	3	15	19	393	1965	2489	131		3	11	15	393	1441	1965	131		3	6	–	393	786	–	393
		131	3	13	19	393	1703	2489	131		3	12	17	393	1572	2227	131		3	5	–	393	655	–	131
		220	2	11	14	440	2420	3080	220		2	8	12	440	1760	2640	440		2	4	–	440	880	–	440
e	w	220	2	12	14	440	2640	3080	440	**ø**	2	7	11	440	1540	2420	220	**o**	2	4	–	440	880	–	440
↗		220	2	10	13	440	2200	2860	220	↗	2	9	12	440	1980	2640	220	↗	2	3	–	440	660	–	220
		262	2	12	15	524	3144	3930	262		2	7	12	524	1834	3144	262		2	4	–	524	1048	–	524
	c	262	2	14	16	524	3668	4192	524		2	8	11	524	2096	2882	262		1	2	–	262	524	–	262
		262	2	10	13	524	2620	3406	262		2	8	10	524	2096	2620	524		2	4	–	524	1048	–	524
		131	5	15	19	655	1965	2489	131		5	9	–	655	1179	–	131								
	m	131	4	13	19	524	1703	2489	131		7	9	–	917	1179	–	131								
		131	3	11	15	393	1441	1965	131		5	10	–	655	1310	–	655								
		220	3	9	13	660	1980	2860	220		4	6	–	880	1320	–	440								
ε	w	220	3	11	14	660	2420	3080	220	**a**	3	5	–	660	1100	–	220								
↗		220	4	10	14	880	2200	3080	440	↗	5	6	–	1100	1320	–	220								
		262	3	10	15	786	2620	3930	262		4	6	–	1048	1572	–	524								
	c	262	3	9	13	786	2358	3406	262		3	5	–	786	1310	–	262								
		262	3	8	13	786	2096	3406	262		4	5	–	1048	1310	–	262								

Table 2. Synthesised sounds related to harmonics at or near spectral peaks of single natural vowel sounds: Results of dominant or prominent vowel quality and pitch level recognition. [M-06-08-T02]

D1 (Hz)		262	262	262	220	220	220	262	262	262
HCF (Hz)		131	131	131	220	220	220	262	262	262
Pitch level (Hz)	≥ 523			i	i	i				
	392–440									
	294–330									
	196–262	i i i i y	i i i i y	i i i y	i i i iy	i i i i	i i y y	i i i i iy	i i i u u	i i i i y
	110–131						i			
i		fo=131 Hz men			fo=220 Hz women			fo=262 Hz children		

		262	131	262	220	220	220	262	262	262
		131	131	262	220	220	220	262	262	262
	≥ 523					e				
	392–440									
	294–330									
	196–262	y y y y y		y y y y y	y y y y y	y y y y	y y y y y	y y y y y	iy y y y y	i iy y y y
	110–131		y y y y y							
y		fo=131 Hz men			fo=220 Hz women			fo=262 Hz children		

D1 (Hz)		262	262	262	220	220	220	262	262	262
HCF (Hz)		262	262	131	220	220	220	262	262	262
Pitch level (Hz)	≥ 523							u		
	392–440					u				
	294–330									
	196–262	o u u u u	u u u u u	u u u u	u u u u u	u u u u	u u u u u	u u u u	ou u u u u	yu u u u u
	110–131			u						
u		fo=131 Hz men			fo=220 Hz women			fo=262 Hz children		

Table 2 (continuation). [M-06-08-T02]

D1 (Hz)		393	393	393	440	440	440	524	524	524
HCF (Hz)		131	131	131	220	440	220	262	524	262
Pitch level (Hz)	≥ 523							i i i i ie	i i i i y	i i i y e
	392–440	y y y e e	y y y y	y y y ø	i i i i	i i i i	y y e			
	294–330									
	196–262		y	y	ø	ei	y y			
	110–131									
e		**fo=131 Hz men**			**fo=220 Hz women**			**fo=262 Hz children**		

D1 (Hz)		262	393	393	440	440	440	524	524	524
HCF (Hz)		262	131	131	440	220	220	262	262	524
Pitch level (Hz)	≥ 523							y y ø ø ø	y y y y	y y y y y
	392–440		y y y ø ø	y y y y ø	y y y ø	yø ø ø	y y y yø ø			
	294–330									
	196–262	y y y y y			y	ø ø			ø	
	110–131									
ø		**fo=131 Hz men**			**fo=220 Hz women**			**fo=262 Hz children**		

D1 (Hz)		262	393	393	440	440	440	524	262	524
HCF (Hz)		131	393	131	440	440	220	524	262	524
Pitch level (Hz)	≥ 523			o				o ou ou u u		uo u u u u
	392–440		ɔ o u u u	u u	o u u u u	u u u u u	o o u			
	294–330									
	196–262	o u		o o			o o		ɔ o o o o	
	110–131	o o o								
o		**fo=131 Hz men**			**fo=220 Hz women**			**fo=262 Hz children**		

Table 2 (continuation). [M-06-08-T02]

D1 (Hz)		655	523	393	660	660	880	786	786	786		655	917	655	880	660	1100	1048	786	1048
HCF (Hz)		131	131	131	220	220	440	262	262	262		131	131	655	440	220	220	524	262	262
Pitch level (Hz)	≥ 659	i i y y e			ø ø y	ei e e	ie ø	i e ø	y ø a	yø ø	≥ 659	a	a a a	a ɔ ɔ ɔ u		ɔ ns	ɛ a		a a a	
	523		y y ø ø								523						a	a a a a a		a a a a
	392–440			y y y ø ø			ɛø e	e e	eø ø	y ø ɛ	392–440		ɔ		a aɔ ɔ ɔ o	ɔ				
	294–300				ø y	e ø					294–300	ɔ ɔ o u				o				
	220–262		y				ø				220–262		o			aɔ	a a		a ɔ	a
	131										131									
ɛ		**fo=131 Hz men**			**fo=220 Hz women**			**fo=262 Hz children**			**a**	**fo=131 Hz men**			**fo=220 Hz women**			**fo=262 Hz children**		

Table 3. Synthesised sounds related to harmonics at or near spectral peaks of single natural vowel sounds: Results of double-vowel (DV) and/or double-pitch (DP) recognition. [M-06-08-T03]

DP	2	2	2	1	2	2	1	1	1		3	3	2	2	2	3	2	1	1		1	–	4	2	–	1	1	1	–
DV	–	–	–	–	–	–	–	–	–		–	–	–	–	–	–	–	–	–		–	–	–	–	–	–	–	–	–
DV&DP	1 i	1 y	1 i	1 i	–	2 i i	1 i	1 y	1 i		–	–	–	–	1 i	–	–	1 i	1 i		–	–	–	–	–	–	–	–	–
i	**fo=131 Hz men**			**fo=220 Hz women**			**fo=262 Hz children**			**y**	**fo=131 Hz men**			**fo=220 Hz women**			**fo=262 Hz children**			**u**	**fo=131 Hz men**			**fo=220 Hz women**			**fo=262 Hz children**		

DP	3	3	2	2	–	1	1	1	–		2	2	1	1	2	1	–	1	1		1	–	2	–	–	2	–	–	1
DV	–	–	–	–	–	–	–	–	–		–	1 y	–	–	–	–	1 y	–	–		–	–	–	–	–	–	–	–	–
DV&DP	1 ø	–	1 i	1 y	1 i	2 i y	–	1 i	1 i		–	1 y	1 y	1 y	1 yø	1 ø	1 y	2 y ø	1 ø		–	–	2 u u	–	–	1 u	–	–	–
e	**fo=131 Hz men**			**fo=220 Hz women**			**fo=262 Hz children**			**ø**	**fo=131 Hz men**			**fo=220 Hz women**			**fo=262 Hz children**			**o**	**fo=131 Hz men**			**fo=220 Hz women**			**fo=262 Hz children**		

DP	2	2	1	2	1	2	1	2	3		1	4	–	4	4	3	3	3	3
DV	–	1 y	1 y	–	–	–	–	–	–		–	–	–	–	–	–	–	–	–
DV&DP	–	1 e	1 y	–	1 i	–	–	–	–		1 a	1 o	–	–	–	1 a	–	1 ə	–
ɛ	**fo=131 Hz men**			**fo=220 Hz women**			**fo=262 Hz children**			**a**	**fo=131 Hz men**			**fo=220 Hz women**			**fo=262 Hz children**		

Table 4. Synthesised sounds related to harmonics at or near spectral peaks of single natural vowel sounds: Further examination of double-vowel and double–pitch recognition. Recognised pitch levels are given in Hz. [M-06-08-T04]

Sounds		Primary and secondary recognition	Pitch shift			Sound		Primary and secondary recognition	Pitch shift		
V	fo (Hz)		1p	ris	fal	V	fo (Hz)		1p	ris	fal
i	131	262 to > 698 = y to i		x		ø	131	392 only = ø and y	x		
	131	262 to > 698 = i to y		x				392 to > 698 = ø to y		x	
	131	262 to 523 = y to i		x			131	392 to > 698 = ø to y		x	
	220	220 to > 698 = iy to i		x			220	440 to > 698 = ø to y		x	
	220	220 to 440 = y to i		x			220	440 to > 698 = ø to yø		x	
		220 to > 698 = y to i		x			220	440 to > 698 = y to ø		x	
	262	262 to > 698 = iy to i		x			262	523 to > 698 = ø to y		x	
	262	262 to > 698 = u to y		x				523 only = ø and y	x		
	262	262 to > 698 = y to i		x			262	262 to 523 = ø to y		x	
y	220	659 to > 698 = e to i		x				523 to > 698 = y to ø		x	
	262	262 to 523 = y to i		x			262	523 to > 698 = y to ø		x	
	262	262 to > 695 = y to i		x							
e	131	396 to > 698 = e to ø		x		a	131	659 to > 698 = ɔ to a		x	
	131	196 to > 698 = y to i		x			131	440 to 523 = ɔ to o		x	
	220	220 to > 698 = ø to y		x			220	> 659 to > 659 = ε to a		ns	
	220	220 to > 698 = ei to i		x			262	698 to 349 = a to ə			x
	220	220 to > 698 = y to i		x							
		440 to > 698 = e to y		x							
	262	523 to > 698 = y to i		x							
	262	523 to > 698 = e to i		x							
o	131	196 to 392 = o to u		x							
		196 to 392 = o to u		x							
	220	440 to 659 = o to u		x							
ε	131	523 only = ø and y	x								
		523 to > 659 = ø to e		x							
	131	440 only = ø and y	x								
		440 to > 698 = ø to y		x							
	220	659 to > 698 = e to i		x							

M6.9 Harmonic Synthesis III – Sinewave-Like Replicas Related to Non-Dominant *H*1 and to Harmonics at or Near Spectral Peaks of Single Natural Vowel Sounds, With Gradual Attenuation of *H*1 Causing Double-Vowel and Double-Pitch Recognition

Introduction

In the synthesis experiments based on a few harmonics discussed in the preceding Chapters M6.5 to M6.8, the same tendency of nonuniform open–close vowel quality shifts with increasing pitch level was observed, as was demonstrated in the chapters on formant pattern and spectral shape ambiguity. At the same time, for many cases of these synthesised sounds, some listeners reported that they recognised either two (or even more) pitch levels or two vowel qualities and two pitch levels. In rare cases, they recognised two vowel qualities only.

As indicated in the previous chapter, sounds for which some listeners may recognise two vowel qualities and two pitch levels offer paradigmatic cases for experimental exploration of the vowel–pitch relation thesis. Above all, experiments addressing a transition from one to another recognised vowel quality and, in parallel, from one to another pitch level may provide crucial experimental evidence on the matter. The experiments reported in this and the next chapter address this matter. Based on the above experiences and reflection, the further developed experimental approach focused on the investigation of sound series with transitions from one vowel quality associated with a lower pitch level to another vowel quality associated with a higher pitch level: Starting from one voiced sound with mostly unambiguous single vowel and pitch recognition, is it possible to create a series of sounds by stepwise lowering the level of a specific harmonic or the levels of a series of harmonics so as to, firstly, create sounds with two competing HCFs (HCF of all harmonics and HCF of the harmonics with unaltered levels) and two recognised pitch levels and possibly also two recognised vowel qualities until, secondly, the vowel quality and the pitch level fully shift to a second sound with mostly unambiguous recognition?

In the first study discussed in the present chapter, this question was investigated with regard to synthesised sounds based on non-dominant *H*1 and the two or three dominant harmonics *D*1–*D*2 or *D*1–*D*2–*D*3 (frequencies and levels of the harmonics) in the spectrum of a natural voiced reference sound, HCF of the dominant harmonics being higher than HCF of the entire *H*1–*D*1–*D*2 or *H*1–*D*1–D2–*D*3 pattern.

An earlier version of the study with limited vowel and pitch recognition tests has already been published as an abstract and as online materials, including illustrations (Maurer et al., 2020). For the present treatise, one natural reference sound of /e/ and the related synthesised sounds were replaced, recognition tests were extended, and the results of this renewed investigation are discussed below. (Note that some text excerpts of this earlier online publication have been integrated into the present text.)

General experimental idea and design

In this section, the main experimental idea and design are outlined in general terms before the method of the actual experiment is explained. (For examples illustrating the experimental design, see the sound links in Tables 1 and 3 in the chapter appendix.)

According to the basic experimental idea, as a first step, natural sounds of open-mid and close-mid front vowels have to be selected for which

- the first three spectral peaks, commonly assumed as vowel-related, are represented by dominant harmonics *D*1–*D*2–*D*3;
- D_1–D_2–D_3 are comparable to F_1–F_2–F_3 patterns as commonly estimated for the sounds in question;
- f_o is ≤ 300 Hz;
- D_1 is above the fundamental H_1;
- D_2–D_3 are integer multiples of D_1.

In parallel, natural sounds of open-mid and close-mid back vowels have to be selected for which

- either the lower spectral peaks or indications of prominent spectral energy < 1.5 kHz, commonly assumed as vowel-related, are represented by dominant or prominent harmonics *D*1–*D*2, or the spectral envelope of the lower harmonics shows only one peak associated with a single harmonic *D*1 and the course of the subsequent harmonic envelope is continuously sloping, with *D*2 representing this slope;
- D_1–D_2 are comparable to F_1–F_2 patterns as commonly estimated for the sounds in question;
- f_o is ≤ 300 Hz;
- D_1 is above the fundamental H_1;
- D_2 is an integer multiple of D_1.

Long vowels may prove to be more applicable than short vowels because the impact of sound duration for vowel recognition is weak. Sounds of the vowel /a/ may also be considered (see below).

In general, the experimental design relates to a common assumption that, for some sounds of some front vowels, three spectral peaks and three estimated formants are needed for an acoustic representation of vowel quality (Ladefoged, 2003, p. 105). However, according to the assumption mentioned, two estimated formants suffice for the quality representation of back vowels. Also, for back vowels, the experimental design accounts for weak or absent second lower peaks often occurring in the sound spectra (see Chapter M7.1).

As a second step, the dominant or prominent harmonics and also the low harmonic(s) below *D*1 are to be extracted from the selected natural reference sounds.

Thirdly, harmonic synthesis has to be applied with patterns of dominant or prominent harmonics kept unchanged but with low harmonic(s) below *D*1 stepwise attenuating until this harmonic or these harmonics are deleted.

Fourthly, vowel and pitch recognition have to be tested for all synthesised sounds.

Replicas that are synthesised based on such patterns of dominant or prominent harmonics supplemented with one or more low harmonic(s) have special characteristics. Concerning opposing replicas without and with full attenuation of the harmonic(s) below *D*1, as said, HCF of the pattern including the harmonic(s) below *D*1 is lower than HCF of the pattern excluding the harmonic(s) below *D*1. Therefore, the periodicity of the two opposing replicas differs, affecting pitch perception and recognition and, in association with pitch, possibly affecting vowel recognition. However, the spectral energy maxima are kept unchanged for both replicas, and if formant patterns are estimated, they will also prove to be unchanged. Further, concerning all synthesised replicas of a series with a stepwise attenuation of the level of the lower harmonic(s) below *D*1, HCF and the lower harmonics compete with HCF related to the pattern of dominant or prominent harmonics only, and therefore, cases of double-pitch recognition can be expected. But if double-pitch recognition can be expected, double-vowel recognition may also occur. Finally, the relation between the calculated f_o of the sound wave and the perceived pitch level may dissociate for some of the synthesised sounds.

Concerning the upper f_o limit of 300 Hz set in the present experimental design, this limitation allows for comparing harmonic configurations with estimated *F*-patterns, filter curves and spectral shapes. The methodological substantiation of estimating these features would be substantially impaired if *f*ₒ were increased further. For the present test, for two reasons, we further limited the selection criteria to close-mid vowels and natural sounds for which *f*ₒ was 200–250 Hz and *D*1 corresponded to *H*2: On the one hand, for sounds of these vowels, the lowest spectral peak of the reference sounds often corresponds to the second harmonic of the vowel spectrum, with peak frequencies being in the range of c. 400–500 Hz. Therefore, for the synthesised replicas for which *H*1 is deleted (full attenuation of its level), the frequencies of *D*1 and HCF of the *D*-pattern, and with them the expected pitch level, are twice the frequency of *H*1 and *f*ₒ of the natural reference sound. However, they are still within a pitch frequency range of everyday speech with easily intelligible vowel sounds. On the other hand, for a pitch level shift within this frequency range, associated vowel quality shifts in an open–close direction were shown to be pronounced for sounds of close-mid vowels in many of the previous experiments described.

The subsequent description of the conducted experiment refers to this limitation. (For another experimental setting including open-mid and open vowels and lower *f*ₒ of the natural sounds < 200 Hz, the first manifest dominant harmonic in a sound spectrum may be *H*3 or higher. In these cases, instead of only attenuating *H*1, the levels of all harmonics below *D*1 would have to be attenuated according to the general experimental design, and the observed pitch level difference would increase accordingly; see also the experiment discussed in the following chapter.)

Note that the abbreviation *D*(i) is again used here to avoid confusion about the number of dominant or prominent harmonics and the number of any harmonic *H*(i) in the spectrum of the natural reference sound.

Experiment conducted

Selection of sounds and assignment of harmonic patterns: According to the experimental idea and general design described and based on the Zurich Corpus, for each of the three Standard German vowels /e, ø, o/, two sounds produced by women with medium vocal effort in V context and nonstyle mode at calculated *f*ₒ in the range of 200–250 Hz were selected, fulfilling the following conditions: (i) The spectral peaks generally assumed to relate to vowel quality were associated with single dominant harmonics near the frequencies of estimated formants, formant estimation being methodologically substantiated; for sounds of /o/, however, a prominent spectral energy above a single lower spectral peak and a correspondingly estimated second formant related to only a prominent harmonic were also accepted; (ii) *H*2 represented the first dominant harmonic *D*1; (iii) the higher dominant (or prominent) harmonic(s) were integer multiples of *D*1. As a result, a sample of six natural reference sounds was created. (Sound duration was disregarded.) For all six sounds, the dominant harmonics *D*1–*D*2–*D*3 (sounds of /e, ø/) and the dominant or prominent harmonics *D*1–*D*2 (sounds of /o/) were assigned, and *H*1 was added to a *D*-configuration to create the harmonic patterns for the subsequent sound synthesis (see Columns 1–10 in Table 1 in the chapter appendix).

Extraction of harmonics and subsequent harmonic synthesis: For each single natural reference sound, the dynamic course of its harmonic spectrum for the entire sound duration was analysed using the analysis function of the HarmSyn tool (default parameter setting). Subsequently, based on this analysis and the selected harmonics *H*1–*D*1–*D*2–*D*3 (sounds of the front vowels /e, ø/) or *H*1–*D*1–*D*2 (sounds of the back vowel /o/), eight replicas applying eight attenuation levels for *H*1 of 0/-5/-10/-15/-20/-30/-50/-100 dB were created using the harmonic synthesis function of the HarmSyn tool. As a result, six series of eight replicas with a one-octave transition of dominant HCF from a lower to a higher level were created, and a total of 48 synthesised sounds were investigated in the listening tests.

Illustration of the experimental design: For an illustration of the experimental design, refer to the sound links in Table 1 (natural reference sounds and the related pairs of opposing synthesised sounds without and with full attenuation of *H*1) and Table 3 (natural reference sounds and the related eight synthesised replicas).

Testing vowel and pitch recognition of the opposing replicas (without and with full attenuation of *H*1): Vowel and pitch recognition of the opposing replicas of a series with unchanged *H*1 and with fully deleted *H*1 were tested in four subtests (hereafter S1–S4) involving the five standard listeners of the Zurich Corpus and applying the standard procedure of the corpus, with additional specifications as given below. In these four subtests, the listeners were asked to label only one vowel quality (the dominant or prominent quality recognised) or only one pitch level (the dominant or prominent level recognised) of a sound (exclusion of double-vowel and double-pitch recognition).

In the first subtest, S1, each test item consisted of a single replica and the listeners were asked to assign the dominant or prominent vowel quality (forced choice, all long Standard German vowels and schwa, no vowel boundaries).

In the subtests S2–S4, each test item consisted of the two opposing replicas of a sound series (separated by a 1 sec. pause), the replica with unchanged *H*1 versus the replica with fully deleted *H*1, or vice versa (sound pairs tested in AB and BA order). In subtest S2, the listeners were asked to assign a vowel quality to the second sound presented (forced choice; for vowel qualities, see above). In subtest S3, the listeners were asked to compare the pitch levels of the first and second sounds presented and to label the corresponding level difference as falling, rising or flat. In subtest S4, the listeners were asked to compare both vowels and both pitch levels of the two sounds presented and to assign the two recognised (dominant or prominent) vowel qualities of the two sounds as well as the recognised (dominant or prominent) pitch level difference according to the order of sound presentation.

Testing vowel and pitch recognition of the replicas with attenuated *H*1 of a series (transitional sounds): Vowel and pitch recognition of the replicas of a series with attenuated *H*1 (hereafter transitional sounds) was tested in three subsequent subtests, S5–S7, involving the five standard listeners of the Zurich Corpus and applying the standard procedure of the corpus, with additional specifications as given below.

In subtest S5, each test item consisted of a single synthesised sound with attenuated *H*1 and the listeners were asked to assign the dominant or prominent vowel quality (forced choice; for vowel qualities, see above). In subtests S6 and S7, each test item consisted of two synthesised replicas of a series (separated by a 1 sec. pause), the replica with unchanged *H*1 versus a second replica with attenuated or deleted *H*1 or, inversely, the replica with deleted *H*1 versus a second replica with unchanged or attenuated *H*1. In subtest S6, the listeners were asked to assign a vowel quality to the second sound presented (forced choice; for vowel qualities, see above). In subtest S7, the listeners were asked to compare the pitch levels of the two sounds presented and to label the corresponding pitch difference as falling, rising or flat.

Testing double-vowel and double-pitch recognition of all replicas of a series: Double-vowel and double-pitch recognition of all replicas of a series were tested in two further subtests, S8 and S9, involving the same listeners of the Zurich Corpus and applying the standard procedure of the corpus, with additional test specifications as given below.

In subtest S8, each test item consisted of a single replica presented and the listeners were asked whether they recognised one or two vowel qualities. For the labelling, the listeners used a prepared paper form listing the sounds (in random order according to sound presentation), and they marked the sounds for which they recognised two vowels with “y” (“yes”). No details of vowel qualities were labelled. Likewise, in subtest S9, single replicas were presented, and the listeners were asked whether they recognised one or two pitch levels. For the labelling, the listeners again used a prepared paper form listing the sounds, and they marked the sounds for which they recognised two pitch levels with “y” (“yes”).

Results

Table 1 in the chapter appendix shows the natural reference sounds investigated, the assigned patterns of *H*1 and dominant or prominent harmonics used for synthesis for the opposing replicas of a series, and the results of separately tested vowel and pitch recognition of these replicas (subtests S1–S3). Table 2 shows the results of simultaneously tested vowel and pitch recognition of the opposing replicas (subtest S4). Table 3 (for details, see also Table 3 online) shows the entire series of replicas resulting from *H*1 attenuation and, for these replicas, the results of separately tested vowel and pitch recognition (subtests S1/S5, S2/S6 and S3/S7) and the results of testing double-vowel and double-pitch recognition (subtests S8 and S9). Table 4 shows the numerical distribution of double-vowel and double-pitch recognition per listener (analysis of listener-specific labelling as given in Table 3). Sound links are included in Tables 1 and 3.

Note that, in Table 3, the results of S1 (synthesised sounds without and with full *H*1 level attenuation) are added to the results of S5 (synthesised sounds with stepwise *H*1 level attenuation). Likewise, the results of S2 are added to the results of S6, and the results of S3 are added to the results of S7; these additions concern the missing positions in the listening tests S5–S7 for the replica with unchanged *H*1 and deleted *H*1. The additions allow for a uniform labelling number among all sounds. (For improved readability, for Series 1 and listener L1, the results of S1–S3 added to S5–S7 are marked with “*”; see Table 3 online.)

Vowel and pitch recognition of the opposing sounds: The results of separately testing vowel and pitch recognition for the opposing replicas of a *D*-pattern (without and with full attenuation of *H*1; see Table 1) showed that unattenuated *H*1 was associated with a lower pitch level and a close-mid vowel quality, and full level attenuation of *H*1 (*H*1 deleted) was associated with a higher pitch level and a close vowel quality. The recognition rate of vowel openness (labelling majority) for subtests S1 and S2 was 80–100% (8 to 10 of 10 identifications in total), and recognised upward pitch shifts were uniform among all listeners.

Highly comparable results were obtained for the simultaneous vowel and pitch recognition test (see Table 2), with only five single vowel identifications of two listeners found to differ comparing the results of S1 and S2 with S4 (see Series 2 and 4, the identifications marked with "*"). Again, the recognition rate of vowel openness was 80–100% and recognised upward pitch shifts were uniform among all listeners.

Vowel and pitch recognition of the opposing and the transitional sounds: The results of the vowel and pitch recognition subtests for the transitional replicas of a *D*-pattern with stepwise attenuation of *H*1 from -5 to -50 dB and their comparison with the results of the opposing replicas (see Table 3, results for subtests S1/S5, S2/S6 and S3/S7) showed that, as a tendency, recognition of close-mid vowel qualities and lower pitch levels was maintained mostly for weak *H*1 attenuation up to -10, -15 or -20 dB, and it markedly shifted to close qualities and higher pitch levels for *H*1 attenuation of -30 or -50 dB. At the same time, in contrast to the results for the opposing sounds of a series, marked between-listener recognition differences occurred in the transition from close-mid vowels and lower pitch levels (no attenuation of *H*1) to close vowels and higher pitch levels (full attenuation of *H*1; see Table 3 online).

Double-vowel and double-pitch recognition: The results of testing double-vowel and double-pitch recognition (see Tables 3 and 4) showed that (i) for each of the sound series investigated, sounds occurred for which two vowel qualities and/or two pitch levels were recognised, (ii) double-vowel and double-pitch recognition was most pronounced for *H*1 attenuation in the range of -15 to -30 dB, (iii) double-pitch recognition occurred more often than double-vowel recognition, (iv) double-vowel recognition unparalleled by double-pitch recognition was rare (for the synthesised sounds of front vowels, it occurred only in the identifications of one of the five listeners, L5; for the synthesised sounds of back vowels, it concerned one single identification of listener L2, four identifications of listener L4 and four identifications of listener L5; note in this context that the recognition consistency of listener L5 was lower than that of the other listeners; for these findings, see Table 3 online).

Listener-specific aspects for testing the recognition of the opposing replicas: For the opposing sounds, consensus on vowel recognition proved to be high and recognition of pitch level differences proved to be uniform among listeners. Thus, between-listener differences were marginal.

Listener-specific aspects for testing the recognition of the transitional replicas and for testing double-vowel and double-pitch recognition: In contrast to the results for the opposing sounds, as mentioned, marked between-listener differences occurred for the transitional replicas with stepwise attenuation of *H*1: Testing vowel and pitch recognition of the sound series investigated resulted in distinct recognition profiles of single listeners, with the labelling of the listeners markedly differing in (i) the attenuation levels of *H*1 associated with a vowel quality and/or a pitch level shift (compare e.g. listeners L1 and L3 with listeners L4 and L5), (ii) the labelling consistency (compare e.g. listeners L1 and L3 with listeners L2, L4 and L5), and (iii) the extent (number of occurring cases) of double-vowel and/or double-pitch recognition (compare all listeners). Besides, a few cases of listener-specific vowel differentiation also occurred (see the /e/-/ø/ alterations labelled by listener L2, and constant labelling of /u/ by listener L4 for the sounds of Series 6).

***D*-pattern-specific aspects:** Because of the small sample of natural reference sounds examined, the question of recognition differences, possibly due to the frequencies and levels of the selected dominant or prominent harmonics, was not further analysed.

Comparison of calculated f_o, HCF and recognised pitch level: For all six series, the shift of calculated f_o from a lower to a higher level preceded the recognised pitch level shift indications given by the labelling majority (see Table 3, Columns 10, 17 and 18), that is, calculated f_o and recognised pitch level were somewhat dissociated. Besides, as expected, the sounds with a labelling majority of double-vowel and/or double-pitch recognition were always found for sounds within the range of competing HCFs. (For HCF, lower values are given in the table for the first two sounds of a series, and higher values are given for the last two sounds. For the other four sounds with attenuation of *H*1 of -10/ -15/-20/-30 dB, lower and higher HCF values are given as competing values.)

Discussion

In the present experiment, three dominant harmonics *D*1–*D*2–*D*3 (sounds of /e, ø/) or two dominant or prominent harmonics *D*1–*D*2 (sounds of /o/) of the spectra of natural close-mid vowel sounds produced by women at f_o in the range of c. 220–250 Hz were assigned, *D*1 being *H*2, the higher dominant or prominent harmonics being multiples of *D*1 and all *D*-patterns corresponding to formant frequency patterns *F*1–*F*2–*F*3 or *F*1–*F*2 commonly estimated for the natural reference sounds investigated. Then, the fundamental *H*1 of the reference sound spectrum in question (the harmonic below the first spectral peak) was added to an assigned *D*-pattern. Based on this type of harmonic pattern, the dynamic course of the selected harmonics was extracted from a natural reference sound and, using a harmonic synthesiser, replicas were produced with stepwise attenuation of the levels of *H*1 with the aim of triggering a low-to-high transition of HCF and the recognised pitch level. Finally, vowel and pitch recognition of the sounds was investigated.

In sum, once again, the results strongly supported the vowel–pitch relation hypothesis, here in terms of a pronounced general tendency for a close-mid–close vowel quality shift associated with an increase in recognised pitch level. Comparable to the experiment described in Chapter M6.7, a change in HCF triggered the pitch level shift. However, in the present experiment, the lower HCF level of a sound series was created by using (non-dominant) *H*1 for synthesis. This experimental condition may explain why, in contrast to the experiment of Chapter M6.7 using intermediate harmonics to alter HCF, listeners unanimously recognised a pitch level difference for the opposing sounds without and with full attenuation of *H*1.

With respect to double-vowel and double-pitch recognition, for almost all synthesised sounds, some listeners recognised two vowel qualities and/or two pitch levels. Further, for five of the six series of replicas and their HCF transition from a lower to a higher level, triggered by a stepwise attenuation of *H*1, two vowel qualities and two pitch levels were simultaneously recognised by the majority of listeners for at least one of the transitional sounds. In these terms, because of the investigation of both opposing and transitional replicas, double-vowel and/or double-pitch recognition could be demonstrated in a more pronounced and systematic way than in the previous experiment. As said, this demonstration, in turn, strongly underpins the vowel–pitch relation hypothesis. Indeed, we interpret the phenomenon as such as a core phenomenon of the vowel sound, its acoustic characteristics and its recognition. We will return to this question in the next chapter.

However, vowel quality and pitch level shifts did not obey strict parallelism. Concerning single identifications of single listeners and in relation to attenuation levels of *H*1, pitch level shifts occurred which were either unassociated with vowel quality shifts or preceding or succeeding them (if looked at from the perspective of increasing *H*1 level attenuation within a sound series). Vice versa, but only rarely, some vowel quality shifts did occur without a simultaneous pitch level shift. Further research is needed to interpret such cases of double-pitch without double-vowel recognition and, inversely, double-vowel without double-pitch recognition. Thereby, the two phenomena may not prove to be directly comparable. Notably, double-pitch without double-vowel recognition occurred in numerous cases and for all listeners, but double-vowel without double-pitch recognition was rare and was not found for all listeners. Besides, some of the labelling variations may have to be expected for the types of sounds investigated and recognition tests conducted (including inconsistent identifications of a listener), whereas others may point to an actual and reproducible recognition strategy of a listener. Also, when interpreting these results, attention must be given to the test procedure applied, according to which vowel qualities and pitch levels were tested separately. Nevertheless, the question of why (numerous) cases of double-pitch without double-vowel recognition and, inversely, (rare) cases of double-vowel without double-pitch recognition occurred is posed, and with it the question of whether these cases are an indication of a process of perception and recognition not imperatively relating vowel quality to a consciously identified and verbally assigned pitch level but, as an alternative, to a comparable perceptual referencing to a sound pattern repetition over time. We address these questions in Chapter 6.11 (see Part II).

As described in the results section, marked between-listener differences occurred in the recognition results for the transitional sounds and for double-vowel and double-pitch recognition (see Table 3 online, details of recognition results; see also Table 4). These differences may be the consequences of several aspects. Above all, listeners possibly differed to some degree in their demarcation of two vowel qualities and their perceived “threshold” of recognising a pitch level shift. Furthermore, the notion of perception as a selective (and therefore listener-specific) process with the focus lying on a specific vibration pattern of a sound, or change of that focus, also has to be taken into account. Furthermore, the order of the stimuli (single sounds or sound pairs, AB or BA order of sound pairs) influenced the recognition results. Finally, the type of sounds investigated and their specific timbre may also have impacted the results, especially when it comes to transitional sounds.

The sounds presented here again highlight the fact that neither the *F*-pattern nor the spectral envelope *per se* acoustically represents vowel quality: Spectral maxima and estimated *F*-patterns did not change for the synthesised sounds investigated in the above series despite occurring vowel quality shifts, and current concepts of spectral envelope estimation do not account for the harmonic configurations of the vowel spectra of these sounds. As a consequence, the spectrograms of the two opposing sounds of a *D*-pattern without and with full attenuation of *H*1, recognised as two different vowel qualities, do not reflect this quality difference in terms of pronounced differences of spectral energy maxima or pronounced differences in an estimated spectral envelope (for illustration, refer to the sound links in Table 1 and compare the spectra, LPC filter curves and spectrograms of the opposing sounds; see also Part II, Chapter 6.9, Figure 4).

With respect to the comparison of calculated f_o, HCF and recognised pitch level for sounds with stepwise attenuation of *H*1, according to the labelling majority of the listeners, upward shifts of calculated f_o preceded the upward pitch level shifts for all sound series possibly associated with ambiguous HCF. Thus, f_o, HCF and pitch have to be clearly distinguished from each other when investigating vowel sounds. (Notably, existing normalisation concepts generally do not account for this differentiation.)

However, some relativisations also have to be made: (i) The number of sounds investigated was again small, and the range of HCF investigated was limited; (ii) the sounds selected had to fulfil very specific conditions concerning their spectral characteristics; (iii) the sound sample presented in the listening test was biased (the sample did not include sounds of many different vowels and was not balanced with regard to vowel qualities and pitch levels); (iv) no simultaneous labelling of double-vowel and double-pitch was tested. Moreover, the sound quality was still somewhat impaired because of the few harmonics the synthesis was based on. Therefore, no detailed prediction is made here for vowel and pitch recognition of other sounds which do not correspond to the selection criteria applied here. Above all, synthesised sound series related to other harmonic configurations may confirm parallel vowel quality and pitch shifts, while others may only be associated with pitch shifts. However, if vowel quality indeed relates to pitch, no pronounced inverse vowel quality shifts in an close–open direction with rising pitch should occur for *D*-patterns of the type described here, given that the assignment of dominant harmonics is methodologically substantiated.

In this context, for future research, special attention should be given to the many indications of the nonuniform relation between spectral characteristics and recognised vowel quality. When designing the present experiment, we have again experienced that the sound quality of this type of synthesis depended on the vowel quality and f_o of the natural reference sounds and that this dependency affects the inter-listener accordance or confusion of both vowel and pitch recognition. For similar experiments, we recommend first investigating natural reference sounds of close-mid vowels produced at f_o of 200–250 Hz because, as mentioned in the experiment section of this chapter, their lower vowel spectrum was found to strongly relate to an increase in f_o and, as a consequence, single formant patterns as well as single spectral envelopes of these sounds are in most cases ambiguous in that they represent close-mid and close vowel qualities (for sounds with different f_o and pitch levels). Subsequently, the investigation may be extended to sounds of other vowel qualities, above all to sounds produced at lower f_o. The investigation of sounds produced at higher f_o above 300 Hz poses methodological problems in defining and assigning dominant harmonics in the natural reference sounds and estimating *F*-patterns of these reference sounds.

Furthermore, concerning future research, the manipulation of one or two selected harmonics of an assigned *D*-pattern in terms of an adjustment of their level(s) for synthesis may be included in the investigation: It may be difficult to configure larger samples of natural reference sounds sufficiently fulfilling the experimental conditions, and adjustments of harmonic levels may help to achieve the required dominant or prominent harmonics and their interrelation in a spectrum. At the same time, the role of the harmonic levels of a *D*-pattern can also be investigated, above all, for sounds of back vowels. (On this matter, see also the next chapter.)

However, the present experiment may serve as a model experiment: The constancy of the spectral maxima of the synthesised sounds of a series and the occurring vowel and pitch shifts as a result of only attenuating a non-prominent *H*1 level can barely be understood and explained outside the framework of a vowel–pitch relation hypothesis (or its alternative).

Chapter appendix

Table 1. Synthesised sound series related to non-dominant *H*1 and harmonics at or near spectral peaks of single natural vowel sounds produced by women, including stepwise *H*1 attenuation: *H*1 and *D*-patterns investigated and results of separate vowel and pitch recognition tests for the opposing sounds of a series (recognition subtests S1–S3). Columns 1–3 = natural reference sounds (S/L = sound series and sound links; note that a given sound link refers to a natural reference sound and the two related opposing synthesised sounds without and with full attenuation of *H*1; V = intended and recognised vowel quality; fo r = calculated f_o levels, in Hz). Columns 4–10 = harmonic configurations used for synthesis (H1 = frequency of the first harmonic *H*1 of the natural reference sound related to its f_o level; AH1 = attenuation of the level of *H*1, in dB; D(i) = dominant or prominent harmonics, in Hz; HCF = HCF of the harmonics used for synthesis, approximate values, in Hz; fo s = calculated f_o for the synthesised sounds, in Hz). Columns 11–16 = recognised dominant or prominent vowel qualities (confusion matrix; summary of subtests S1 and S2, with ten identifications per sound in total). Columns 17–18 = recognised dominant or prominent pitch levels (l = lower, h = higher; summary of subtest S3, with five identifications per sound in total). Columns 19–33 = listener-specific details of the listening tests. Colour code: Dark blue = recognition rates of ≥ 80% for close-mid vowel qualities associated with a lower pitch level; dark red = recognition rates of ≥ 80% for close vowels associated with a higher pitch level; light blue = single identifications of a close-mid vowel quality and/or of a lower pitch level; light red = single identifications of a close vowel quality and/or of a higher pitch level.
[M-06-09-T01]

Table 2. Synthesised sound series related to non-dominant *H*1 and harmonics at or near spectral peaks of single natural vowel sounds produced by women, including stepwise *H*1 attenuation: Results of simultaneous vowel and pitch recognition for the opposing sounds of a series (recognition subtest S4). For columns and colour code, see Table 1. (Note that sound presentation was in AB and BA order, with ten identifications per sound in total.) In addition, labelling differences in comparison with the results of Table 1 are marked with "*", and the occurring cases of listener-specific recognitions of a pitch shift without a vowel quality shift are marked in grey.
[M-06-09-T02]

Table 3. Synthesised sound series related to non-dominant *H*1 and harmonics at or near spectral peaks of single natural vowel sounds produced by women, including stepwise *H*1 attenuation: Vowel and pitch recognition results for all sounds of a series (all recognition subtests). For Columns 1–18, see Table 1. A given sound link refers to a natural reference sound and all eight related synthesised sounds. Vowel recognition is given as the sum of subtests S1/S5 and S2/S6, with 15 identifications per sound in total. Pitch recognition is given as the sum of subtests S3/S7, with ten identifications per sound in total. In Column 9, competing HCFs are given for the sounds with *H*1 level attenuation of -10 to -30 dB. Columns 19–20 = cases of two vowel qualities and/or two pitch levels simultaneously recognised for a single sound (double-vowel recognition, DV, and/or double-pitch recognition, DP), marked with "y" for "yes" (summaries of subtests S8 and S9, with five identifications per sound in total for each of the subtests). Extended online table: Columns 21 ff. = listener-specific details of the recognition results. To improve readability and give an example, the results of subtests S1–S3 (added to the results of subtests S5–S7) are marked with "*" for Series 1 and listener L1. Colour code, see Table 1. In addition, in Columns 19 and 20, the most pronounced parallel double-vowel and double-pitch recognition found among the listeners for one or two synthesised sounds of a series are marked in dark green. In Columns 21 ff. (extended online table), parallel double-vowel and double-pitch recognition, as well as double-pitch recognition only, are marked in light green, and double-vowel recognition only is marked in grey.
[M-06-09-T03]

Table 4. Synthesised sound series related to non-dominant *H*1 and harmonics at or near spectral peaks of single natural vowel sounds, including stepwise *H*1 attenuation: Sounds per listener with double-vowel and/or double-pitch recognition. Column 1 = listeners. Column 2 = number of sounds with parallel double-vowel and double-pitch recognition (DV&DP). Column 3 = number of sounds with double-pitch recognition only (DP). Column 4 = number of sounds with double-vowel recognition only (DV).
[M-06-09-T04]

Table 1. Synthesised sound series related to nondominant H1 and harmonics at or near spectral peaks of single natural vowel sounds produced by women, including stepwise H1 attenuation: H1 and D-patterns investigated and results of separate vowel and pitch recognition tests for the opposing sounds of a series (recognition subtests S1–S3). [M-06-09-T01]

Sounds											Recognition								Recognition (details; subtests S1–S3)														
References				Synthesis							Vowel						Pitch		L1			L2			L3			L4			L5		
S/L		V	fo r	H1	AH1	D1	D2	D3	HCF	fo s	close-mid			close			level		S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3
			Hz	Hz	dB	Hz	Hz	Hz	Hz	Hz	e	ø	o	i	y	u	l	h	V	V	P	V	V	P	V	V	P	V	V	P	V	V	P
1	↗	e	218	218	0	436	2616	3052	220	218	10						5		e	e	l	e	e	l	e	e	l	e	e	l	e	e	l
				–	-100				440	436				10				5	i	i	h	i	i	h	i	i	h	i	i	h	i	i	h
2	↗	e	246	246	0	492	2460	2952	250	246	9	1					5		e	e	l	e	ø	l	e	e	l	e	e	l	e	e	l
				–	-100				500	492	1			9				5	i	i	h	i	i	h	i	i	h	i	i	h	e	i	h
3	↗	ø	221	221	0	442	1768	2652	220	221		10					5		ø	ø	l	ø	ø	l	ø	ø	l	ø	ø	l	ø	ø	l
				–	-100				440	442					10			5	y	y	h	y	y	h	y	y	h	y	y	h	y	y	h
4	↗	ø	221	221	0	424	1768	2652	220	221		10					5		ø	ø	l	ø	ø	l	ø	ø	l	ø	ø	l	ø	ø	l
				–	-100				440	442		1			9			5	y	y	h	y	y	h	y	y	h	y	y	h	ø	y	h
5	↗	o	223	223	0	446	892	–	220	223			10				5		o	o	l	o	o	l	o	o	l	o	o	l	o	o	l
				–	-100				440	446						10		5	u	u	h	u	u	h	u	u	h	u	u	h	u	u	h
6	↗	o	215	215	0	430	860	–	220	215			8			2	5		o	o	l	o	o	l	o	o	l	u	u	l	o	o	l
				–	-100				440	429						10		5	u	u	h	u	u	h	u	u	h	u	u	h	u	u	h

Table 2. Synthesised sound series related to nondominant H1 and harmonics at or near spectral peaks of single natural vowel sounds produced by women, including stepwise H1 attenuation: H1 and D-patterns investigated and results of simultaneous vowel and pitch recognition for the opposing sounds of a series (recognition subtest S4). [M-06-09-T02]

Sounds										Recognition								Recognition (details; subtest S4)									
References			Synthesis							Vowel						Pitch		L1		L2		L3		L4		L5	
S/L	V	fo r	H1	AH1	D1	D2	D3	HCF	fo s	close-mid			close			level		S4		S4		S4		S4		S4	
		Hz	Hz	dB	Hz	Hz	Hz	Hz	Hz	e	ø	o	i	y	u	l	h	AB	BA	AB	BA	AB	BA	AB	BA	AB	BA
1	e	218	218	0	436	2616	3052	220	218	10						5		e l	e l	e l	e l	e l	e l	e l	e l	e l	e l
			–	-100				440	436				10				5	i h	i h	i h	i h	i h	i h	i h	i h	i h	i h
2	e	246	246	0	492	2460	2952	250	246	8	2					5		e l	e l	ø* l	ø l	e l	e l	e l	e l	e l	e l
			–	-100				500	492	2			6	2			5	i h	i h	y* h	y* h	i h	i h	i h	i h	e h	e* h
3	ø	221	221	0	442	1768	2652	220	221		10					5		ø l	ø l	ø l	ø l	ø l	ø l	ø l	ø l	ø l	ø l
			–	-100				440	442					10			5	y h	y h	y h	y h	y h	y h	y h	y h	y h	y h
4	ø	221	221	0	424	1768	2652	220	221		10					5		ø l	ø l	ø l	ø l	ø l	ø l	ø l	ø l	ø l	ø l
			–	-100				440	442					10			5	y h	y h	y h	y h	y h	y h	y h	y h	y* h	y h
5	o	223	223	0	446	892	–	220	223			10				5		o l	o l	o l	o l	o l	o l	o l	o l	o l	o l
			–	-100				440	446						10		5	u h	u h	u h	u h	u h	u h	u h	u h	u h	u h
6	o	215	215	0	430	860	–	220	215			8			2	5		o l	o l	o l	o l	o l	o l	u l	u l	o l	o l
			–	-100				440	429						10		5	u h	u h	u h	u h	u h	u h	u h	u h	u h	u h

Table 3. Synthesised sound series related to nondominant H1 and harmonics at or near spectral peaks of single natural vowel sounds produced by women, including stepwise H1 attenuation: Vowel and pitch recognition results for all sounds of a series (all recognition subtests). [M-06-09-T03]. Extended online table: ↗

Sounds										Recognition									
References			Synthesis							Vowel						Pitch			
S/L	V	fo r	H1	AH1	D1	D2	D3	HCF	fo s	close-mid			close			single		double	
		Hz	Hz	dB	Hz	Hz	Hz	Hz	Hz	e	ø	o	i	y	u	l	h	DV	DP
1 ↗	e	218	218	0	436	2616	3052	220	218	13			2			10		3	4
			(218)	-5				220	218	10			5			9	1	3	4
			(218)	-10				220 vs 440	437	11			4			9	1	4	4
			(218)	-15					437	6			8	1		8	2	4	5
			(218)	-20					437	5			10			7	3	4	5
			(218)	-30					437				15				10	1	2
			(218)	-50				440	437				15				10	1	1
			–	-100				440	436				15				10		1
2 ↗	e	246	246	0	492	2460	2952	250	246	13	1		1			10		3	4
			(246)	-5				250	246	13	2					10		3	4
			(246)	-10				250 vs 500	246	13	1		1			10		2	5
			(246)	-15					246	13			2			7	3	3	5
			(246)	-20					492	9			6			6	4	4	5
			(246)	-30					492	9			4	2		6	4	4	5
			(246)	-50				500	492	3			12				10	2	2
			–	-100				500	492	3			12				10	2	1
3 ↗	ø	221	221	0	442	1768	2652	220	221		15					10		2	3
			(221)	-5				220	221		15					10		1	4
			(221)	-10				220 vs 440	221		14			1		10		2	4
			(221)	-15					221		13			2		9	1	2	3
			(221)	-20					442		11			4		8	2	2	5
			(221)	-30					442		8			7		5	5	4	5
			(221)	-50				440	442					15			10	2	1
			–	-100				440	442					15			10		1
4 ↗	ø	221	221	0	442	1768	2652	220	221		15					10		3	4
			(221)	-5				220	221		15					10		2	4
			(221)	-10				220 vs 440	221		14			1		10		2	4
			(221)	-15					442		11			4		8	2	2	4
			(221)	-20					442		9			6		6	4	3	4
			(221)	-30					442		7			8		5	5	3	5
			(221)	-50				440	442		1			14			10	1	2
			–	-100				440	442		1			14			10	1	1

Table 3 (continuation). [M-06-09-T03].

Sounds										Recognition									
References			Synthesis							Vowel						Pitch			
S / L	V	fo r	H1	AH1	D1	D2	D3	HCF	fo s	close-mid			close			single		double	
		Hz	Hz	dB	Hz	Hz	Hz	Hz	Hz	e	ø	o	i	y	u	l	h	DV	DP
5	o	223	223	0	446	892	–	220	223			15				10		2	1
			(223)	-5				220	223			14			1	10		2	1
			(223)	-10				220 vs 440	446			12			3	9	1	3	2
			(223)	-15					446			10			5	7	3	2	3
			(223)	-20					446			7			8	6	4	2	4
			(223)	-30					446			6			9	4	6	2	5
			(223)	-50				440	446						15		10	1	
			–	-100				440	446						15		10	1	
6	o	215	215	0	430	860	–	220	215			12			3	10		1	
			(215)	-5				220	215			12			3	10		1	
			(215)	-10				220 vs 440	215			12			3	9	1	2	2
			(215)	-15					215			10			5	9	1	2	2
			(215)	-20					429			8			7	8	2	3	4
			(215)	-30					429			6			9	6	4	3	5
			(215)	-50				440	429						15		10	1	
			–	-100				440	429						15		10		

Table 4. Synthesised sound series related to nondominant H1 and harmonics at or near spectral peaks of single natural vowel sounds produced by women, including stepwise H1 attenuation: Sounds per listener with double-vowel and/or double-pitch recognition. [M-06-09-T04]

Listeners	DV&DP	DP	DV
L1	8	3	0
L2	13	16	1
L3	15	21	0
L4	27	4	4
L5	25	8	10

M6.10 Harmonic Synthesis IV – Replicas Related to Harmonics at or Near Spectral Peaks of Single Natural Vowel Sounds, With Gradual Attenuation of Selected Intermediate Harmonics Causing Double-Vowel and Double-Pitch Recognition

Introduction

Pursuing the investigation of synthesised sound series with vowel quality shifts related to pitch level shifts, and pursuing harmonic synthesis related to natural reference sounds keeping vowel-related spectral maxima unchanged, we further developed the experimental design of the previous experiment and conducted a new study.

Development of the experimental design

With respect to the experimental design, sounds with the same spectral peak structure as described in the previous chapter were investigated, but in contrast to the previous experiment, the entire harmonic spectrum of natural reference sounds was manipulated in order to create sound transitions: The harmonic analysis and subsequent synthesis were conducted producing, firstly, a synthesised replica related to the entire calculated harmonic spectrum of a selected natural reference sound (harmonic resynthesis), secondly, a series of synthesised sounds with stepwise attenuation of the harmonics lying in between the multiple integers of the first dominant harmonic *D*1 and, thirdly, a synthesised sound based on only the harmonics as multiple integers of *D*1.

As a further part of the development of the experimental design, speaker groups, vowel qualities and the f_o range of the reference sounds were extended, *D*1 was not limited to *H*2, editing of the calculated harmonic spectra of the natural reference sounds was applied, and one of the listening tests described in the previous chapter was also extended: Sounds of open-mid and close-mid vowels produced by speakers different in age and gender in an f_o range of c. 100–300 Hz were investigated; depending on the spectra of the reference sounds, *D*1 was assigned as *H*2, *H*3 or *H*4; if necessary, the harmonic spectra of the sounds were edited (adjustment of harmonic levels) to fulfil the spectral specificity that was searched for; additional natural sounds were added to the synthesised sounds of the first listening subtest to balance vowel qualities and pitch ranges of the sounds in the task. Details are given in the following paragraph.

Experiment conducted

Selection of sounds, assignment of dominant or prominent harmonics, adjustment of single harmonic levels (spectral editing): On the basis of the Zurich Corpus, for each of the five Standard German vowels /e, ø, o/ and /ɛ, ɔ/ and each of the speaker groups of men, women and children, a natural reference sound produced in V context and nonstyle mode at calculated f_o in the range of c. 100–300 Hz was selected (vocal effort was disregarded). Sound selection further accorded to the conditions as set by the general experimental design described in the previous chapter, with the additional option of spectral editing:

- For sounds of front vowels, either the three spectral peaks generally assumed to relate to vowel quality were associated with single dominant harmonics *D*1–*D*2–*D*3 near the frequencies of estimated formants, *D*1 being *H*2 or a higher harmonic and the frequencies of *D*2–*D*3 being integer multiples of *D*1 frequency, or this condition could be generated with a very limited adjustment in terms of attenuating or amplifying the level of single harmonics for the harmonic synthesis of the replicas.
- For back vowels, sounds with two different types of spectra were included; for the first type, either the two spectral peaks or a low spectral peak and an additional subsequent spectral enhancement < 1.5 kHz were associated with single dominant or prominent harmonics *D*1–*D*2 near the frequencies of estimated formants *F*1–*F*2, *D*1 being *H*2 or a higher harmonic and *D*2 frequency being an integer multiple of *D*1 frequency, or this condition could be generated with a very limited adjustment of single harmonics; for the second type, either only one spectral peak was manifest (*D*1 being *H*2 or a higher harmonic) with the remainder of the upper harmonic envelope showing a continuous slope (for numerous examples of sounds of this type, see Chapter M7.1), or this condition could be generated with a very limited level adjustment of single harmonics (see above).

Although the vowel /ɔ/ is short in Standard German, for three reasons, sounds of that vowel were included in the investigation: (i) /ɔ/ is an open-mid vowel, and the experiment aimed at also producing open-mid–close vowel quality shifts, that is, shifts exceeding adjacent qualities; (ii) the vowel quality distance of the long Standard German vowels /o/ and /a/ is pronounced and when creating the Zurich Corpus, therefore, we also recorded sustained sounds of /ɔ/ produced by some of the speakers investigated (see Chapter 1.1); (iii) the recognition of /ɔ/ as a quality in between /o/ and /a/ was not difficult to develop for the standard listeners of the Zurich Corpus even though the sound duration was long and corresponded to the duration of long vowels.

As mentioned, if necessary, single harmonic levels (limited to very few harmonics of a spectrum) of a natural reference sound were adjusted for synthesis in terms of assigning dB values for attenuating or amplifying these levels in synthesis, with the aim of creating sounds for which the harmonic spectra represented exemplary cases for the experimental conditions in question (for details, see Table 1, extended online version; for comparison of the natural reference sounds and the synthesised replicas with adjusted harmonic levels but without attenuation of the harmonics in between the integer multiples of *D*1, see the sound links in this table).

As a result of the selection process, a sample of 15 natural reference sounds was created, and their *D*1–*D*2–*D*3 or *D*1–*D*2 patterns (frequencies and levels) and, if necessary, the dB values for the adjustments of single levels of the dominant harmonics were assigned.

Extraction of harmonics and subsequent harmonic synthesis: For each single natural reference sound, the dynamic course of its harmonic spectrum (frequencies and levels) for the entire sound duration was analysed using the analysis function of the HarmSyn tool (default parameter setting). Subsequently, based on this analysis and the assigned *D*-pattern and, if necessary, including level adjustments of single harmonics, five replicas applying five attenuation levels of 0/-12/-24/-36/-100 dB for the harmonics that are not integer multiples of the *D*1 frequency were created using the harmonic synthesis function of the HarmSyn tool. Harmonics that are integer multiples of *D*1 frequency were kept unchanged. Thus, for each natural reference sound, a series of synthesised replicas was created, with the first and last sound representing two opposing sounds with two different H_1 and HCF (unchanged harmonics of the reference sound versus multiples of *D*1 only) and with three transitional sounds in between (harmonics as integer multiples of *D*1 unattenuated, all other harmonics stepwise attenuated). In these terms, 15 series of five sounds each were created, and a total of 75 synthesised sounds were investigated in the listening tests (see Table 1).

Note on terminology: For the present study, no differentiation of the terms resynthesis and synthesis is made in the text, and all sounds produced with the harmonic synthesiser are termed synthesised sounds. (However, according to the terminological differentiation of resynthesis and synthesis made in the Introduction, the replicas with no attenuation of harmonic levels and no level adjustments represented resynthesised sounds.)

Testing vowel and pitch recognition: The same nine subtests, S1–S9, described in the previous chapter were also conducted for the sounds of the present experiment: Vowel and pitch recognition of the opposing sounds of a series, presented as single sounds or as sound pairs (subtests S1–S4), vowel and pitch recognition of all sounds of a series, presented as single sounds or as sound pairs (subtests S5–S7 combined with the results of subtests S1–S3), and double-vowel and double-pitch recognition of all sounds of a series, presented as single sounds (subtests S8 and S9).

The first subtest, S1, however, was further developed: If only the synthesised sounds of the described sample were presented in a listening test and if, with increasing pitch, vowel quality shifts in an open–close direction were found, this could be interpreted as partly due to a sound presentation bias in that no sounds of open-mid and close-mid vowels were presented for higher f_o or pitch levels, and no sounds of close vowels and only a few sounds of close-mid vowels were presented for lower f_o and pitch levels (see also the discussion section in the previous chapter). Therefore, for each investigated pair of opposing synthesised sounds and of expected open–close shifts due to an increase in pitch, an additional sound pair was added with inverse relations of expected vowel qualities and f_o and pitch levels. (Note that the speaker-group relation could not be fully balanced for the additional natural sounds.) Thus, the investigated 30 opposing synthesised sounds and 30 additional natural sounds were presented in random order in subtest S1. Table 1 in the chapter appendix lists the additional sounds (see the extended online version of the table).

For subtests S8 and S9, the online screen of the listening test was adjusted. The listeners performed the two tests online by selecting the button "yes" ("y" as a result) for double-vowel or double-pitch recognition and the button "no" ("n" as a result) for single vowel or single pitch recognition.

Illustration of the experimental design: For an illustration of the experimental design, see the sound series and corresponding sound links in the chapter appendix: In Table 1, each link presents the natural reference sound and the opposing synthesised replicas without and with full attenuation of the harmonics in between the integer multiples of *D*1. If level adjustments of single harmonics of the reference sounds were made, they are listed in Column 36 (for these cases, compare the natural reference sound with the first synthesised replica). In Table 3, each link presents the natural reference sound and the five synthesised sounds of a series, that is, the sound without and with stepwise attenuation of the harmonics between the integer multiples of *D*1.

Conceptual note: Similarly to the previous experiment, replicas synthesised based on these kinds of harmonic patterns have special characteristics. Because of the paradigmatic character of the experiment, these characteristics are repeated in the following paragraphs.

Concerning the opposing replicas without and with full attenuation of the harmonics that are not integer multiples of *D*1 frequency, the recognised pitch of the two sounds can be expected to be different because of the unambiguous difference of HCF. Furthermore, the recognition of different pitches may in turn affect vowel recognition in terms of a vowel quality shift, even if the vowel-related spectral energy maxima are kept unchanged for both sounds. (Note that if formant patterns are estimated, they will also prove to be unchanged for the two sounds of comparison.) Concerning all synthesised replicas of a series in between the two opposed sounds, that is, concerning the transitional sounds with stepwise attenuation of the intermediate harmonics that are not integer multiples of *D*1, the lower HCF related to *H*1 frequency competes with the higher HCF related to *D*1 frequency and, therefore, cases of double-pitch recognition can be expected. But if two pitch levels are recognised for a single sound, this may again also affect vowel recognition in terms of two vowel qualities being simultaneously recognised, related to the two pitch levels. This attempt to produce sounds with double-pitch and double-vowel recognition is at the core of the experimental purpose and design. Besides, for the transitional sounds, the relation between the calculated f_o and the recognised pitch level of a sound is again in question because they may dissociate.

Comparable to the previous experiment, the rationale to investigate replicas of natural reference sounds of open-mid and close-mid vowels produced in the f_o range of c. 100–300 Hz is as follows: For sounds of open-mid vowels, depending on f_o, their lowest spectral peak often corresponds to the fourth or third harmonic of the vowel spectrum, with estimated spectral peak or formant frequencies being in the range of c. 450–750 Hz (dependent on f_o of sound production). For sounds of close-mid vowels, depending on f_o, their lowest spectral peak often corresponds to the third or second harmonic of the vowel spectrum, with peak frequencies being in the range of c. 300–600 Hz. In consequence, the frequencies of *D*1 and HCF of the replicas with deleted intermediate harmonics that are not integer multiples of *D*1 are four, three or two times the frequency of *H*1 and f_o of the natural reference sound in question, but they are still within the frequency range of f_o or HCF – and expected pitch level – of recognisable vowel sounds. Yet, for a pitch level shift within this frequency range, vowel quality shifts in an open–close direction were shown to be pronounced for sounds of close-mid vowels in many of the previous experiments described, and these shifts also occurred for sounds of open-mid vowels.

Note again that, concerning the set upper f_o limit of 300 Hz, this limitation allows for a comparison of harmonic configurations with estimated *F*-patterns, filter curves and spectral shapes: The methodological substantiation of the estimation of these features would be substantially impaired if f_o were increased further.

Results

Table 1 in the chapter appendix shows the natural reference sounds investigated, the assigned patterns of dominant or prominent harmonics (including the adjusted harmonic levels used for synthesis) for the opposing replicas of a series and the results of separately tested vowel and pitch recognition (subtests S1–S3). Table 2 shows the results of simultaneously testing vowel and pitch recognition of the opposing replicas (subtest S4). Table 3 shows the entire series of replicas resulting from the attenuation of harmonics that are not integer multiples of *D*1 and, for these replicas, the results of separately tested vowel and pitch recognition of these replicas (subtests S1/S5, S2/S6 and S3/S7) and the results of testing double-vowel and double-pitch recognition (subtests S8 and S9). Table 4 shows the numerical distribution of double-vowel and double-pitch recognition per listener (analysis of listener-specific results as given in Table 3). Sound links are included in Tables 1 and 3.

As was the case for the experiment described in the previous chapter, in Table 3, some of the results of S1, S2 and S3 are added to the results of S5, S6, and S7 (for details, see Chapter M6.9, the results section and Table 3).

Vowel and pitch recognition of the opposing sounds: The results of separate vowel and pitch recognition tests for the opposing sounds of a series without and with full attenuation of the harmonics in between the integer multiples of *D*1 (see Table 1, the sum of subtests S1 and S2) showed that, with few exceptions of single vowel quality identifications of single listeners, no attenuation of the harmonics in between the integer multiples of *D*1 was associated with a lower pitch level and with an open-mid or close-mid vowel quality in correspondence to the natural reference sound, and full attenuation of the harmonics in between the integer multiples of *D*1 was associated with a higher pitch level and a vowel quality shift in an open–close direction when compared with the natural reference sound.

Concerning the replicas without attenuation of harmonic levels, the recognition rate according to vowel intention of the natural reference sounds was 100% for 13 series and 80% for two series. Concerning the replicas with full attenuation of the harmonics in between the integer multiples of *D*1 of Series 1–9 (close-mid reference vowel sounds), ignoring unrounded–rounded differences, the recognition rate for vowel quality shifts in an open–close direction was 100% for eight series and 90% for one series. Concerning the replicas with full attenuation of the harmonics between the integer multiples of *D*1 of Series 10–15 (open-mid reference vowel sounds), ignoring unrounded–rounded differences, the recognition rate for vowel quality shifts in an open–close direction was 100% for five series and 80% for one series (note also the occurring single back–front confusion for this series). Open-mid–close shifts prevailed in Series 13 and 14, and open-mid–close-mid shifts prevailed in Series 11 and 12.

For all 15 sound pairs, recognition of upward pitch level shifts was uniform among all listeners.

Comparing the results of subtests S1 and S2 with S4 for Series 1–9 (Tables 1 and 2, replicas of the reference sounds of close-mid vowels), highly comparable vowel recognition results were obtained, with only four single vowel identifications of two listeners found to differ (see Series 1 and 7, extended online version of the table, listeners L2 and L5, the identifications marked with "*"). Notably, the results of the subtest S4 showed simultaneous vowel quality and pitch level shifts, recognised uniformly among all sounds and all listeners. Thereby, the recognised qualities of the synthesised sounds without attenuating the harmonics in between the integer multiples of *D*1 matched directly with the qualities of the natural reference sounds.

Comparing the results of subtests S1 and S2 with S4 for Series 10–15 (Tables 1 and 2, replicas of the reference sounds of open-mid vowels), more pronounced differences appeared, with 22 single vowel identifications of listeners being different (see identifications marked with "*"). However, the recognition rates for vowel quality shifts in an open–close direction barely changed: The recognition rate was 100% for Series

10–13, 90% for Series 14 and 80% for Series 15. In these terms, the results of subtests S1–S3 and S4 for the sounds in Series 10–15 were still comparable to those of Series 1–9. However, the differences in results indicated an impact of sound context and listening tasks on vowel quality recognition. The recognition of upward shifts of pitch levels was again uniform among all listeners.

Vowel and pitch recognition of the transitional sounds: For Series 1–9 (replicas of the reference sounds of close-mid vowels), in general, the results of the vowel and pitch recognition tests for the sounds with stepwise attenuation of the levels of the harmonics that are not integer multiples of the *D*1 frequency showed transition patterns from close-mid to close vowels and from lower to higher pitch levels (see Table 3, results for subtests S1+S5, S2+S6 and S3+S7). Thereby, the transition details of these patterns proved to be somewhat dependent on the natural reference sounds and the listeners. Furthermore, in the transition, recognition sometimes proved to be unstable or inconsistent: Within-listener differences in both vowel quality and pitch level recognition for AB and BA presentation of sound pairs occurred, and pitch level shifts occurred that were not associated with vowel quality shifts (for details, see the extended online version of Table 3, subtests S6 and S7; note that vowel quality and pitch levels were tested separately).

The same held true for Series 10–14 (replicas of the reference sounds of open-mid vowels) and the transitions from open/open-mid to close-mid/close vowel qualities and from lower to higher pitch levels. The results of Series 15 were also in accordance with these transition patterns, but some back–front confusion occurred in the results of listeners L4 and L5.

Double-vowel and double-pitch recognition: The results of testing double-vowel and double-pitch recognition (subtests S8 and S9, see Table 3 for details and Table 4 for a numerical distribution) showed comparable but more pronounced results as found in the previous experiment discussed in Chapter M6.9: (i) For each of the sound series investigated, sounds occurred for which two vowel qualities and/or two pitch levels were recognised. Notably, a labelling majority for simultaneous double-vowel and double-pitch recognition for at least one of the transitional replicas was found for all nine series of front vowel sounds and three of the six series of back vowel sounds. Further, in contrast to the previous experiment, simultaneous double-vowel and double-pitch recognition occurred only for the transitional sounds with stepwise attenuation of the harmonics that are not integer multiples of the *D*1 but not for the opposing sounds without and with full attenuation of these harmonics. (ii) Double-pitch recognition without simultaneous double-vowel recognition also occurred. (iii) Again in contrast to the previous experiment, with one exception of a single labelling, double-vowel recognition without simultaneous double-pitch recognition did not occur in the present investigation.

Listener-specific aspects: As said, for the opposing sounds, consensus on vowel recognition proved to be high and recognition of pitch differences proved to be uniform among the listeners. Thus, between-listener differences were again marginal. However, as was the case in the previous experiment, more pronounced between-listener differences occurred for the transitional sounds: The testing of vowel and pitch recognition revealed distinct recognition profiles for each of the listeners, with similar differences as found in the previous experiment (see Chapter M6.9).

Specific aspects of the reference sounds: According to a comparison of the results of the series, some of the occurring differences were indicated to relate to the individual harmonic configuration of a sound.

Comparison of calculated f_o, HCF and recognised pitch: In Table 3, HCF (see Column 9), calculated f_o (see Column 10) and the results of the listening tests for pitch recognition (see Columns 19–20) are listed for all synthesised sounds. According to the labelling majority of ≥ 80%, the shift of calculated f_o from a lower to a higher level occurred preceding or succeeding the shift from a lower to a higher pitch level for 5 of 15 series.

Discussion

The entire line of experimentation on the question of whether vowel recognition relates to pitch (or to a comparable perceptual referencing) produced consistent indications for such a relation. It finally culminated in the evidence of this relation provided by the results of the previous and the present study: By manipulating the harmonic spectrum of a natural reference vowel sound, two sounds could be produced with equal spectral maxima but with two different recognised vowel qualities associated with two different recognised pitch levels, and the vowel quality shift direction in relation to pitch proved to be consistent for these sounds, that is, rising pitch levels were associated with vowel quality shifts in an open–close direction; in addition, through the above manipulation, it was even possible to produce single sounds for which two vowel qualities and two pitch levels were identified.

As stated, the double-vowel and double-pitch recognition phenomenon is understood here as representing a core phenomenon of vowel acoustics and recognition. Single sounds for which two vowel qualities and two pitch levels can be recognised and for which associated shift directions are consistent and predictable – rising pitch levels associated with a vowel quality shift in an open–close direction – represent a crucial phenomenon for any statement about the vowel-related acoustic characteristics: The phenomenon would not occur if the perceptual process did not relate vowel recognition to pitch (or to a comparable perceptual referencing). At the same time, in a definitive manner, the phenomenon contradicts the thesis of spectral maxima or filter curves as being vowel quality-specific *per se.*

However, four main aspects that possibly relativise a simple formulation of a vowel–pitch relation have to be considered: (i) Vowel quality shifts related to pitch level shifts were somewhat dependent on the vowel qualities, the individual spectral energy configuration for sounds of a given vowel and the levels and ranges of f_o investigated; (ii) double-pitch recognition occurred more often than simultaneous double-vowel and double-pitch recognition; (iii) vowel quality shifts with only a weak indication of simultaneous pitch level shifts occurred for the sounds of the first experiment described in Chapter M6.7, and a few single cases of vowel quality shifts without pitch level shifts also occurred; (iv) finally, differences in the recognition profiles for different listeners were also found.

The finding that simultaneous double-vowel and double-pitch recognition was not uniform among vowel qualities and levels and ranges of f_o may be understood within the general observation that the spectral representation of vowel quality is nonuniform. (For an illustration of the nonuniform indication of the vowel–pitch relation, see also Figure 1 in the chapter appendix. Note in this context that, when creating these experiments, we have again observed that a one-octave pitch level shift below 200 Hz is generally not associated with a vowel quality shift, contrary to pitch level shifts in the frequency ranges as documented here.) The finding that simultaneous shifts sometimes also depended on the individual spectral energy configuration of sounds of a given vowel may be understood within the foreground–background hypothesis (see the excursus on vowel quality and harmonic spectrum). In these terms, the nonuniform indication of the vowel–pitch relation observed in the above experiments is interpreted here not as a counterargument but as a specification of that relation.

The finding that double-pitch recognition occurred more often than simultaneous double-vowel and double-pitch recognition may indicate that, for transitional sounds, perception and recognition are more sensitive to sound pattern repetition over time than to a differentiation of the energy distribution within the repeating pattern, that is, the vibration form of the repeating pattern itself and, with it, the vowel quality.

Remarkably, except for one labelling, no double-vowel recognition occurred in this experiment without double-pitch recognition.

The finding that marked between-listener recognition differences occurred for transitional sounds underpins the indication of a vowel–pitch relation (or its alternative) in that it points towards the corresponding perceptual referencing: If vowel recognition indeed includes a perceptual referencing to pitch (or its alternative), this perceptual operation is assumed here as being to some degree dependent on the general ability, sensibility and recognition experience of the listeners and their individual strategies for giving attention to and focussing on different sound qualities and separating vowel quality from the sound timbre. In addition, as mentioned in the previous chapter, borders of adjacent vowels are also listener-specific in vowel recognition. Therefore, for transitional sounds, listener-specific recognition profiles are expected from the perspective of a vowel–pitch relation (or its alternative).

In conclusion, evidence is provided here for a general relation of vowel recognition to pitch or to a comparable perceptual referencing to a sound pattern repetition over time, the experimental results depending on the vowel quality, f_o level and spectral energy configuration of the natural reference sound investigated. Further, when investigating this relation, the influence of listening test conditions (including the context of sound presentation) and individual recognition strategies of the listeners also have to be considered. The difference between the vowel–pitch relation and the previously discussed formant pattern and spectral shape ambiguity lies in the evidence provided that it is not f_o – not $H1$, not HCF, not f_o measured based on an algorithm – but pitch to which vowel recognition relates.

With regard to future experiments, some further remarks are worth adding. Obviously, the selection of natural sounds fulfilling the experimental conditions and allowing for the creation of *D*-patterns as described above is laborious, and it requires a large sample of natural reference sounds as a basis, including vocal effort variation in sound production. Also, the investigation requires a tool for harmonic synthesis that produces sounds with a natural-like quality.

In future experiments, *D*-patterns may be investigated for which HCF does not only relate to the frequencies of either *H*1 of the natural reference sound or *D*1 but also to the frequencies of an intermediate value in between if *D*1 is equal to *H*4 or a higher harmonic even in number: To give an example, if *D*1 of a natural reference sound is equal to *H*4, HCF can be investigated for frequency levels *H*1, *H*2 and *H*4 of that sound, with correspondingly deleting intermediate harmonics (for some indications, see Figure 1 in the chapter appendix, third and fourth sound triplet). Most importantly, this allows for the investigation of sounds of all vowels and all f_o levels below c. 300 Hz. (Note in this context that the investigation of sounds of /a/ produced at f_o below 300 Hz is difficult to conduct: The frequency of the first dominant harmonic *D*1 of sounds of this vowel is in most cases above 600 Hz, *D*2 at a doubled frequency level of *D*1 is rare, and deleting the intermediate harmonics that are not integer multiples of *D*1 produces a high-pitched sound.)

As indicated in the results section of this chapter, some vowel recognition differences occurred when comparing the results of subtests S1 and S2 with S4 (see Tables 1 and 2). The same held true for vowel and pitch recognition when comparing the results of subtests S1/S5 with S2/S6 and the two presentation orders. These differences indicate an impact of sound context and test procedure of the listening test on vowel quality and pitch recognition, which has to be considered for future research.

The experimental design of the present experiment allowed for more conclusive results when compared with the results of the previous experiment discussed in Chapter M6.9 because of two reasons: On the one hand, a markedly higher inter-listener consensus of interrelated double-vowel and double-pitch recognition occurred, and on the other hand, (disregarding one single labelling) there were no vowel quality shifts without associated pitch level shifts. Both findings, in their turn, indicate the impact of the spectral fine structure of the sounds – and possibly of the related sound quality – on double-vowel and double-pitch recognition, which also has to be considered for future research.

Finally, in future experiments, the steps of harmonic level attenuation may be increased: It can be expected that the attenuation level interacts with different harmonic configurations of the sounds investigated and different f_o levels.

Chapter appendix

Table 1. (Re-)synthesised sounds related to harmonics at or near spectral peaks of single natural vowel sounds, including stepwise attenuation of harmonics that are not integer multiples of *D*1: *D*-patterns investigated and separate vowel and pitch recognition results of the opposing sounds of a series (recognition subtests S1–S3). Columns 1–5 = natural reference sounds (S/L = sound series and sound links; each sound link refers to a natural reference sound and the two related opposing synthesised sounds without and with full attenuation of the harmonics that are not integer multiples of *D*1; V = intended and recognised vowel quality; SP = speaker group, where m = men, w = women, c = children; fo r = calculated f_o of the natural reference sound, in Hz; VE = vocal effort). Columns 6–10 = spectral specification used for synthesis (*D*-pattern = vowel-related dominant harmonics as multiple integers of *D*1, in numbers and frequencies in Hz; AH(i) = attenuation of the levels of the intermediate harmonics that are not integer multiples of *D*1, in dB; ΔHCF = HCF difference of opposing sounds, in semitones ST; fo s = calculated f_o of the synthesised sounds, in Hz). Columns 11–18 = recognised dominant or prominent vowel qualities (confusion matrix, summary of subtests S1 and S2, with ten identifications per sound in total; vowel openness is given as o = open, o-m = open-mid, c-m = close-mid, c = close). Columns 19–20 = recognised dominant or prominent lower or higher pitch level (summary of subtest S3, with five identifications per sound in total). Extended online table: Columns 21–35 = listener-specific details of the recognition results. Column 36 = spectral editing (adjustments of the levels of single harmonics *H*(i) applied to the spectrum of the natural reference sounds for sound synthesis, given for a harmonic number in dB). Columns 37 and 38 = additional natural sounds included in subtest S1 (V = vowel intended and recognised; fo = calculated f_o). Colour code (including extended online table): Dark blue = recognition rates of ≥ 80% according to vowel intention or vowel openness for the synthesised replicas of close-mid vowels (Series 1–9) and of open-mid vowels (Series 10–15) associated with a 100% recognition rate for lower pitch levels (all series); dark red = recognition rates of ≥ 80% for synthesised replicas of close vowels (Series 1–9) and for close-mid and close vowels (Series 10–15) associated with a 100% recognition rate for higher pitch levels (all series); light blue = recognition of close-mid vowels (Series 1–9) or of open and open-mid vowels (Series 10–15) and/or recognition of lower pitch levels; light red = recognition of close vowels (Series 1–9) or of close-mid and close vowels (Series 10–15) and/or recognition of higher pitch levels; grey = back–front confusion.
[M-06-10-T01]

Table 2. (Re-)synthesised sounds related to harmonics at or near spectral peaks of single natural vowel sounds, including stepwise attenuation of harmonics that are not integer multiples of *D*1: Results of simultaneous vowel and pitch recognition of the opposing sounds of a series (recognition subtest S4). For the columns, see Table 1. (Note that the sound presentation was in AB and BA order, with ten identifications per sound in total.) The colour code also corresponds to Table 1, with the exception that the code for the recognition rate for the synthesised replicas of open/open-mid vowels in Series 10–15 was set to ≥ 70%. In addition, labelling differences in comparison to the results of Table 1 (compare the recognition details in the online tables) are marked with "*", and the recognition of schwa is also marked in grey (extended online table only).
[M-06-10-T02]

Table 3. (Re-)synthesised sounds related to harmonics at or near spectral peaks of single natural vowel sounds, including stepwise attenuation of harmonics that are not integer multiples of *D*1: Vowel and pitch recognition results of all sounds of a series (all recognition subtests). Columns 1–20 = see Table 1 (vowel recognition as the summary of subtests S1/S5 and S2/S6, with 15 identifications per sound in total; pitch recognition as the summary of subtests S3/S7, with ten identifications per sound in total; note that, in Column 9, competing HCFs are given for the sounds with attenuation of -12 to -36 dB for the levels of the harmonics that are not integer multiples of *D*1). Columns 21–22 = cases of two vowel qualities simultaneously recognised for a single sound (double-vowel recognition, DV) and/or two pitches simultaneously recognised for a single sound (double-pitch recognition DP), marked with "y" for "yes" (summary of subtests S8 and S9, each with five identifications per sound in total). Extended online table: Columns 23 ff. = listener-specific details of the recognition results. Colour code, see Table 1. In addition, in Columns 21 and 22, the most pronounced double-vowel and double-pitch recognition found among the listeners for a sound or for sounds of a series of synthesised replicas are marked in dark green; in Columns 23 ff., double-vowel and/or double-pitch recognition of a listener for a single sound are marked in light green. To improve readability and give an example, the results of subtests S1–S3 (added to the results of subtests S5–S7) are marked with "*" for Series 1 and listener L1. Note also the single exception of a double-vowel recognition without a double-pitch recognition for the second sound of Series 9, indicated as (+1) and marked in grey in the extended online table.
[M-06-10-T03]

Table 4. (Re-)synthesised sounds related to harmonics at or near spectral peaks of single natural vowel sounds, including stepwise attenuation of harmonics that are not integer multiples of *D*1: Sounds per listener with double-vowel and/or double-pitch recognition. Column 1 = listeners. Column 2 = number of sounds with parallel double-vowel and double-pitch recognition (DV&DP). Column 3 = number of sounds with double-pitch recognition only (DP). Column 4 = number of sounds with double-vowel recognition only (DV).
[M-06-10-T04]

Figure 1. (Re-)synthesised sounds related to harmonics at or near spectral peaks of single natural vowel sounds, including stepwise attenuation of harmonics that are not integer multiples of *D*1: Illustration of the nonuniform indication of the vowel–pitch relation. The first sound triplet (see Series 9 in Table 1, sounds 1–3) shows a natural reference sound of /o/ produced by a child at an intended f_o of 262 Hz, and the two opposing synthesised replicas without and with full attenuation of the harmonics that are not integer multiples of *D*1. As discussed above, the two replicas differ in vowel quality and pitch level. The second sound triplet (sounds 4–6, additional sounds not investigated in the experiment) shows a second natural reference sound of /o/ of a child, produced at the same intended f_o level, and the two opposing synthesised replicas (produced with harmonic synthesis including enhancement of the level of *H*4 of the reference spectrum). The two replicas differ in pitch level but not vowel quality (author's estimate). Thus, the harmonic level configuration (the spectral energy distribution) of a sound has to be accounted for by the vowel–pitch relation. The third sound triplet (sounds 7–9, additional sounds not investigated in the experiment) shows a natural reference sound of /a/ produced by a woman at a similar f_o level as for the preceding reference sounds of /o/, and the two synthesised replicas without and with full attenuation of the harmonics that are not integer multiples of 0.5×*D*1. For these replicas, in contrast to the first sound triplet of /o/, the vowel quality is maintained despite a comparable pitch level shift (author's

estimate). Thus, the vowel quality of the sounds also has to be accounted for by the vowel–pitch relation. The fourth sound triplet (sounds 10–12, additional sounds not investigated in the experiment) shows a natural reference sound of /o/ produced by a man at an intended f_o of 98 Hz, and the two synthesised replicas without and with full attenuation of the harmonics that are not integer multiples of 0.5×*D*1. The pitch level for these replicas again shifts by one octave while vowel quality is maintained (author's estimate). Thus, the levels and ranges of f_o and the corresponding pitch levels of reference sounds and their replicas have to be accounted for by the vowel–pitch relation, too.
[M-06-10-F01]

Table 1. (Re-)synthesised sounds related to harmonics at or near spectral peaks of single natural vowel sounds, including stepwise attenuation of harmonics that are not integer multiples of D1: D-patterns investigated and separate vowel and pitch recognition results of the opposing sounds of a series (recognition subtests S1–S3). [M-06-10-T01] Extended online table: ↗

Sounds											Recognition									
References						Synthesis					Vowel								Pitch	
S/L		V	SP	fo r	VE	D–pattern		AH(i)	ΔHCF	fo s	o-m/c-m				c				level	
				Hz		Numbers	Hz	dB	ST	Hz	e	ø	ɔ	o	i	y	–	u	l	h
1	↗	e	m	123	low	3–21–24	369–2583–2952	0	19	123	10								5	
								-100		370					9	1				5
2	↗	e	w	218	med	2–12–14	436–2616–3052	0	12	218	10								5	
								-100		435					10					5
3	↗	e	c	264	med	2–10–12	528–2640–3168	0	12	264	10								5	
								-100		528					10					5
4	↗	ø	m	147	med	2–10–14	294–1470–2058	0	12	147		10							5	
								-100		293						10				5
5	↗	ø	w	221	med	2–8–12	442–1768–2652	0	12	221		10							5	
								-100		443						10				5
6	↗	ø	c	244	med	2–8–12	488–1952–2928	0	12	244		10							5	
								-100		488						10				5
7	↗	o	m	131	med	3–6	393–786	0	19	131			2	8					5	
								-100		393				1				9		5
8	↗	o	w	223	med	2–4	446–892	0	12	223				10					5	
								-100		446								10		5
9	↗	o	c	269	med	2–4	538–1076	0	12	269				10					5	
								-100		539								10		5

Table 1 (continuation). [06-10-T01]

Sounds										Recognition									
References					Synthesis					Vowel								Pitch	
S/L	V	SP	fo r	VE	D–pattern		AH(i)	ΔHCF	fo s	o/o-m				c-m/c				level	
			Hz		Numbers	Hz	dB	ST	Hz	ə	ε	a	ɔ	i/y	e/ø	o	u	l	h
10 ↗	ε	m	147	med	4–12–16	588–1764–2352	0	24	147		10							5	
							-100		589					5	5				5
11 ↗	ε	w	150	med	4–16–20	600–2400–3000	0	24	150		10							5	
							-100		600					1	9				5
12 ↗	ε	c	243	med	3–9–12	729–2187–2916	0	19	243		10							5	
							-100		730					2	8				5
13 ↗	ɔ	m	156	med	3–6	468–936	0	19	156				10					5	
							-100		469							2	8		5
14 ↗	ɔ	w	172	high	3–6	516–1032	0	19	172				10					5	
							-100		515								10		5
15 ↗	ɔ	c	245	med	3–6	735–1470	0	19	245			2	8					5	
							-100		735				1		1	3	5		5

Table 2. (Re-)synthesised sounds related to harmonics at or near spectral peaks of single natural vowel sounds, including stepwise attenuation of harmonics that are not integer multiples of D1: Results of simultaneous vowel and pitch recognition results of the opposing sounds of a series (recognition subtest S4).
[M-06-10-T02. Extended online table: ☐↗

Sounds										Recognition									
References					Synthesis					Vowel								Pitch	
S/L	V	SP	fo r	VE	D–pattern		AH(i)	∆HCF	fo s	o-m/c-m				c				level	
			Hz		Numbers	Hz	dB	ST	Hz	e	ø	ɔ	o	i	y	–	u	l	h
1	e	m	123	low	3–21–24	369–2583–2952	0	19	123	10								10	
							-100		370					10*					10
2	e	w	218	med	2–12–14	436–2616–3052	0	12	218	10								10	
							-100		435					10					10
3	e	c	264	med	2–10–12	528–2640–3168	0	12	264	10								10	
							-100		528					10					10
4	ø	m	147	med	2–10–14	294–1470–2058	0	12	147		10							10	
							-100		293						10				10
5	ø	w	221	med	2–8–12	442–1768–2652	0	12	221		10							10	
							-100		443						10				10
6	ø	c	244	med	2–8–12	488–1952–2928	0	12	244		10							10	
							-100		488						10				10
7	o	m	131	med	3–6	393–786	0	19	131				10*					10	
							-100		393								10*		10
8	o	w	223	med	2–4	446–892	0	12	223				10					10	
							-100		446								10		10
9	o	c	269	med	2–4	538–1076	0	12	269				10					10	
							-100		539								10		10

Table 2 (continuation). [06-10-T02]

Sounds										Recognition									
References					Synthesis					Vowel								Pitch	
S/L	V	SP	fo r	VE	D–pattern		AH(i)	ΔHCF	fo s	o/o-m				c-m/c				level	
			Hz		Numbers	Hz	dB	ST	Hz	ə	ε	a	ɔ	i/y	e/ø	o	u	l	h
10	ε	m	147	med	4–12–16	588–1764–2352	0	24	147		10							10	
							-100		589					6*	4*				10
11	ε	w	150	med	4–16–20	600–2400–3000	0	24	150		10							10	
							-100		600					2*	8*				10
12	ε	c	243	med	3–9–12	729–2187–2916	0	19	243		10							10	
							-100		730					5*	5*				10
13	ɔ	m	156	med	3–6	468–936	0	19	156				7*			3*		10	
							-100		469								10*		10
14	ɔ	w	172	high	3–6	516–1032	0	19	172	1*			9*					10	
							-100		515								10		10
15	ɔ	c	245	med	3–6	735–1470	0	19	245	2*			8*					10	
							-100		735						1	1*	8*		10

Table 3. (Re-)synthesised sounds related to harmonics at or near spectral peaks of single natural vowel sounds, including stepwise attenuation of harmonics that are not integer multiples of D1: Vowel and pitch recognition results of all sounds of a series (all recognition subtests). [M-06-10-T03] Extended online table: ↗

Sounds										Recognition											
References					Synthesis					Vowel								Pitch			
S/L	V	SP	fo r	VE	D–patter		AH(i)	HCF	fo s	o-m/c-m				c				single		double	
			Hz		Harmonics	Hz	dB		Hz	e	ø	ɔ	o	i	y	–	u	l	h	DV	DP
1 ↗	**e**	m	123	low	3–21–24	369–2583–2952	0	123	123	15								10		–	1
							-12	low	123	4				11				2	8	5	5
							-24	versus	370					15					10	2	5
							-36	high	370	1				14					10	–	4
							-100	369	370					14	1				10	–	–
2 ↗	**e**	w	218	med	2–12–14	436–2616–3052	0	218	218	15								10		–	–
							-12	low	435	11				4				4	6	3	4
							-24	versus	435	1				14					10	1	4
							-36	high	435					15					10	–	–
							-100	436	435					15					10	–	–
3 ↗	**e**	c	264	med	2–10–12	528–2640–3168	0	264	264	15								10		–	–
							-12	low	264	14				1				7	3	4	5
							-24	versus	528	5				10				2	8	3	5
							-36	high	528	2				13					10	2	5
							-100	528	528					15					10	–	–

Table 3 (continuation). [M-06-10-T03]

Sounds										Recognition											
References					Synthesis					Vowel								Pitch			
S/L	V	SP	fo r	VE	D–patter		AH(i)	HCF	fo s	o-m/c-m				c				single		double	
			Hz		Harmonics	Hz	dB		Hz	e	ø	ɔ	o	i	y	–	u	l	h	DV	DP
4 ↗	ø	m	147	med	2–10–14	294–1470–2058	0	147	147		15							10		–	–
							-12	low	147		5				10			1	9	5	5
							-24	versus	293		1				14				10	1	5
							-36	high	293						15				10	–	2
							-100	294	293						15				10	–	–
5 ↗	ø	w	221	med	2–8–12	442–1768–2652	0	221	221		15							10		–	–
							-12	low	221		12				3			7	3	2	4
							-24	versus	443		4				11				10	4	5
							-36	high	443						15				10	1	3
							-100	442	443						15				10	–	–
6 ↗	ø	c	244	med	2–8–12	488–1952–2928	0	244	244		15							10		–	1
							-12	low	244		10				5			6	4	3	4
							-24	versus	488	(1 ə)	5				9				10	3	5
							-36	high	488						15				10	1	4
							-100	488	488						15				10	–	–

Table 3 (continuation). [M-06-10-T03]

Sounds										Recognition											
References					Synthesis					Vowel								Pitch			
S/L	V	SP	fo r	VE	D–patter		AH(i)	HCF	fo s	o-m/c-m				c				single		double	
			Hz		Harmonics	Hz	dB		Hz	e	ø	ɔ	o	i	y	–	u	l	h	DV	DP
7 ↗	**o**	m	131	med	3–6	393–786	0	131	131			3	12					10		–	–
							-12	low	131			1	6				8	3	7	5	5
							-24	versus	393				2				13		10	3	5
							-36	high	393								15		10	2	4
							-100	393	393				1				14		10	–	–
8 ↗	**o**	w	223	med	2–4	446–892	0	223	223				15					10		–	–
							-12	low	446				13				2	9	1	1	4
							-24	versus	446				6				9	4	6	–	3
							-36	high	446				1				14		10	2	2
							-100	446	446								15		10	–	–
9 ↗	**o**	c	269	med	2–4	538–1076	0	269	269				15					10		–	–
							-12	low	538				15					8	2	2(+1)	3
							-24	versus	538				9				6	2	8	2	4
							-36	high	538				5				10	2	8	1	2
							-100	538	539								15		10	–	–

Table 3 (continuation). [M-06-10-T03]

Sounds										Recognition											
References					Synthesis					Vowel								Pitch			
S/L	V	SP	fo r	VE	D–patter		AH(i)	HCF	fo s	o/o-m				c-m/c				single		double	
			Hz		Harmonics	Hz	dB		Hz	ə	ε	a	ɔ	i/y	e/ø	o	u	l	h	DV	DP
10 ↗	ε	m	147	med	4–12–16	588–1764–2352	0	147	147		15							10		–	–
							-12	low	147		8			3	4			3	7	4	5
							-24	versus	589					4	11				10	4	5
							-36	high	589	1				5	9				10	–	4
							-100	588	589	1				6	8				10	–	–
11 ↗	ε	w	150	med	4–16–20	600–2400–3000	0	150	150		15							10		–	–
							-12	low	150		10				5			6	4	3	5
							-24	versus	600						15				10	1	5
							-36	high	600					1	14				10	–	5
							-100	600	600					2	13				10	–	–
12 ↗	ε	c	243	med	3–9–12	729–2187–2916	0	243	243		15							10		–	–
							-12	low	243		7			1	7			6	4	4	5
							-24	versus	730		1			3	11				10	3	5
							-36	high	730					2	13				10	3	4
							-100	729	730					3	12				10	–	–

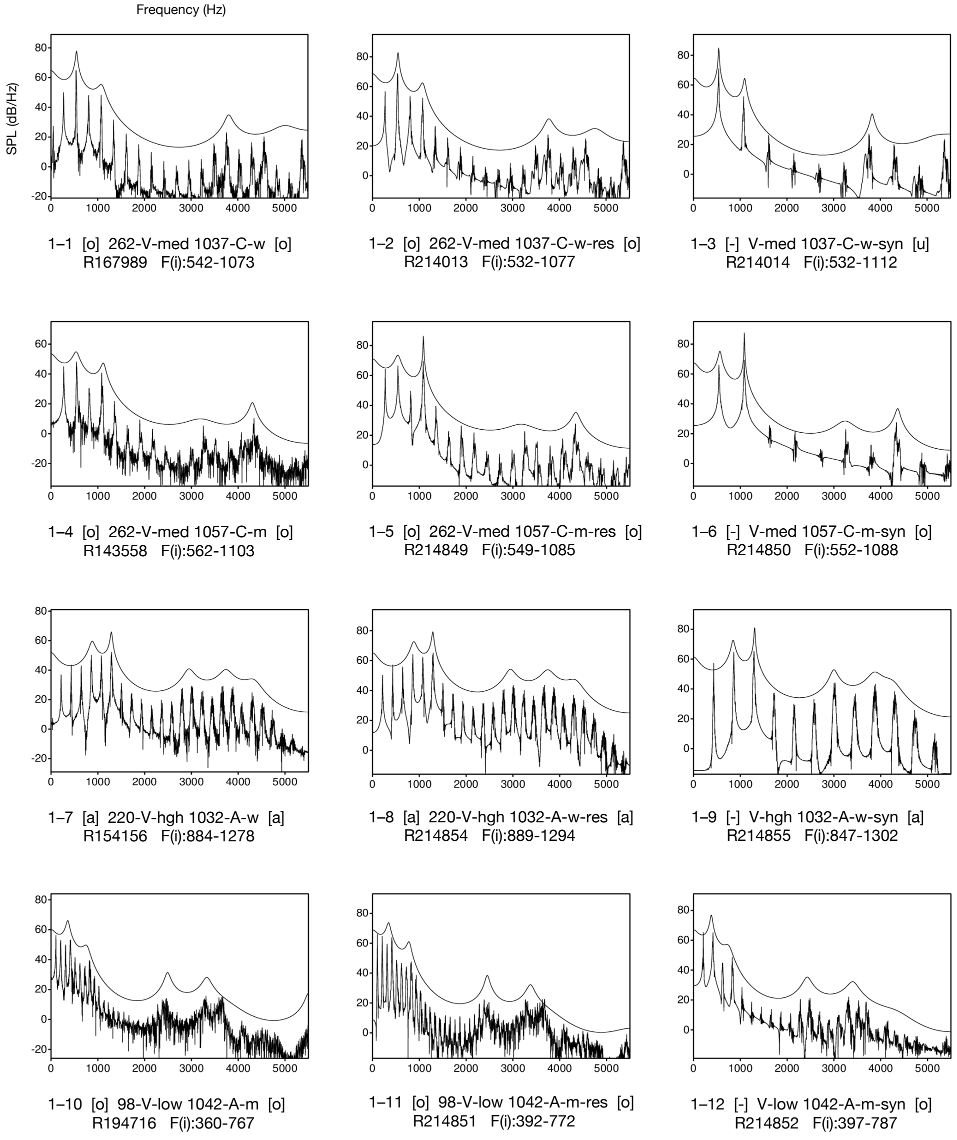
Table 3 (continuation). [M-06-10-T03]

Sounds										Recognition											
References					Synthesis					Vowel								Pitch			
S/L	V	SP	fo r	VE	D–patter		AH(i)	HCF	fo s	o/o-m				c-m/c				single		double	
			Hz		Harmonics	Hz	dB		Hz	ə	ɛ	a	ɔ	i/y	e/ø	o	u	l	h	DV	DP
13 ↗	ɔ	m	156	med	3–6	468–936	0	156	156				15					10		–	…
							-12	low	156				3			9	3	5	5	3	5
							-24	versus	469				2			1	12		10	3	5
							-36	high	469							3	12		10	3	5
							-100	468	469							2	13		10	–	–
14 ↗	ɔ	w	172	high	3–6	516–1032	0	172	172				15					10		–	1
							-12	low	172				15					9	1	1	2
							-24	versus	515				7			2	6	1	9	4	5
							-36	high	515				1			1	13		10	2	4
							-100	516	515								15		10	–	–
15 ↗	ɔ	c	245	med	3–6	735–1470	0	245	245			3	12					10		–	–
							-12	low	245	1		4	9			1		7	3	2	4
							-24	versus	735	1			6		3	5		2	8	2	4
							-36	high	735				3		2	3	7	2	8	2	5
							-100	735	735				1		2	5	7		10	–	–

Table 4. (Re-)synthesised sounds related to harmonics at or near spectral peaks of single natural vowel sounds, including stepwise attenuation of harmonics that are not integer multiples of D1: Sounds per listener with double-vowel and/or double-pitch recognition. [M-6-10-T04]

Listeners	DV&DP	DP	DV
L1	15	17	0
L2	13	26	1
L3	31	14	0
L4	19	14	0
L5	24	18	0

Figure 1. (Re-)synthesised sounds related to harmonics at or near spectral peaks of single natural vowel sounds, including stepwise attenuation of harmonics not being integer multiples of D1: Illustration of the nonuniform indication of the vowel–pitch relation. [M-06-10-F01]



1–1 [o] 262-V-med 1037-C-w [o]
R167989 F(i):542-1073

1–2 [o] 262-V-med 1037-C-w-res [o]
R214013 F(i):532-1077

1–3 [-] V-med 1037-C-w-syn [u]
R214014 F(i):532-1112

1–4 [o] 262-V-med 1057-C-m [o]
R143558 F(i):562-1103

1–5 [o] 262-V-med 1057-C-m-res [o]
R214849 F(i):549-1085

1–6 [-] V-med 1057-C-m-syn [o]
R214850 F(i):552-1088

1–7 [a] 220-V-hgh 1032-A-w [a]
R154156 F(i):884-1278

1–8 [a] 220-V-hgh 1032-A-w-res [a]
R214854 F(i):889-1294

1–9 [-] V-hgh 1032-A-w-syn [a]
R214855 F(i):847-1302

1–10 [o] 98-V-low 1042-A-m [o]
R194716 F(i):360-767

1–11 [o] 98-V-low 1042-A-m-res [o]
R214851 F(i):392-772

1–12 [-] V-low 1042-A-m-syn [o]
R214852 F(i):397-787

M7 Spectral Variation of Vowel Sounds and its Nonuniform Character – Broadening the Documentation of the Variation Extent

M7.1 Different Vowel-Related Spectral Peak Numbers

Introduction

As was discussed when describing the relation of the lower vowel spectrum to f_o and the resulting formant pattern and spectral shape ambiguity (see Chapters M2 and M3), and as was also indicated in the documentation of vocal effort-related spectral variation (see Chapter M5.4), the vowel spectrum exhibits a nonuniform character concerning these three aspects. Earlier, we have also discussed further aspects of nonuniform spectral variation observed for natural sounds of a given vowel in vowel-related frequency ranges, such as (i) an inconstant number of vowel-related spectral peaks, (ii) occurring inversions of vowel-related relative spectral energy maxima and minima, (iii) occurring flat or sloping vowel-related spectral energy distribution and (iv) the fine structure of spectral energy distribution having an impact on the relation of the lower vowel spectrum to f_o, without and with vocal effort variation (see Maurer and Landis, 1995; Maurer et al., 2019; Preliminaries, Chapters 7 and 8 and related materials), these aspects obstructing the formulation of a general concept of relating recognised vowel quality to a specific spectral peak pattern or an average spectral shape, even if f_o is included. All these observations are taken up, completed, discussed and documented anew in the following chapters based on the natural sounds of the Zurich Corpus, with the exception of the illustration of occurring flat or sloping vowel-related spectral energy distribution, which is provided for both natural and synthesised vowel sounds.

The first observation, discussed in this chapter, concerns an inconstant number of spectral peaks for sounds of a given vowel in their vowel-related frequency range.

It is well known that sounds of back vowels can manifest only one spectral peak < 1–1.5 kHz instead of the expected two peaks. This phenomenon is generally understood as being a consequence of two formants close in frequency (formant merger): “When the formants are close together […] neither the wide- nor the narrowband spectrum gives a good indication of the formant frequencies. […] The first two formants appear as a single peak below 1000 Hz. Their frequencies cannot be determined from these spectra.” (Ladefoged, 2003, pp. 119–120). Often, the recognition of vowel sounds related to this type of spectra is understood within the concept of a “centre of gravity” effect in terms of an auditory spectral averaging process (Chistovich and Lublinskaya, 1979, Chistovich, 1985; however, see Assmann, 1991, and de Cheveigné and Kawahara, 1999, for findings and arguments that run counter to such a general concept). Within this concept, two formants within a frequency distance not exceeding 3.5 Bark are assumed to be represented by a spectral peak in between the assumed formant frequencies.

Comparably, attempts were made to relate the *F*-patterns of synthesised sounds of front vowels that are based on either two or three formants, with *F*2 of the two-formant sounds (often termed *F*2-prime) being in between *F*2–*F*3 of the three-formant sounds, often also related to a spectral “centre of gravity” effect (for an overview and critical review, see Kiefte et al., 2013).

Further, according to the literature, in some sound spectra of some speakers, an additional spectral envelope peak may occur between the expected first and second or second and third formant. According to prevailing methodological rules for determining formants, this maximum is not interpreted as vowel specific but as a specific characteristic of the voice of the speaker in question and it is referred to as a spurious formant (see e.g. Ladefoged, 1996, p. 210–212; 2003, pp. 114–115; see also the Preliminaries, p. 33).

Finally, as discussed in Chapter M5.1, the spectra of vowel sounds produced with breathy phonation may manifest an increased amplitude of the first harmonic, which sometimes shows the highest energy level in the spectrum.

However, these types of spectral peak patterns that deviate from the general assumption of vowel-specific peak numbers were barely investigated systematically, including a variation of basic production parameters, and our own earlier investigations did not support a general explanation of different formant numbers of vowel sounds as being a phenomenon of formant merging or *F*2-prime or spurious formants (see Maurer and Landis, 1995; Preliminaries, p. 56–58 and 132–157). Therefore, and to again document the possible variability of the vowel spectrum based on the newly compiled Zurich Corpus and to embed it into the line of argument of this treatise, a corresponding study was conducted addressing two questions: What is the variation of spectral peak numbers that can be observed for natural vowel sounds, including different phonation types and different levels of f_o for the voiced sounds? Do sound spectra manifesting "unexpected" spectral peak numbers generally comply with the concept of formant merging or of spurious formants? Based on the Zurich Corpus and addressing these two questions, sounds of the eight long Standard German vowels produced by speakers of different speaker groups were inspected, for which the spectra manifested a varying number of vowel-related spectral peaks. For each vowel, exemplary sound series were compiled for documentation and illustration in this treatise.

Experiment

Vowel sounds and speakers – voiced, breathy and creaky sounds: Sounds of the eight long Standard German vowels of the Zurich Corpus produced by the speakers documented in the corpus with voiced, breathy and creaky phonation were taken as a basis of investigation, including a variation of f_o, vocal effort, vowel context (V and CVCV context) and production style.

Vowel sounds and speakers – whispered sounds: Likewise, sounds of the same vowels produced by the speakers documented in the corpus with whispered phonation were taken as a second, separate basis of investigation, including a variation of vowel context (V and CVCV context).

Inspection of the spectra of voiced, breathy and creaky sounds, and sound selection: The spectra of sounds with voiced, breathy and creaky phonation, that is, sounds manifesting a harmonic or quasi-harmonic spectrum, were analysed and rated first: For back vowels and /a/, examples of sounds with either one or two or even more spectral peaks < c. 1.5 kHz were selected; for front vowels, examples of sounds with either one or two spectral peaks > c. 1.3 kHz were selected. With few exceptions, the listening test conducted when creating the corpus provided a 100% recognition rate (matching vowel intention) for the selected sounds. As a result, eight compilations of sounds of the eight vowels were created (for the number of sounds per vowel, see Table 1 in the chapter appendix).

Inspection of the spectra of whispered sounds, and sound selection: Subsequently, the spectra of sounds with whispered phonation, that is, sounds lacking a harmonic or quasi-harmonic spectrum, were analysed and rated separately: For back vowels and /a/, examples of sounds with either one, two or three spectral peaks < c. 1.5 kHz were selected; for front vowels, examples of sounds with either one or two spectral peaks < c. 1.3 kHz or one or two spectral peaks > 1.5 kHz were selected. All the selected sounds were fully recognised in the standard listening test conducted when creating the corpus (100% recognition rate matching vowel intention). As a result, eight compilations of sounds of the eight vowels were created (for the number of sounds per vowel, see Table 2 in the appendix to this chapter).

Spectral peak estimation: In general, peak patterns were investigated in relation to frequency ranges that are assumed to be vowel quality-related. For voiced sounds, spectral peaks were interpreted in terms of identifiable and distinct relative spectral energy maxima of a single harmonic (all sounds) or of two neighbouring harmonics (sounds at f_o < c. 250 Hz). For breathy and creaky sounds, spectral peaks were interpreted either according to the voiced sounds or in terms of pronounced relative spectral energy maxima of narrowly defined frequency bands, in most cases with a single manifest tip, above all for peaks > 1 kHz. For whispered sounds, spectral peaks were interpreted in terms of identifiable relative spectral energy maxima of frequency bands not exceeding c. 150 Hz. However, the definition of spectral peaks for whispered sounds appertains to the general methodological problem of spectral peak estimation of vowel sounds. For all sounds, independent of their phonation type, only spectral peak patterns but not calculated formant patterns were interpreted.

Additional note: The inspection, spectral peak estimation and sound selection was made based on previous experiences regarding different numbers of spectral peaks in vowel sounds and was focused on finding examples best suited to documenting and illustrating the phenomenon in the context of the present treatise. The selection did not further consider the statistical distribution of the different pattern types. Sounds of back vowels and /a/ with flat or sloping spectral energy distribution < c. 1.5 kHz, sounds of front vowels with flat or sloping spectral energy distribution > c. 1.5 kHz and sounds produced at high levels of f_o will be addressed separately in the following chapters.

Results

For each of the eight vowels investigated, separating voiced, breathy and creaky sounds from whispered sounds, Tables 1 and 2 in the chapter appendix list the entire compilation of selected sounds and present extracts of these compilations. The sound series given in the tables illustrate specific aspects of the occurring spectral peak number variation according to the following description.

Sounds of /u/ (Tables 1 and 2, Series 1): Concerning voiced, breathy and creaky sounds, besides the cases with two distinct spectral peaks < c. 1.5 kHz (as generally expected), cases of sounds with only one peak in this frequency range often occurred, above all for sounds produced with medium or low vocal effort or with middle or higher f_o levels. Comparing sounds with one and two peaks produced at similar f_o levels shows that the second peak is “absent” for the former. Thus, these cases cannot be understood as cases of formant merging, but they indicate that, for the sounds presented, the second spectral peak may have an effect on sound timbre yet not on vowel quality. Note also cases of creaky sounds with prominent energy in the low frequency range < 250 Hz. Besides, the levels of second peaks (if manifest) markedly varied. Concerning whispered sounds, most spectra exhibited the expected two lower spectral peaks. Only rare cases with either one or three peaks < 1.5 kHz were observed.

Sounds of /o/ (Tables 1 and 2, Series 2): Concerning voiced, breathy and creaky sounds, besides the cases with two distinct spectral peaks < c. 1.5 kHz (as generally expected), cases of sounds with other types of peak patterns for this frequency range occurred as follows: Sounds with only one spectral peak; one, two or three peaks < 1.5 kHz for creaky sounds; dominant *H*1 and weak or “undetectable” first expected peak; dominant *H*1 and one or two peaks < 1.5 kHz. The comparison of sounds with one and two peaks produced at similar f_o levels again shows that the second peak is “absent” for the former. Thus, as for the sounds of /u/, these cases cannot be understood within the formant merging concept. Levels of second peaks (if manifest) varied markedly. With regard to whispered sounds, cases of one, two or three peaks < 1.5 kHz occurred.

Sounds of /a/ (Tables 1 and 2, Series 3): Concerning voiced, breathy and creaky sounds, besides the cases with two distinct spectral peaks < c. 1.5 kHz (as generally expected), cases of sounds with other types of peak patterns for this frequency range occurred as follows: Rippled spectrum in the entire frequency range up to 1–1.5 kHz, sometimes including a dominant *H*1 or dominant *H*1 and *H*2 (note that for spectra of this type, more than two peaks can be interpreted); dominant *H*1 or *H*1 and *H*2 or a noise peak at very low frequencies followed by a prominent frequency band or by two peaks; dominant *H*1 and a subsequent (sometimes weak) single peak < 1.5 kHz; only one peak < 1.5 kHz. Comparing sounds with one and two peaks produced at similar f_o levels proved to be difficult and did not provide uniform indications (see also the Discussion). With regard to whispered sounds, cases of one, two or three peaks or a prominent frequency band < 1.5 kHz occurred.

Sounds of /ɛ/ (Tables 1 and 2, Series 4): Concerning voiced, breathy and creaky sounds, besides the cases with one distinct spectral peak < c. 1.5 kHz and two distinct peaks > c. 1.5 kHz (as generally expected), cases of sounds with other types of peak patterns for these frequency ranges occurred as follows: Weak or barely definable first peak structure < 1 kHz in terms of flat or sloping spectral energy, often extended over a frequency range exceeding 500 Hz and sometimes covering prominent spectral energy up to 1 kHz; dominant *H*1 or *H*1 and *H*2 or low-frequency noise preceding the expected first peak, sometimes with a weak or "absent" second peak; weak or "absent" second or third peak. Comparing cases of sounds with "absent" second or third spectral peaks and cases with expected peak patterns, no indication was found in the sample of selected sounds that, in the spectra, two higher peaks correspond with a single peak in between them in general terms. Thus, as was the case for the close and close-mid back vowels, no indication was found for different peak numbers as a phenomenon of formant merging. Although spectral peaks of recognisable sounds produced at high f_o levels cannot be estimated, some sounds indicated a first spectral energy maximum above 1 kHz. Again, manifest spectral peak levels varied markedly. With regard to whispered sounds, above all, cases of a "split" lower peak (two or three peaks in the frequency range of an expected single peak < 1.2 kHz) occurred.

Sounds of /ø/ (Tables 1 and 2, Series 5): Concerning voiced, breathy and creaky sounds, besides the cases with one distinct spectral peak < c. 1 kHz and two distinct peaks > c. 1.3 kHz (as generally expected), cases of sounds with other types of peak patterns for these frequency ranges occurred as follows: Weak or barely definable first peak structure < 1 kHz in terms of sloping spectral energy; dominant *H*1 preceding the expected first peak; two low peaks for creaky sounds; no separating peak structure for the frequency range of the first two expected peaks; weak or "absent" expected second or third peak. Comparing cases of sounds with "absent" expected second or third spectral peaks and cases with expected peak patterns, no general indication is given by the examples selected that two higher peaks correspond with a single peak in between (no formant merging indication). Again, the peak levels varied markedly. With regard to whispered sounds, above all, cases of a "split" lower peak (two peaks in the frequency range of an expected single peak < 1 kHz) occurred.

Sounds of /e/ (Tables 1 and 2, Series 6): Concerning voiced, breathy and creaky sounds, besides the cases with one distinct spectral peak < c. 1 kHz and two distinct peaks > c. 1.7 kHz (as generally expected), cases of sounds with other types of peak patterns for these frequency ranges occurred as follows: Weak or barely definable first peak structure < 1 kHz in terms of sloping spectral energy; dominant *H*1 or *H*1–*H*2 preceding the expected first peak; prominent rippled low spectral energy or two low peaks for creaky sounds; weak or "absent" expected second or third expected peak, with no general formant merging indication. Again, the peak levels varied markedly. With regard to whispered sounds, above all, cases of a "split" lower peak (two peaks in the frequency range of an expected single peak < 1.2 kHz) occurred. In addition, cases of sounds with only one spectral peak > 1 kHz in the expected vowel-related frequency range were also observed.

Sounds of /y/ (Tables 1 and 2, Series 7): Concerning voiced, breathy and creaky sounds, besides the cases with one distinct spectral peak < c. 1 kHz and two distinct peaks > c. 1.5 kHz (as generally expected), cases of sounds with other types of peak patterns for these frequency ranges occurred as follows: No spectral peak < 1 kHz for some sounds at high f_o levels; weak or "absent" expected second or third peak, with no general formant merging indication. Again, manifest peak levels varied markedly (for an extreme spectral variation, see Table 1, Series 7e). With regard to whispered sounds, a few cases of "split" lower peaks (two peaks in the frequency range of an expected single peak < 1 kHz) and a few cases of sounds with only one spectral peak > 1 kHz in the expected vowel-related frequency range occurred.

Sounds of /i/ (Tables 1 and 2, Series 8): Concerning voiced, breathy and creaky sounds, besides the cases with one distinct spectral peak < c. 1 kHz and two distinct peaks > c. 1.5 kHz (as generally expected), above all two cases of sounds with other types of peak patterns occurred for these frequency ranges: No relative spectral energy maximum < 1 kHz for some sounds at high f_o levels, and weak or "absent" expected third peak. Again, manifest peak levels varied markedly (for an extreme spectral variation, see Table 1, Series 8d). With regard to whispered sounds, with a few exceptions of cases of a barely defined spectral peak < 1 kHz, most sound spectra were found to correspond to the expected peak patterns.

Discussion

The inspection of the Zurich Corpus indicated a very pronounced variation of vowel-related spectral peak numbers, and the sound compilations presented in this chapter illustrate this variation. On this basis, the following conclusions are made.

Sounds of /u, o, a/ produced with voiced, breathy or creaky phonation can manifest either one or two lower spectral peaks < 1.5 kHz or, for /o, a/, even three peaks. In addition, sounds with dominant *H*1 and weak or "undetectable" first (expected) peak, or with a rippled spectrum < 1–1.5 kHz, or with dominant *H*1 or *H*1 and *H*2 and a subsequent single peak < 1.5 kHz, also occur. Sounds of /o, a/ produced with whispered phonation can manifest one to three spectral peaks or a frequency band of prominent spectral energy.

Sounds of /ɛ, ø, e, y, i/ produced with voiced, breathy or creaky phonation can manifest many different types of peak structures such as (i) weak or barely definable first spectral peak structure < 1 kHz in terms of sloping spectral energy, (ii) dominant *H*1 or *H*1 and *H*2 for voiced or breathy sounds or a low spectral peak for creaky sounds preceding the first expected peak, (iii) prominent rippled low spectral energy or two low spectral peaks for creaky sounds, (iv) no separating peak structure for the frequency range of the first two expected spectral peaks, (v) "absent" expected second or third spectral peak, (vi) weak higher peak structure and low energy in the corresponding frequency range, and (vii) sounds at high f_o levels with unassessable spectral peak structure. Sounds of these front vowels produced with whispered phonation can manifest "split" lower spectral peaks < 1.5 kHz in terms of two peaks in the frequency range of an expected single peak (in some cases of sounds of /ɛ/, even three peaks were indicated), and they can also manifest only one peak in the vowel-related frequency range > c. 1.3 kHz.

The results of this experiment confirmed our earlier observations regarding spectral peak variation for vowel sounds (see the introduction to this chapter). They give reason to assume that any phenomenological investigation of vowel sounds will provide evidence that, firstly, there is no constant number of spectral peaks for sounds of one vowel quality *in general* and, secondly, inconstant peak numbers as documented here cannot *generally* be explained by formant merging or *F*2-prime spurious formants. Thus, a vowel sound does not relate to a specific number of vowel quality-related spectral peaks, and vowel perception cannot be approached within a spectral peak-picking concept.

In the experiment, the most apparent contradiction to formant merging as an explanation for different spectral peak numbers was observed concerning sounds of the back vowels /u, o/: As mentioned above, comparing sounds of /u/ or /o/ produced at similar f_o levels which manifested either only one or two (expected) spectral peak(s) showed that the first peak frequencies were similar but, for the sounds with only one peak, an (expected) second peak was "missing". This finding indicated that, for sounds of these two vowels, the second spectral peak might affect sound timbre but not vowel quality. Concerning sounds of front vowels, sounds with three expected vowel-related spectral peaks and sounds with only two peaks occurred for both cases of an "absent" second or third peak. This observation contradicted, in its turn, the concepts of formant merging or an *F*2-prime in between (expected) *F*2 and *F*3.

As for the notion that assumed spurious formants could explain a higher number of peaks than expected based on phonetic knowledge, sounds with prominent *H*1 preceding an expected low spectral peak, creaky sounds with three lower peaks, and whispered sounds of front vowels with two lower peaks < 1 kHz can hardly be attributed to rare effects of a speaker's individual production characteristics. Rather, they have to be accounted for as part of the occurring general spectral variation of vowel sounds. Besides, there may be cases of natural vowel sounds in support of the concept of formant merging or spurious formants. Still, these cases do not relativise the general objections made here.

In this context, the results of the vowel synthesis study of Ito et al. (2001) are worth noting. According to their results, the vowel quality of synthesised sounds of /i, e, a, o, u/ (Klatt synthesiser; f_o = 125 Hz) could be maintained even if *F*1 or *F*2 was suppressed, but the whole spectral shape remained unchanged. In their experiment, vowel quality confusion only occurred within the boundaries of two adjacent vowels.

Comparable to the results of many other experiments reported in this treatise, spectral peak number alterations proved to be nonuniform. This finding is illustrated in Figure 1 in the chapter appendix: Sounds 1 and 2 in the figure exemplify the observation that, for two sounds of /o/ produced at similar f_o levels but with different spectral peak numbers < 1.5 kHz, in most cases the first peak frequencies corresponded to each other (if vocal effort was not varied extensively). Sounds 3 and 4 exemplify that this was mostly untrue for corresponding sounds of /a/. Further, sounds 5 to 7 exemplify the observation that, if for the above two sounds of /o/, a third sound produced at a similar f_o but manifesting a spectral peak in between the two lower peaks found for the first sound is added to the comparison, the recognised vowel quality of this third sound is /ɔ/ (author's estimate). Finally, in contrast to sounds 3 and 4, sounds 8 and 9 exemplify the observation that, although rarely, the first peak frequencies of two sounds of /a/ produced at similar f_o levels but with different spectral peak numbers < 1.5 kHz may sometimes correspond to each other.

For many sounds whose peak patterns deviate from the general expectation, there is either only a weak or no methodological substantiation for formant pattern estimation. However, no detailed analysis of the estimation difficulties is discussed here since, as we argue, a future acoustic theory of the vowel principally cannot rely on formants.

Chapter appendix

Table 1. Compilation of voiced, breathy and creaky vowel sounds: Illustration of occurring vowel-related spectral peak number variation. Columns 1–5 = sounds (V = vowel intended and recognised; S/L = Series number and sound links; P = phonation type, where v = voiced, b = breathy, c = creaky; N = number of sounds; E = sound compilation the extracted sounds relate to). Column 6 = Content of the sound series (aspects of occurring spectral peak patterns).
[M-07-01-T01]

Table 2. Compilation of whispered vowel sounds: Illustration of occurring vowel-related spectral peak number variation. For the columns, see Table 1.
[M-07-01-T02]

Figure 1. Illustration of nonuniform spectral peak number variation for sounds of a vowel. Sounds 1–4 = a sound pair of /o/ produced at an intended f_o of 220 Hz with different numbers of lower spectral peaks, whose first peak frequencies correspond to each other, and a sound pair of /a/ produced at intended f_o of 247 and 262 Hz with different numbers of lower spectral peaks, whose first peak frequencies do not correspond to each other. Sounds 5–7 = two sounds of /o/ produced at an intended f_o of c. 220 Hz with either two or just one distinct spectral peak < 1.5 kHz, whose first peak frequencies correspond to each other (see sounds 1 and 2, again presented), and a third sound produced at this f_o level with a single peak at a frequency in between the two peaks found for the first sound, with vowel quality shifted to /ɔ/. Sounds 8 and 9 = a sound pair of /a/ produced at intended f_o of 147 and 220 Hz with different numbers of lower spectral peaks, whose first peak frequencies correspond to each other, contrary to the first sound pair of /a/ (see sounds 3 and 4).
[M-07-01-F01]

Table 1. Compilations of voiced, breathy and creaky vowel sounds: Illustration of occurring vowel-related spectral peak number variation. [M-07-01-T01]

Sounds						Content
V	S/L		P	N	E	
u	1	a ↗	v b c	52	–	Entire sample of selected voiced, breathy and creaky sounds.
		b ↗	v	4	a	Sound pairs illustrating present or absent expected second spectral peak. No "formant merging" indicated.
		c ↗	c	3	a	Creaky sounds, prominent energy in the frequency band of c. 0–250 Hz.
		d ↗	v	6	a	Sounds at higher fo with a single prominent H1 < 2 kHz.
o	2	a ↗	v b c	67	–	Entire sample of selected voiced, breathy and creaky sounds.
		b ↗	v	4	a	Sound pairs illustrating present or absent expected second spectral peak. No "formant merging" indicated.
		c ↗	c	11	a	One, two or three peaks < 1 kHz for creaky sounds.
		d ↗	v b	5	a	Dominant H1 and weak or "undetectable" first expected spectral peak.
		e ↗	v	3	a	Dominant H1 and one or two spectral peaks < 1 kHz.
a	3	a ↗	v b c	73	–	Entire sample of selected voiced, breathy and creaky sounds.
		b ↗	v c	10	a	Rippled spectra in the entire frequency range up to 1–1.5 kHz.
		c ↗	v b c	4	a	Dominant H1 or H1–H2 or a low frequency noise peak followed by two spectral peaks < 1.5 kHz.
		d ↗	v b	6	a	Dominant H1 or H1–H2 and only one peak < 1.5 kHz.
		e ↗	v	14	a	Only one peak < 1.5 kHz.
ɛ	4	a ↗	v b c	57	–	Entire sample of selected voiced, breathy and creaky sounds.
		b ↗	v b	5	a	Weak or barely definable first peak structure < 1 kHz, i.e., flat or sloping spectral energy, often extended over a frequency range exceeding 500 Hz and sometimes covering prominent spectral energy up to 1 kHz.
		c ↗	v b c	3	a	Dominant H1 or H1 and H2 for voiced or breathy sounds or a low spectral peak for creaky sounds preceding the expected first peak.
		d ↗	v b	4	a	Sound pairs illustrating present or absent expected second spectral peak.
		e ↗	v	4	a	Sound pairs illustrating present or absent expected third spectral peak.
		f ↗	v	4	a	Sounds at high fo, for which an expected spectral peak structure cannot be estimated. However, they sometimes indicate a first spectral energy maximum above 1 kHz.

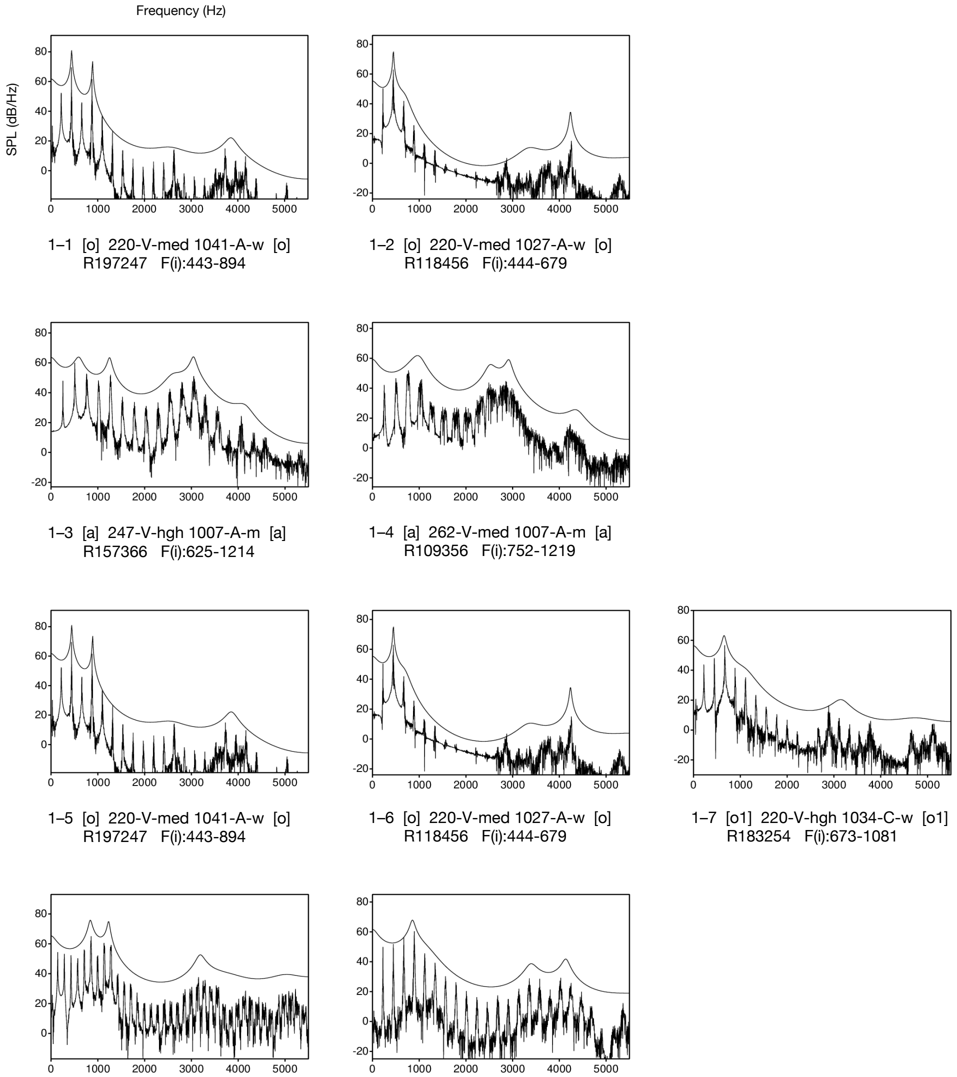
Table 1 (continuation). [M-07-01-T01]

Sounds						Content
V	S/L		P	N	E	
ø	5	a ↗	v b c	59	–	Entire sample of selected voiced, breathy and creaky sounds.
		b ↗	v b	3	a	Weak or barely definable first spectral peak structure < 1 kHz in terms of sloping spectral energy.
		c ↗	v b	4	a	Dominant H1 preceding the expected first spectral peak.
		d ↗	c	2	a	Two low spectral peaks for creaky sounds.
		e ↗	v	5	a	No separating peak structure for the frequency range of the first two expected spectral peaks.
		f ↗	v	6	a	Sound pairs illustrating present or absent expected second spectral peak.
		g ↗	v c	4	a	Sound pairs illustrating present or absent expected third spectral peak.
		h ↗	v b	2	a	Two sounds illustrating weak higher peak structure and low enegry in the corresponding frequency range.
e	6	a ↗	v b c	59	–	Entire sample of selected voiced, breathy and creaky sounds.
		b ↗	v b	4	a	Weak or barely definable first spectral peak structure < 1 kHz in terms of sloping spectral energy.
		c ↗	v	4	a	Dominant H1 or H1–H2 preceding the expected first peak.
		d ↗	c	3	a	Prominent rippled low spectral energy or two low spectral peaks for creaky sounds.
		e ↗	v	4	a	Sound pairs illustrating present or absent expected second spectral peak.
		f ↗	v	4	a	Sound pairs illustrating present or absent expected third spectral peak.
y	7	a ↗	v b c	67	–	Entire sample of selected voiced, breathy and creaky sounds.
		b ↗	v b	6	a	Sound pairs illustrating present or absent expected second spectral peak.
		c ↗	v	4	a	Sound pairs illustrating present or absent expected third spectral peak.
		d ↗	v	2	a	No relative spectral energy maximum < 1 kHz for sounds at high fo.
		e ↗	v	2	a	Extreme level variation for the first and second expected peaks.
i	8	a ↗	v b c	76	–	Entire sample of selected voiced, breathy and creaky sounds.
		b ↗	v	4	a	Sound pairs illustrating present or absent expected third spectral peak.
		c ↗	v	2	a	No relative spectral energy maximum < 1 kHz for sounds at high fo.
		d ↗	v	3	a	Extreme level variation for the first and second expected peaks.

Table 2. Compilations of whispered vowel sounds: Illustration of occurring vowel-related spectral peak number variation. [M-07-01-T02]

Sounds						Content
V	S/L		P	N	E	
u	1	a ↗	w	16	–	Entire sample of selected whispered sounds.
		b ↗	w	3	a	Rare cases with either one or three peaks < 1 kHz.
o	2	a ↗	w	33	–	Entire sample of selected whispered sounds.
		b ↗	w	10	a	One or two or three spectral peaks < 1 kHz (no “formant merging”).
a	3	a ↗	w	31	–	Entire sample of selected whispered sounds.
		b ↗	w	6	a	One or two or three spectral peaks or a frequency band of prominent spectral energy < 1.5 kHz.
ɛ	4	a ↗	w	41	–	Entire sample of selected whispered sounds.
		b ↗	w	4	a	“Split” spectral peaks (two or three peaks in the frequency range of expected single peaks).
ø	5	a ↗	w	44	–	Entire sample of selected whispered sounds.
		b ↗	w	5	a	Two peaks in the frequency range of expected single peaks.
e	6	a ↗	w	51	–	Entire sample of selected whispered sounds.
		b ↗	w	4	a	Two peaks in the frequency range of expected single peaks).
y	7	a ↗	w	40	–	Entire sample of selected whispered sounds.
		b ↗	w	2	a	Two peaks in the frequency range of expected single peaks.
		c ↗	w	2	a	Only one peak > 1 kHz in the expected vowel-related frequency range.
i	8	a ↗	w	46	–	Entire sample of selected whispered sounds.
		b ↗	w	3	a	Sounds with a barely defined spectral peak < 1 kHz.

Figure 1. Illustration of nonuniform spectral peak number alterations. [M-07-01-F01]



1–1 [o] 220-V-med 1041-A-w [o]
R197247 F(i):443-894

1–2 [o] 220-V-med 1027-A-w [o]
R118456 F(i):444-679

1–3 [a] 247-V-hgh 1007-A-m [a]
R157366 F(i):625-1214

1–4 [a] 262-V-med 1007-A-m [a]
R109356 F(i):752-1219

1–5 [o] 220-V-med 1041-A-w [o]
R197247 F(i):443-894

1–6 [o] 220-V-med 1027-A-w [o]
R118456 F(i):444-679

1–7 [o1] 220-V-hgh 1034-C-w [o1]
R183254 F(i):673-1081

1–8 [a] 147-V-hgh 1052-A-w [a]
R160980 F(i):823-1230

1–9 [a] 220-V-med 1052-A-w [a]
R140705 F(i):837-1104

M7.2 Inversions of Vowel-Related Relative Spectral Energy Maxima and Minima

Introduction

The second observation discussed in this main chapter concerns the occurrence of inversions of relative spectral energy maxima and minima for sounds of a given vowel in vowel-related frequency ranges.

As explained in the previous chapter, comparing sounds of the back vowels /u, o/ of the Zurich Corpus, the sounds produced at similar f_o levels with either two (expected) spectral peaks or only one single peak < 1.5 kHz showed that the first peak frequencies for those sounds were similar in most cases but, for the sounds with only one peak, the second peak was "missing". However, as discussed in the Preliminaries (see p. 62 and pp. 183–186), if sounds of these vowels produced at lower or middle levels of f_o and manifesting two spectral peaks were compared with other sounds produced at middle or higher f_o with only one spectral peak, inverse relative spectral maxima and minima in the form of inverse spectral envelope curves ≤ 1.5 kHz occurred, without any change in vowel recognition: Thus, whereas a relative energy minimum in between two peaks in the spectrum can be manifest for one sound of a vowel, a single spectral energy maximum can be manifest for another sound of that vowel. The same holds true for comparisons between the respective calculated filter curves and estimated formant patterns (if the estimation is methodologically substantiated). Similar observations were made for sounds of /a/, but they did not systematically relate to f_o variation. For sounds of front vowels, such inversions were, in their turn, observed for the vowel-specific frequency range > 1 kHz, but they were often related to marked vocal effort variations.

As was the case for different spectral peak numbers in vowel-related spectral frequency ranges, to document the possible variability of the vowel spectrum on the new basis of the Zurich Corpus and to embed it in the line of argument of this treatise, a corresponding study was conducted: Based on the inspection of the corpus, for back vowels and /a/, sound pairs of a vowel produced by speakers of different speaker groups were investigated for which the spectra manifested inversions of relative spectral energy maxima and minima in their vowel-related frequency ranges. On this basis, exemplary sound series were compiled for documentation and illustration in this treatise.

The study was limited to sounds of back vowels and /a/ because, for sounds of these vowels, occurring inversions concern the entire vowel-related spectral frequency range and, according to our previous experiences in the context of the investigation discussed in the Preliminaries, the inversions can often be observed not only for sounds with marked vocal effort variation but also for sounds produced without intended effort variation.

Experiment

Vowel sounds and speakers: Voiced sounds of the three long Standard German vowels /u, o, a/ of the Zurich Corpus produced by the speakers documented in the corpus in nonstyle mode at various f_o levels with medium vocal effort in V context were taken as the basis of investigation, for which the listening test conducted when creating the corpus provided a 100% recognition rate (matching vowel intention).

Inspection of sound spectra and selection of sound pairs: For each of the three vowels investigated, the occurrence of sound spectra with either two distinct peaks or one single peak < c. 1.5 kHz in the corpus was investigated, and a sample of numerous sound pairs was compiled for which the first sound manifested two distinct relative spectral energy maxima < 1.5 kHz and the second sound manifested a single distinct relative spectral energy maximum in between the two maxima of the first sound. On this basis, an exemplary documentation of the phenomenon was created: For each of the three vowels investigated, three sound pairs produced by men, women and children were selected. For each single sound pair, the first sound manifested two distinct relative spectral energy maxima < 1.5 kHz, including two single harmonics forming the tips of the peaks, and the second sound manifested a single distinct relative spectral energy maximum, including a single harmonic forming the peak tip, this single peak lying in between the two peaks of the first sound. As a result, a compilation of three sound pairs per vowel (9 sound pairs and 18 single sounds in total) was created.

Indications of spectral peak frequencies and peak distances: Spectral peak frequencies < 1.5 kHz and related peak frequency distances were determined based on the frequencies of the dominant harmonics that form the peak tips. Values were calculated in Hz and Bark (for the conversion algorithm, see Traunmüller, 1990).

Additional note: The inspection and comparison of sound spectra and the selection of exemplary sound pairs were again carried out based on previous experiences regarding the inversion phenomenon and were focused on finding examples best suited to documenting and illustrating the phenomenon in the context of the present treatise.

Results

Table 1 in the chapter appendix lists the selected sound pairs and presents the estimated peak frequencies and their interrelations.

With regard to the inspection of the Zurich Corpus, a large number of sound pairs with inverted spectral energy minima and maxima were found for each of the three vowels investigated. For the sound pairs of the two back vowels /u, o/, the sound manifesting two distinct relative spectral energy maxima was generally produced at a markedly lower f_o level than the sound with a single distinct relative spectral energy maximum in between the two maxima of the first sound. For the sound pairs of /a/, this was also often the case, although not in a systematic way.

With regard to the exemplary sound selection and documentation presented here, all sound pairs illustrate marked inversions of relative spectral energy maxima in their vowel-related spectral range since spectral energy maxima were represented by single dominant harmonics forming the tips of the peaks. For all sounds compared, the above f_o level differences were associated with the inversions of spectral maxima and minima.

For the three documented sounds of /u/ with two lower spectral peaks, the frequency distance between the two peak-related dominant harmonics was found as > 4.4 Bark (see Series 1–3); for the three sounds of /o/ with two lower spectral peaks, this frequency distance was found in the range of c. 3.1–3.4 Bark (see Series 4–6); for two out of the three sounds of /a/ with two lower spectral peaks, this frequency distance was found as > 4 Bark, and for the remaining sound, a 2.46 Bark frequency distance was found (see Series 7–9). Thus, for five of the nine sounds with two spectral peaks, the frequency distance between these peaks exceeded 3.5 Bark. Hence, the distance exceeded the frequency range discussed within the "centre of gravity" concept (see Chapter M7.1).

Note also that, for three pairs, both sounds compared were produced by a single speaker (see Series 1, 4 and 9).

Discussion

In the present examination of the sounds of /u, o, a/ of the Zurich Corpus, the comparison of two sounds with either two spectral peaks or only one peak in the vowel-related frequency range confirmed again that numerous sounds with a single peak < 1.5 kHz lying in between two peaks < 1.5 kHz of another sound of the same vowel occur, if substantial f_o level differences of the sounds (above all of the back vowels) are included in the investigation. This phenomenon is documented here anew in an exemplary manner. Further, such inversions of spectral maxima and minima cannot be explained by an auditory spectral averaging process for two reasons: In part, these spectral inversions occurred for sounds for which the frequency distance exceeded the 3–3.5 Bark limit (this limit often being assumed as the frequency range of the supposed averaging process), and in a resynthesis based on the spectral envelope of sounds of the back vowels with only one spectral peak < 1 kHz but applying the lower f_o of the two-peak counterexamples, vowel quality shifts occurred (for details, see below).

Concerning the sounds of /u, o/, it was shown in the previous chapter that if the f_o levels of sounds with two spectral peaks or only one peak were similar, the lowest peak related to vowel quality only. In the present study, however, two aspects interacted for the sounds of these vowels: The relation of the vowel sound to its f_o level (and pitch level) and the possible number alteration of vowel-related spectral peaks. Because of this interaction, a single spectral peak of one sound lying between two peaks of another can occur without changing vowel quality.

For sounds of /a/, the examples presented in Table 1 seem to point in a similar direction since the sounds of a pair with only one lower spectral peak were all produced at substantially higher f_o levels than those with two lower peaks. However, in the inspection of the Zurich Corpus, this kind of f_o difference occurred less systematically than for the sounds of the back vowels, and further examination involving LP filtering and Klatt synthesis showed that, for sounds of /a/, f_o and pitch do not play the same role regarding inversions as for sounds of /u, o/. To exemplify this observation, two tests were performed by the author. The author's estimate for the sounds in Table 1 is verifiable via the SpecFilt and KlattSyn tools in the Zurich Corpus: If the first sound of a pair of /u/ or /o/ (produced at the lower f_o level of the pair, see Table 1, Series 1–6) with two lower spectral peaks < 1.5 kHz was LP filtered with a CF in between these two peaks, the vowel quality as such did not change, confirming the notion put forward in the previous chapter that the first peak is of primary importance for the vowel qualities in question. Contrarily, for /a/, if the first sound of a pair (see Table 1, Series 7–9) with two lower spectral peaks < 1.5 kHz was LP filtered with a CF in between these two peaks, the vowel quality changed in an open–close direction. Furthermore, if the second sound of a pair of /u/

or /o/ was resynthesised (Klatt synthesis) at the lower f_o level of the first sound of that pair, the vowel quality changed in a close–open direction. Contrarily, for /a/, this type of resynthesis did not result in a vowel quality shift exceeding the vowel boundary of the original reference sound. This observation was in line with the finding that the relation of vowel sounds and their spectrum to the f_o level is pronounced for close and close-mid vowels but often weak or lacking for the open vowel /a/ (see the second and third main chapters).

Chapter appendix

Table 1. Compilation of sound pairs of the vowels /u/, /o/ and /a/: Illustration of occurring inversions of vowel-related relative spectral energy maxima and minima. Columns 1–5 = sounds (V = intended and recognised vowel quality; S/L = sound pairs and sound links; SG = speaker group, where m = men, w = women, c = children; SP = speaker ID in the Zurich Corpus; fo = calculated f_o, in Hz). Columns 6–9 = spectral peak frequencies < 1.5 kHz in terms of frequencies of dominant harmonics, and peak frequency distances (*D*1 and *D*2 = first and second spectral peaks for the first sound of a sound pair; *D*1* = single peak frequency of the second sound of a pair; *D*2-*D*1 = frequency distance between the two spectral peaks of the first sound of a sound pair; all values in Hz). Columns 10–13 = values for spectral peak frequencies and their distances in Bark.
[M-07-02-T01]

Table 1. Compilation of sound pairs of the vowels /u, o, a/: Illustration of occurring inversions of vowel-related relative spectral energy maxima and minima. [M-07-02-T01]

Sounds					Dominant harmonics (Hz)				Dominant harmonics (Bark)			
V	**S/L**	**SG**	**SP**	**fo (Hz)**	**D1**	**D1***	**D2**	**D2-D1**	**D1**	**D1***	**D2**	**D2-D1**
u	1 ↗	m	1051	258	258	–	774	516	2.59		7.06	4.47
			1051	521	–	521	–	–	–	5.10	–	–
	2 ↗	w	1004	259	259	–	777	518	2.60		7.08	4.48
			1032	527	–	527	–	–	–	5.15	–	–
	3 ↗	c	1057	382	382	–	1146	764	3.84	–	9.36	5.52
			1034	814	–	814	–	–	–	7.34	–	–
o	4 ↗	m	1069	196	392	–	784	392	3.94	–	7.13	3.19
			1069	289	–	578	–	–	–	5.58	–	–
	5 ↗	w	1004	189	378	–	756	378	3.80		6.93	3.13
			1071	262	–	524	–	–	–	5.13	–	–
	6 ↗	c	1098	220	440		880	440	4.39	–	7.78	3.39
			1058	325		650	–	–	–	6.15	–	–
a	7 ↗	m	1008	111	555	–	1221	666	5.39	–	9.76	4.37
			1063	449	–	898	–	–	–	7.89	–	–
	8 ↗	w	1088	217	651		1302	651	6.15	–	10.17	4.02
			1046	338		1014	–	–	–	8.61	–	–
	9 ↗	c	1056	215	860	–	1290	430	7.65	–	10.11	2.46
			1056	515	–	1030	–	–	–	8.71	–	–

M7.3 Flat or Sloping Vowel-Related Spectral Energy Distribution in Natural Vowel Sounds

Introduction

The third observation discussed in this main chapter concerns sounds with flat or sloping spectral energy distribution in the entire vowel-related frequency range or in the upper part of that range.

In early vowel synthesis experiments (using an early type of harmonic synthesiser), Carpenter and Morton (1962) and Morton and Carpenter (1962) showed that a stepwise increase in the number of harmonics from *H*1 (its frequency set to 180 Hz) to *H*10 with equal harmonic levels caused a step-by-step shift in vowel quality from /u/ to /o/ to /ɔ/ to /ɑ/ and finally to /a/. The same held true for a stepwise increase in the number of harmonics with decreasing harmonic levels (with a spectral slope). Since these sounds did not exhibit any spectral peaks and formant structure, the authors concluded that vowel quality recognition does not relate to the discrimination of spectral peaks, at least for back vowels and /a/.

Carpenter and Morton (1962) were also able to show that front vowels were recognised on the basis of the two lowest harmonics (their frequencies set to 180 and 360 Hz, with the second harmonic level decreased by approximately 18 dB) and a band of harmonics in the higher frequency range with the harmonic levels stepwise decreasing (with a spectral slope): The sound synthesised on the basis of *H*1–*H*2 and *H*10–*H*13 was labelled by a majority of trained phoneticians as /y/, and the sound synthesised on the basis of *H*1–*H*2 and *H*14–*H*40 was labelled by a majority of these trained phoneticians as /i/. Thus, the study also showed that sounds of front vowels are recognisable even if there is a lack of spectral peaks and formant structure in the frequency range > 1 kHz. Notably, Morton and Carpenter (1962) concluded that "[…] the selection of harmonics according to formant theory is not the only, and perhaps not always the best means of synthesizing isolated vowels […]".

Note in this context that Dubno and Dorman (1987) showed recognisable synthesised sounds of front vowels with only *F*1 as a well-specified spectral peak, combined with a broad higher frequency region of energy with no marked peaks (Klatt synthesis as indicated in the figures presented, sounds produced at f_o of 100 Hz). They concluded that when the first formant of a vowel is "[…] well specified, the presence of a broad region of energy in the higher frequencies is sufficient for normal-hearing listeners to correctly identify front vowels".

Concerning natural sounds, in the Preliminaries (pp. 57–58 and 147–157), we have discussed and documented cases of sounds of back vowels and /a/ whose spectra exhibited a series of harmonics with quasi-identical or with continuously decreasing amplitudes in the lower frequency range < c. 1.5 kHz (flat or sloping spectral energy distribution in vowel-related frequency ranges). We have also discussed and documented cases of sounds of front vowels that manifested series of harmonics with quasi-identical amplitudes in the higher frequency range > c. 1.3 kHz (flat spectral energy distribution in the vowel-related portions of higher frequency ranges).

In a subsequent investigation of vowel recognition for synthesised sounds based on series of harmonics equal in amplitude, recognisable vowel sounds related to flat harmonic configurations were found for all long Standard German vowels except /u/, although with different recognition rates for the different vowels (Maurer and Suter, 2017a; f_o of the investigation was 200 Hz). As reported in Chapter M6.7, investigating vowel recognition for synthesised sounds based on series of equal amplitude harmonics > 1 kHz combined with a single lower harmonic < 1 kHz, such kind of harmonic configurations also proved to be related to recognisable front vowel sounds (Maurer et al., 2017b; the range of f_o of the investigation was 150–600 Hz; note that this finding was somewhat comparable to the above results of Dubno and Dorman).

As was the case in the previous chapters, to once again document the possible variability of the vowel spectrum on the new basis of the Zurich Corpus and to embed it into the line of argument of this treatise, a corresponding new study was conducted addressing the documentation of exemplary sound compilations of flat or sloping spectral energy distribution in either the entire or the higher part of vowel-related frequency ranges of natural vowel sounds: Inspecting the Zurich Corpus, voiced and breathy sounds of the eight long Standard German vowels produced by speakers of different speaker groups were investigated for which the spectra manifested flat or sloping spectral energy distribution. On this basis, for each vowel, a larger sound sample and a short extract thereof with exemplary cases were compiled for documentation and illustration.

Experiment

Vowel sounds and speakers: Sounds of all eight long Standard German vowels of the Zurich Corpus produced by the speakers of all speaker groups documented in the corpus with voiced or breathy phonation and including the variation of the production parameters f_o, vocal effort, vowel context (V and sVsV context) and production style were taken as the basis of investigation.

Inspection of sound spectra and sound selection: For each vowel, the occurrence of flat vowel-related sound spectra was investigated, that is, spectra which manifested either series of harmonics with quasi-identical or with continuously decreasing levels throughout the entire vowel-related frequency range (all vowels) or series of harmonics with quasi-identical or with continuously decreasing levels in the higher frequency range > c. 1.3 kHz (front vowels). In the course of this inspection of the corpus, a larger sample of numerous sounds with these types of spectra was compiled. All the selected sounds were fully recognised in the standard listening test conducted when creating the corpus (100% recognition rate matching vowel intention). On this basis, a few cases per vowel were selected in terms of exemplary representations of the phenomenon in question. As a result, for each vowel, a larger sound compilation and a short extract of it were created (for numerical indications, see Table 1 in the chapter appendix).

Additional note: The inspection of sound spectra and the selection of exemplary sounds for the present documentation were again made based on previous experiences regarding flat or sloping vowel spectra and were focused on finding examples best suited to documenting and illustrating the phenomenon in the context of the present treatise.

Results

Table 1 in the chapter appendix lists the sound samples compiled and presents extracts of these compilations in terms of a few examples illustrating the main types of the observed spectral energy configurations discussed below. In the Zurich Corpus, a high number of sounds were found exhibiting both flat and sloping spectra or spectral parts in vowel-related frequency ranges. Only some of these examples were selected for the present documentation in order to demonstrate the variation of this kind of spectral manifestation.

For the sounds of /u, o, a/, two main types of either flat or sloping energy distribution < c. 1.5 kHz were observed: As a tendency, for the lower range of f_o < c. 250 Hz, the selected sounds related to low vocal effort in voiced phonation or to breathy phonation. For the frequencies above this range, no vocal effort-specific relation of the selected sounds was manifest.

For the sounds of /ɛ/, /ø/ and /e/, dependent on the f_o level of the sounds, three main types of spectral manifestations were observed for the vowel-related spectral frequency range: A spectral peak or prominent frequency band in the lower frequency range and a flat energy distribution in the higher frequency range; only flat energy distribution; or only sloping energy distribution.

For the sounds of /y/ and /i/, also dependent on the f_o level of the sounds, two main types of spectral manifestations were observed for the vowel-related spectral frequency range: A spectral peak in the lower frequency range associated with flat energy distribution in the higher frequency range or only flat energy distribution.

Discussion

The inspection of the Zurich Corpus indicated a high number of sounds without manifest, distinct spectral peaks (relative spectral energy maxima) in the frequency ranges assumed to be vowel-related, and the sound compilations presented in this chapter illustrate this phenomenon. Thus, flat or sloping energy distribution in a vowel spectrum proved not to be a rare phenomenon of vowel sounds, and it was not limited to a specific type of vowel production. In conclusion, the phenomenon in question can be expected to occur in any investigation of a large sound sample that includes an extensive variation of basic production parameters.

Given the previous analyses in the Preliminaries and the above results of vowel synthesis experiments, finding flat or sloping vowel spectra for sounds of /o/ and /a/ was to be expected. The same holds true for finding a peak in the lower part of the spectrum and flat energy distribution in the higher part, as observed for the sounds of front vowels. Additional observations of entirely flat or sloping spectra at higher f_o, as documented here for sounds of front vowels (see also Chapter M2.2), provide a further example of occurring types of vowel spectra with no distinct peak structures. (On the matter, see also the extensive citation of Ito et al., 2001, in Chapter M8.1.)

In this context, examples of recognisable synthesised vowel sounds based on series of harmonics equal in amplitude, as documented in the study of Maurer and Suter (2017a), have to be included when considering the matter. Since they were not discussed in detail in the previous chapters, they will be presented in the following chapter.

As was the case for vowel sounds with different vowel-related spectral peak numbers, methodological substantiation for formant pattern estimation for vowel sounds with flat or sloping spectral energy distribution often proved to be weak or lacking. Further, for some comparisons of sounds of different vowels, the vowel-related spectral difference was barely understood based on existing phonetic knowledge. (For an exemplary illustration, see Part II, Chapter 7.3, Figure 4.)

Finally, an early valuation of Carpenter and Morton (1962) when discussing sounds with flat or sloping spectra or higher spectral parts shall be cited: "Thus far it seems possible to claim that complex sounds can have a fairly consistent vowel quality, even when there are no peaks in the harmonic structure. […] The fact that human vowels are, for any individual, reasonably discriminable on analysis in terms of formant positions alone, is a phenomenon relating to the method of production of the vowels, and it does not follow that the mechanism of speech recognition proceeds in a similar way." This early questioning of production and perception as not related to each other in a simple and direct (unmediated) manner is important to note. The question may prove to be at the core of a future acoustic theory of the vowel. We will return to this matter below.

Chapter appendix

Table 1. Compilation of natural sounds of the long Standard German vowels: Illustration of occurring flat or sloping spectral energy distribution in vowel-related frequency ranges. Columns 1–4 = sounds (S/L = sound pairs and sound links; V = intended and recognised vowel quality; fo = minimum and maximum of calculated f_o of the sounds, in Hz; N = number of sounds). Column 5 = content of the sound series (aspects of occurring flat or sloping vowel-related spectral energy distribution).
[M-07-03-T01]

Table 1. Compilations of natural sounds of the long Standard German vowels: Illustration of occurring flat or sloping spectral energy distribution in vowel-related frequency ranges. [M07-03-T01]

Sounds						Content
S/L		V	fo (Hz)		N	
			min	max		
1	a ↗	u	142	671	49	Entire sample of selected sounds
	b ↗		166	254	3	Examples (extract)
2	a ↗	o	139	588	49	Entire sample of selected sounds
	b ↗		168	294	3	Examples (extract)
3	a ↗	a	140	527	35	Entire sample of selected sounds
	b ↗		140	438	4	Examples (extract)
4	a ↗	ɛ	239	822	32	Entire sample of selected sounds
	b ↗		251	674	6	Examples (extract)
5	a ↗	ø	205	790	43	Entire sample of selected sounds
	b ↗		205	625	3	Examples (extract)
6	a ↗	e	220	878	35	Entire sample of selected sounds
	b ↗		265	757	3	Examples (extract)
7	a ↗	y	246	923	36	Entire sample of selected sounds
	b ↗		263	923	3	Examples (extract)
8	a ↗	i	203	927	32	Entire sample of selected sounds
	b ↗		203	546	3	Examples (extract)

M7.4 Flat Vowel-Related Spectral Energy Distribution in Synthesised Vowel Sounds

Introduction

In the context of discussing vowel-related flat or sloping spectral energy distribution, the vowel recognition of correspondingly synthesised sounds is worth considering: As mentioned in the introduction of the previous chapter, early synthesis experiments already demonstrated sounds of back vowels and of /a/ with flat spectra in terms of series of consecutive harmonics equal in amplitude or with sloping harmonic levels. Further, in a more recent study (Maurer and Suter, 2017a) addressing the question of vowel recognition for synthesised sounds with entirely flat harmonic configurations (vowels synthesised based on series of harmonics equal in amplitude in various frequency bands, with reference *H*1 frequency = 200 Hz), we have shown examples of recognisable sounds of this type for all long Standard German vowels except /u/ (recognisable sounds of /u/ related to only a single harmonic). Since sounds of this type are important to consider for the extent of the vowel-related spectral variation, the study and its results are described below, and sound examples in terms of "best cases" are given.

Experiment

Harmonic configurations investigated: Three different types of harmonic series were investigated and used for subsequent vowel synthesis and vowel recognition tests:

- Type 1 (LP filter-like) = harmonic series resulting from a stepwise increase in the number of consecutive harmonics from *H*1 to *H*1–*H*20.
- Type 2 (HP filter-like) = harmonic series resulting from a stepwise decrease in the number of consecutive harmonics from *H*1–*H*20 to *H*20.
- Type 3 = (BP filter-like) harmonic series resulting from a stepwise increase in the number of consecutive harmonics from a middle harmonic to a series of 11 harmonics at a maximum (e.g., from *H*2 to *H*12, or from *H*3 to *H*13 and so forth, with the last series being from *H*15 to *H*20).

Harmonic frequencies were always multiples of 200 Hz (*H*1 frequency was the reference) and were equal in amplitude. As a result, a sample of 190 harmonic configurations was created.

Vowel synthesis: Monotonous sounds of 1.2 sec. with a 0.1 sec. fade in/out were produced using the SinSyn tool (all phases set to 0). As a result, a sample of 190 synthesised sounds was created.

Listening test: Vowel recognition of the synthesised sounds was tested in a listening test according to the standard procedure of the Zurich Corpus and involving the five standard listeners (forced choice, excluding vowel boundaries but including "no quality recognised"; note that, according to the standard procedure, the vowel /ɔ/ was included as a labelling option). The test was subdivided into subtests according to the three types of harmonic configurations and, for the configurations of type 3, also according to the harmonic series related to a given initial harmonic.

Results

The entire sample of the harmonic configurations and the vowel confusion matrix resulting from the listening test are given in Maurer and Suter (2017a; see additional materials). According to the listening test results, configurations related to recognisable sounds were found for all long Standard German vowels and /ɔ/. However, the number of harmonic configurations and synthesised sounds per vowel, as well as the related recognition rates, varied strongly, with the most opposite findings found for the comparison of sounds of /ɛ, a/ with those of /ø/: Numerous sounds related to different harmonic configurations were recognised as /ɛ/ or /a/ by all listeners, but only one sound related to a single harmonic configuration was recognised as /ø/ with a weak labelling majority (3/5 listeners).

For this treatise, the sounds recognised as the same vowel by all five listeners were selected from the entire sample investigated and are presented in Tables 1 (sounds per type of harmonic series) and 2 (sounds per vowel quality) in the chapter appendix. The sound compilation reflects the above two main findings: Recognisable vowel sounds could be synthesised based on entirely flat harmonic series, but the number of such sounds varied markedly for different vowel qualities. (Note that all sounds recognised as /u/ were related only to a single harmonic.)

Besides, when analysing the listening results of the entire sample shown in Maurer and Suter (2017a, additional materials, confusion matrix), these results indicated that some sounds were recognised inconsistently by single listeners due to either an ambiguous sound quality or the presentation context of the listening test. For example, sounds with only one harmonic were unanimously recognised as /u/ for the harmonic frequencies of 200, 600 and 800 Hz, but for 400 Hz, only four listeners labelled /u/. Similarly, the sound related to the harmonic configuration *H*1–*H*20 (type 1) was recognised by four listeners as /ɛ/ when testing the sounds of type 1 but only by two listeners when testing the sounds of type 2.

Discussion

As the main result, the experiment confirmed earlier indications that it is possible to synthesise recognisable vowel sounds based on entirely flat harmonic spectra in terms of series of consecutive equal amplitude harmonics in various frequency bands. However, for the sounds investigated, the number of clearly recognisable sounds was found to strongly relate to vowel qualities. The most impressive sounds and sound spectra were those of the vowels /a/ and /ɛ/ since a 100% recognition rate was found for several sounds of these vowels and since the related frequency bands of equal amplitude harmonics were large and opposed to any concept of spectral peak structure.

The strong variation in the number of clearly recognised sounds for different vowel qualities is difficult to interpret since only harmonics as multiples of 200 Hz were investigated. Future studies should address the question of the role of reference H_1 (and corresponding HCF) in this type of experimentation. However, to give a first indication, additional synthesised sounds applying reference H_1 (and corresponding HCF) ≤ 150 Hz in synthesis are presented in Table 3 in the chapter appendix: According to the author's estimate, these sounds demonstrate flat harmonic configurations for recognisable sounds of /u/ (with f_o and HCF of 100 Hz), /o/ (with f_o and HCF of 125 and 150 Hz), /ø/ (with f_o and HCF of 100 Hz) and /e/ (with f_o and HCF of 150 Hz). For sounds of these vowels, when investigating harmonic synthesis related to reference H_1 (and corresponding HCF) of 200 Hz, either no sound with a flat harmonic spectrum (sounds of /u/) or no sounds with a 100% recognition rate (sound of /ø/) or only one sound with a 100% recognition rate (sounds of /e/ and /o/) were found. Thus, it can be expected that different reference H_1 and corresponding HCF – and also different recognised pitch levels – impact the results of a synthesis experiment of this type.

Notably, the findings of the present experiment and the related phenomena of synthesised sounds based on flat harmonic series > 1 kHz combined with a single lower harmonic < 1 kHz (see Chapter M6.7) were in line with the results of early synthesis studies and studies of suppressed *F*2 and they again supported the notion that a spectral peak structure is not an imperative prerequisite of vowel recognition.

Chapter appendix

Table 1. Compilation of synthesised sounds of the long Standard German vowels and /ɔ/: Illustration of occurring flat spectral energy distribution in vowel-related frequency ranges. Sounds with a recognition rate of 100% are shown. Column 1 = type of harmonic configuration (T) and sound links (L). Column 2 = reference numbers of the sounds in the Zurich Corpus (Ref). Column 3 = recognised vowel quality (V). Column 4 = number of sounds (N). Column 5 = harmonics used in synthesis (*H*(i), with *H*1 reference frequency = 200 Hz). Column 6 = frequency range of the harmonics used in synthesis (FR).
[M-07-04-T01]

Table 2. Compilation of synthesised sounds of the long Standard German vowels and /ɔ/: Summary of sounds per vowel as listed in Table 1. Column 1 = sound series and sound links (S/L). Column 2 = recognised vowel quality (V). Column 3 = number of sounds (N).
[M-07-04-T02]

Table 3. Additional attempts to synthesise recognisable sounds related to flat harmonic spectra for the vowels /u, o, ø, e/. For details, see the discussion section of this chapter. Column 1 = sounds and sound links (S/L). Column 2 = vowel quality (V; author's estimate). Column 3 = harmonics used in synthesis (*H*(i); for *H*1 reference frequencies, see text). Column 4 = frequency range of the harmonics used in synthesis (FR).
[M-07-04-T03]

Table 1. Compilation of synthesised sounds of the long Standard German vowels and /ɔ/: Illustration of occurring flat spectral energy distribution in vowel-related frequency ranges. [M-07-04-T01]

T/L	Ref	V	N	H(i)	FR (Hz)
1 ↗	180699	u	1	1 (only)	200
	180705	a	2	1–7	200–1400
	180706			1–8	200–1600
	180711	ɛ	1	1–13	200–2600
2 ↗	180735	i	2	17–20	3400–4000
	180737			19–20	3800–4000
3 ↗	180750	u	2	3 (only)	600
	180761			4 (only)	800
	180741	o	1	2–4	400–800
	180762	ɔ	3	4–5	800–1000
	180751			3–4	600–800
	180752			3–5	600–1000
	180744	a	7	2–7	400–1400
	180754			3–7	600–1400
	180763			4–6	800–1200
	180774			5–7	1000–1400
	180764			4–7	800–1400
	180775			5–8	1000–1600
	180765			4–8	800–1600
	180759	ɛ	8	3–12	600–2400
	180768			4–11	800–2200
	180778			5–11	1000–2200
	180758			3–11	600–2200
	180771			4–14	800–2800
	180779			5–12	1000–2400
	180781			5–14	1000–2800
	180780			5–13	1000–2600
	180844	e	1	11–17	2200–3400
	180816	y	2	9 (only)	1800
	180829			10–12	2000–2400
	180867	i	2	14–16	2800–3200
	180874			15–17	3000–3400

Table 2. Compilation of synthesised sounds of the long Standard German vowels and /ɔ/: Summary of sounds per vowel as listed in Table 1. [M-07-04-T02]

S/L	V	N
1 ↗	u	3
2 ↗	o	1
3 ↗	ɔ	3
4 ↗	a	9
5 ↗	ɛ	9
6	ø	0
7 ↗	e	1
8 ↗	y	2
9 ↗	i	4

Table 3. Additional attempts to synthesise recognisable sounds related to flat harmonic spectra for the vowels /u, o, ø, e/. [M-07-04-T03]

S/L	V	H(i)	FR (Hz)
1 ↗	u	1–3	100–300
2 ↗	o	2–5	250–625
3 ↗	o	2–4	300–600
4 ↗	ø	13–18	1300–1800
5 ↗	e	12–17	1800–2550

M7.5 Sounds of Close and Close-Mid Vowels for Which Marked f_o Variation < 250 Hz Does Not Affect Estimated Formant Patterns and Spectral Envelopes

Introduction

In the context of discussing the nonuniform character of spectral variation when observing sounds of a given vowel, a further observation concerns the nonuniform relation of the lower vowel spectrum to f_o for different frequency ranges of f_o variation.

As indicated in the Preliminaries (p. 159) and discussed in Chapter M2.1, vocalises of close vowels showed a marked relation of the lower vowel spectrum to f_o (spectral envelope and peaks, frequency ranges of prominent spectral energy) but only from f_o levels above c. 200–300 Hz (depending on vowel quality). Correspondingly, as shown in Chapter M3.3, the same held true for f_o variation in vowel synthesis resulting in formant pattern and spectral shape ambiguity (see also f_o levels and ranges of the comparison of natural vowel sounds with ambiguous *F*-patterns and spectral envelopes in Chapter M3.5). Taking, in addition, our general experiences of the extensive inspection of the Zurich Corpus into account, we assumed that, for sounds of close and close-mid vowels, formant pattern and spectral shape ambiguity may generally occur for an approximately one-octave (or more) increase in f_o, if (and only if) the higher f_o levels of comparison are ≥ c. 300 Hz. (However, for some exceptions of synthesised sounds, see Chapters M3.1 and M3.2.; see also Chapter M7.7 for cases of sounds of open-mid and open vowels that indicate possible vowel quality shifts due to f_o variation including an upper f_o frequency of 300 Hz, and Chapter M7.8) In contrast, no differences in the spectral energy distribution and the related maxima may be manifest for one-octave differences of f_o if all f_o levels of compared sounds are below c. 250 Hz. To some extent, we expected that the same holds true for sounds of close-mid vowels, above all, if all f_o levels of the compared sounds are below c. 200 Hz. (Note in this context that, when summarising observations of the lacking influence of low f_o < 150 Hz on vowel openness, Traunmüller, 1988, pointed to an "anomalous influence of low f_o" as an "end of scale effect".)

To document the nonuniform relation of the lower vowel spectrum to f_o, in an earlier study, we presented a sample of six sound pairs of close and close-mid vowels, each sound pair produced by a single male speaker at two different f_o levels approximately one octave apart, with all f_o levels being ≤ 250 Hz (see Maurer et al., 2019, Chapter M7.3.1 in the online presentation): In contrast to sound comparisons including higher f_o levels (substantially surpassing 250 Hz), no marked change in the vowel spectrum was indicated due to increasing f_o in these examples. However, variations in vocal effort and production style were disregarded for these pairs. Against this background, this earlier investigation and documentation of the nonuniform relation of the lower vowel spectrum to f_o was renewed for the present treatise, restricting the production parameters of the sounds to nonstyle mode and medium vocal effort: Inspecting the Zurich Corpus for close and close-mid vowels, sound pairs of a vowel produced by men were investigated for which the f_o levels differed by approximately one octave or more, but with all f_o levels ≤ 250 Hz. On this basis, exemplary sound pairs were compiled for documentation and illustration in this treatise.

Experiment

Vowel sounds and speakers: Voiced sounds of the long close and close-mid Standard German vowels /i, y, u/ and /e, ø, o/ of the Zurich Corpus produced by men in nonstyle mode with medium vocal effort in V context at f_o ≤ 250 Hz were taken as the basis of investigation, for which the listening test conducted when creating the corpus provided a 100% recognition rate (matching vowel intention).

Inspection of sound spectra and sound selection: Sound pairs and related spectra of individual speakers were inspected, and a sample of numerous pairs per vowel was compiled for which the f_o levels differed by approximately one octave or more, but the spectral envelope and the estimated F-patterns < 1 kHz did not indicate marked differences. As a further selection criterion, for each of the two sounds of a pair, vowel quality was investigated in resynthesis using the KlattSyn tool, resynthesis based on the estimated F-pattern of a sound but applying both f_o levels of a pair. Sounds were selected for which no marked vowel quality shift occurred in resynthesis for both f_o levels applied (author’s estimate). On this basis, an exemplary documentation and illustration of the phenomenon was created for this treatise in the form of one sound pair per vowel, resulting in six exemplary pairs. For this reduced sound sample, vowel recognition of the resynthesised sounds applying both f_o levels of a pair was further investigated in a listening test involving the five standard listeners (see below).

Estimation of *F*-patterns: The estimation of F-patterns accorded to the standard acoustic analysis of the Zurich Corpus and included a visual crosscheck of the calculated F-patterns based on the spectrum, spectrogram and formant tracks. F_1–F_2 were estimated for the sounds of the back vowels, and F_1–F_2–F_3 were estimated for the sounds of the front vowels. If discontinuous formant tracks occurred, LPC parameters were lowered until continuous tracks were obtained. (Note that for two sounds, F_2 or F_3 estimation was problematic; see the corresponding calculated values given in parentheses in Table 1 in the chapter appendix.)

Additional note: Inspection and comparison of sound spectra and F-patterns, the first examination of resynthesised sounds and the subsequent selection of exemplary sounds for the present documentation were focused on finding examples best suited to documenting and illustrating the phenomenon of a lack of vowel quality shifts due to f_o variation for $f_o \leq 250$ Hz for sounds of close and close-mid vowels in the context of the present treatise.

Klatt resynthesis experiment for the six selected sound pairs: Using the KlattSyn tool (default parameters, 1 sec. sound duration including a 0.05 sec. fade in/fade out), every single natural sound was resynthesised based on its estimated F-pattern but applying both f_o of the sound pair it belonged to. As a result, a sample of 24 resynthesised sounds was created.

Listening test: The vowel recognition of the resynthesised sounds was tested in a listening test according to the standard procedure of the Zurich Corpus and involving the five standard listeners (forced choice, excluding vowel boundaries but including schwa), with an additional test specification: Each prompt consisted of the original natural sound followed by one of the resynthesised sounds (separated by a 0.5 sec. pause), and the listeners were asked to label the vowel quality of the second sound.

Results

Table 1 in the chapter appendix lists the selected sound pairs and estimated F-patterns and shows the vowel recognition results for the resynthesised sounds. According to the estimated F-patterns, no distinct spectral differences < 1 kHz occurred for the selected sound pairs /i, y, u/ and /e, ø, o/ despite an f_o variation of approximately one octave or more, with all f_o levels of the first sound of a pair being below or equal to c. 110 Hz and all calculated f_o levels of the second sound of a pair being below or equal to c. 250 Hz: For all sound pairs, differences in F_1 were found as < 50 Hz; the same held true for F_2 of the two pairs of sounds of the back vowels and for three of the four pairs of sounds of front vowels; the differences in F_2 for the remaining pair was found as 111 Hz; for three of the four sound pairs of front vowels, differences in F_3 were also found as ≤ 50 Hz, and for the remaining sound pair, the difference was 130 Hz. In parallel, according to the labelling majority, no marked vowel quality shifts were found for the resynthesised sounds when the f_o of one sound of a pair was substituted with that of the other, neither for an upward nor for a downward f_o shift of approximately one octave or more.

Discussion

As demonstrated extensively in this treatise, formant pattern and spectral shape ambiguity occur almost regularly for sounds of close and close-mid vowels if the f_o levels of the sounds compared are approximately one octave (or more) apart and if the higher f_o levels of sound comparison are above 300 Hz. The actual values depend on the vowel quality in question. Thereby, the ambiguity phenomenon is primarily a consequence of the relation of the lower vowel spectrum to f_o.

On the contrary, when inspecting the Zurich Corpus, very similar vowel-related spectral peaks, calculated *F*-patterns and entire spectral envelopes occurred for most of the sounds of close and close-mid vowels produced in nonstyle mode with a medium vocal effort at f_o levels that were approximately one octave (or more) apart and were below or equal to c. 250 Hz – more precisely, below or equal to c. 200 Hz for sounds of close-mid vowels, and below or equal to c. 250 Hz for sounds of close vowels. The selected examples illustrate this finding. Thus, the relation of the lower vowel spectrum to f_o proved to be nonuniform with regard to the frequency range of f_o: For the sounds presented, a one-octave shift of f_o (upward or downward) in the lower frequency range mentioned did not result in a distinct vowel quality shift, in stark contrast to f_o variation including higher frequency ranges.

Chapter appendix

Table 1. Sound pairs of close and close-mid vowels produced by single male speakers at $f_o \leq 250$ Hz: Illustration of comparable vowel-related lower spectral characteristics despite an f_o difference of one octave or more. Columns 1–4 = sounds (S/L = sound series and sound links; SP = speaker ID in the Zurich Corpus; V = intended and recognised vowel quality of the natural sounds; fo = calculated f_o, in Hz). Columns 5–11 = estimated formant frequencies (F(i)), related frequency differences (ΔF1) and LPC parameter setting applied in resynthesis (Par). Note that, for two sounds, F_2 is given in parentheses in order to indicate estimation problems (see sound spectrogram and calculated formant tracks). Columns 12 and 13 = vowel recognition of the resynthesised sounds for unaltered f_o levels of the natural reference sounds (Maj = labelling majority; details = the five single labellings). Columns 14 and 15 = vowel recognition of the resynthesised sounds for altered f_o levels, applying f_o of the opposing sound of a pair. Colour code: Blue = formant frequency differences ≤ 50 Hz and/or equal vowel recognition for the two sounds of a pair. For a cross-examination of the resynthesis, use the KlattSyn tool in the corpus, taking into account the parameter settings for the LPC analysis indicated in Column 11.
[M-07-05-T01]

Table 1. Sound pairs of close and close-mid vowels produced by single male speakers at fo ≤ 250 Hz: Illustration of comparable vowel-related lower spectral characteristics despite an fo difference of one octave or more. [M-07-05-T01]

Sounds				F-patterns (Hz)							Vowel recognition in resynthesis			
S/L	SP	V	fo (Hz)	F1	ΔF1	F2	ΔF2	F3	ΔF3	Par	fo org		fo exchanged	
											Maj	details	Maj	details
1 ↗	1002	i	95	269	7	2325	33	2965	12	P6	i	i i i i e	i	i i i i i
			250	262		(2358)		2953		P6	i	i i i i i	i	i i i i e
2 ↗	1002	y	83	231	20	1996	41	2323	130	P6	y	y y y y y	y	y y y y y
			219	251		1955		2453		P6	y	y y y y y	y	y y y y y
3 ↗	1002	u	98	255	10	737	12	–	–	P6	u	u u u u u	u	u u u u u
			249	245		749		–		P5	u	u u u u u	u	u u u u u
4 ↗	1077	e	95	353	46	2014	35	2618	50	P6	e	e e e e e	e	e e e e e
			204	399		2049		2568		P5	e	e e e e ø	e	e e e e ɛ
5 ↗	1069	ø	99	396	22	1650	111	2195	22	P6	ø	ø ø ø ø ø	ø	ø ø ø ø ø
			196	418		1761		(2173)		P6	ø	e ø ø ø ø	ø	e ø ø ø ɛ
6 ↗	1069	o	109	411	7	831	41	–	–	P5	o	o o o o ɔ	o	o o o o o
			196	418		872		–		P5	o	o o o o o	o	o o o ɔ ɔ

M7.6 The Role of Vowel Quality With Respect to the Relation of the Lower Vowel Spectrum to f_o

Introduction

As discussed in the second and third main chapters, the relation of the lower vowel spectrum to f_o – and, with it, formant pattern and spectral shape ambiguity – does not only differ for different frequency ranges of f_o variation but also for different vowel qualities and the individual course of the spectral envelope or the harmonic configuration of sounds of a vowel. These two additional aspects of the nonuniform character of the vowel spectrum are brought into focus in this chapter and the following one.

Concerning the first aspect, when investigating vocalises, the relation of the lower vowel spectrum to f_o proved to be dependent on vowel qualities (see Chapter M2.1; see also some indication in Maurer et al., 2019, Chapter M7.3.2 in the online presentation), with the most pronounced differences observed when comparing sounds of close and open vowels. Similarly, formant pattern and spectral shape ambiguity occurred far less often for sounds of the open vowel /a/ than for close and close-mid vowel sounds (see Chapter M3.5, and Chapter 3.6 in Part II for a conclusion). In order to exemplify the nonuniform relation of the lower vowel spectrum to f_o concerning vowel quality and its impact on formant pattern and spectral shape ambiguity, a corresponding documentary study was conducted: Sound pairs of /u/ and sound pairs of /a/ produced by single speakers were compiled, with a difference in f_o levels of one octave at a minimum for the two sounds of a pair and the higher f_o levels of sound comparison exceeding 400 Hz, and with pronounced lower spectral differences with increasing f_o occurring only for the pairs of /u/ but not for the pairs of /a/. Further, this nonuniform impact of f_o variation was also investigated in Klatt resynthesis.

Experiment

Vowel sounds and speakers, and sound selection: Based on the sounds presented in Chapter M2.1 (see Table 1 in that chapter, vocalises of voiced sounds produced in nonstyle mode with medium vocal effort in V context), for each of the three speakers (man, woman and child) and for the close vowel /u/ and the open vowel /a/, two sounds were selected fulfilling two conditions for comparison: the f_o difference of the sounds of a pair was one octave at a minimum, and the f_o level of the higher sound of a pair was > 400 Hz. In addition, in the selection process for sounds of /a/, high spectral similarity of the two sounds of a pair was attempted. As a result, a sample of six sound pairs (12 natural reference sounds in total) was created.

Estimation of *F*-patterns: For all selected natural sounds, *F*-patterns were estimated according to the standard acoustic analysis of the Zurich Corpus, with speaker group-specific default LPC parameters (P6 for men, P5 for women, P4 for children) and including a visual crosscheck of the calculated formant frequency values based on the spectrum, the spectrogram and the formant tracks. If discontinuous formant tracks occurred, the LPC parameters were altered until continuous tracks were obtained (see Table 1 in the chapter appendix, indication for LPC parameters). Note that formant frequencies were also estimated for sounds at higher f_o levels despite the methodological estimation problems.

Visual comparison of the spectra: The two spectra of the natural reference sounds of a pair were compared visually, and the spectral envelope differences < 1.5 kHz (differences in the spectral energy distribution in general and spectral peaks in particular) were labelled as either marked or marginal.

Klatt resynthesis experiment: Each natural reference sound was resynthesised based on its estimated *F*-pattern and calculated average f_o using the KlattSyn tool (default parameters). In addition, the sound with the higher f_o level of a pair was also resynthesised based on its estimated *F*-pattern but applying the f_o level of the opposing lower sound of the pair. The duration of the resynthesised sounds was 1 sec., including a 0.05 sec. fade in/fade out. As a result, a sample of 18 resynthesised sounds was created.

The rationale of the resynthesis experiment was as follows: If a natural vowel sound produced at a middle or higher f_o level of a speaker's vocal range retains its vowel quality when resynthesised at the original f_o level, but shows a clear shift in vowel quality when resynthesised at the lower f_o level of an opposing sound of the same vowel, then this points towards a substantial vowel-related spectral variation with altering f_o. Conversely, if a vowel quality can be maintained in resynthesis independently of the f_o level applied, then this indicates no substantial vowel-related spectral variation.

Listening test of the resynthesised sounds: Vowel recognition of the resynthesised sounds was tested in a listening test according to the standard procedure of the Zurich Corpus and involving the five standard listeners (forced choice, excluding vowel boundaries but including schwa), with an additional test specification: Each prompt consisted of the original natural sound followed by one of the two resynthesised replicas (separated by a 0.5 sec. pause), and the listeners were asked to label the vowel quality of the second sound.

Results

Table 1 in the chapter appendix lists the selected sound pairs (including sound links) and the estimated *F*-patterns (including the LPC parameters applied), indicates the estimation of spectral differences < 1.5 kHz related to f_o variation and shows the vowel recognition results for the resynthesised sounds. (Note that estimated F_1 for sounds produced at higher f_o levels are given in parentheses to point towards the methodological estimation problem.)

For all three natural sound pairs of /u/, an increase in f_o levels of one octave or more resulted in a pronounced spectral variation < 1.5 kHz (marked differences in general spectral energy distribution and spectral peaks), and Klatt resynthesis based on the estimated *F*-patterns of the natural reference sounds produced at higher f_o levels but applying the lower f_o level of the opposing sound of the pair in question resulted in a marked vowel quality shift in a close–open direction, the shift including non-adjacent vowel qualities for the sounds of the adult speakers. In contrast, no comparable indication of a pronounced spectral variation or of a distinct vowel quality shift in resynthesis was found for the sounds of /a/.

Concerning estimated *F*-patterns and comparing the two sounds of a pair of /u/, F_1–F_2 varied strongly for all three pairs (differences in F_1 = 174–489 Hz; differences in F_2 = 371–970 Hz). These differences corresponded to pronounced differences in the general spectral energy distribution and the occurring spectral energy maxima. Comparing the two sounds of a pair of /a/, estimated F_1–F_2 did not vary markedly for two of the three pairs (differences in F_1 = 12–26 Hz; differences in F_2 = 81–96 Hz). For the third pair, marked differences were found, but no marked difference in the general spectral energy distribution was indicated in the direct spectral comparison.

However, frequency ranges and f_o variation of the sound pairs of a single speaker somewhat differed, with the range of f_o variation being higher for the sound pairs of /u/ than of /a/ for adults (differences of vowel-related upper f_o levels of the sound pairs = five semitones at a maximum). This difference was the result of two aspects: The applied selection criterion concerning the sounds of /a/ (aiming for high spectral similarity of the sounds of a pair) and the aim to demonstrate very pronounced vowel quality shifts for the sounds of /u/ produced at higher f_o levels when resynthesised applying lower f_o levels.

Discussion

In the present study, a direct visual comparison of the investigated sound spectra for the sound pairs of /u/ indicated very pronounced lower spectral differences with increasing f_o, contrary to the investigated sounds of /a/. Similarly, for the sounds of /u/, Klatt resynthesis based on estimated LPC filter curves indicated a pronounced impact on the recognised vowel quality if f_o was decreased from the higher to the lower level of a sound pair, contrary to the sounds of /a/.

The facts that *F*-pattern estimation was not methodologically substantiated for the natural reference sounds produced at middle and higher f_o levels, and that the frequency ranges and f_o variation of the sound pairs of a single speaker somewhat differed, do not relativise the demonstration of the nonuniform relation of the lower vowel spectrum to f_o concerning vowel quality and its impact on formant pattern and spectral shape ambiguity, for the following reasons: Firstly, according to the results of the vowel recognition test (labelling majority), resynthesis of the natural reference sounds produced at middle and higher f_o levels did not result in distinct vowel quality shifts when applying these middle or higher f_o levels of the reference sounds. Secondly, in all of the previous experiments concerning the relation of the lower vowel spectrum to f_o, marked spectral variation was found for the comparison of natural sounds of /u/ produced at f_o levels in the range of 220–440 Hz (these levels approximately corresponding to the levels of the sound pairs of /a/ presented here), in contrast to the comparison of natural vowel sounds of /a/.

In these terms, the extracts of the vocalises of the three speakers discussed here exemplify the nonuniform relation of the lower vowel spectrum to f_o concerning vowel quality and its impact on formant pattern and spectral shape ambiguity.

Chapter appendix

Table 1. Compilation of sound pairs of /u/ and /a/ produced by single speakers: Illustration of the nonuniform relation of the lower vowel spectrum to *f*ₒ with regard to vowel quality. Columns 1–5 = sounds (SP = speaker ID in the Zurich Corpus; SG = speaker group, where m = men, w = women, c = children; S/L = sound pairs and sound links; V = intended and recognised vowel quality of the natural reference sounds; fo = calculated *f*ₒ, in Hz). Columns 6–10 = estimated *F*-patterns (F(i) = formant frequencies, ΔF(i) = formant frequency differences; Par = LPC parameter setting for maximum formant number; estimated formant frequencies for sounds produced at higher fo levels are given in parentheses to point towards the methodological estimation problem). Column 11 = estimation of spectral differences for the sounds of a pair. Columns 12–15 = vowel recognition results for the resynthesised sounds (fo org = *f*ₒ of the natural reference sound applied in resynthesis; fo lowered = *f*ₒ of the opposing natural sound of a pair applied in resynthesis; maintained = vowel quality of the natural reference sound maintained in resynthesis; close–open = shift direction comparing the vowel qualities of the natural reference sound and the resynthesised sound). Colour code: Red = marked spectral differences for two sounds of a vowel associated with a recognised vowel quality shift (labelling majority) in resynthesis; blue = marginal spectral differences for two sounds of a vowel associated with maintained recognised vowel quality in resynthesis. For a cross-examination of the resynthesis, use the Klatt synthesiser in the corpus, taking into account the parameter settings for LPC analysis as indicated in Column 10.
[M-07-06-T01]

Table 1. Compilation of sound pairs of /u/ and /a/ produced by single speakers: Illustration of the nonuniform relation of the lower vowel spectrum to fo with regard to vowel quality. [M-07-06-T01]

Sounds					F-patterns (Hz)					Spectral difference	Vowel recongition for resynthesis			
											Summary		Details	
SP	SG	S/L	V	fo (Hz)	F1	ΔF1	F2	ΔF2	Par		fo org	fo lowered	fo org	fo lowered
1069	m	1 a ↗	u	125	294	375	653	667	P6	marked	maintained	–	o u u u u	–
				674	(669)		(1320)		P4		maintained	close–open	o u u u u	a a a ɛ ɛ
		b ↗	a	159	782	309	1266	270	P5	marginal	maintained	–	a a a a a	–
				459	(1091)		(1536)		P5		maintained	maintained	a a a a a	a a a a a
1068	w	2 a ↗	u	132	278	489	662	970	P5	marked	maintained	–	u u u u u	–
				772	(767)		(1632)		P5		maintained	close–open	e o u u u	ɛ ɛ ə a a
		b ↗	a	174	1041	26	1416	81	P5	marginal	maintained	–	a a a a a	–
				488	1015		1497		P5		maintained	maintained	a a a a a	a a a a a
1056	c	3 a ↗	u	262	376	174	745	371	P4	marked	maintained	–	o o u u u	–
				534	(550)		(1116)		P4		maintained	close–open	u u u u u	o o o o o
		b ↗	a	215	889	12	1309	96	P5	marginal	maintained	–	a a a a ɔ	–
				441	(877)		(1405)		P4		maintained	maintained	a a a a a	a a a a a

M7.7 The Role of the Fine Structure of Spectral Energy Distribution With Respect to the Relation of the Lower Vowel Spectrum to f_o

Introduction

Besides frequency ranges of f_o and vowel qualities, the fine structure of spectral energy distribution (the individual course of the estimated spectral envelope or the harmonic level configuration) also has an impact on the relation of the lower vowel spectrum to f_o, as was indicated in the investigation of the formant pattern and spectral shape ambiguity phenomenon (see Chapter M3; also consider the vocal effort-related spectral differences discussed in Chapter M5.4). In order to exemplify this impact, too, a corresponding documentary study was conducted: Sounds of the two vowels /ɛ/ and /a/ were selected from the Zurich Corpus for which vowel quality in resynthesis differed when applying increased f_o levels because of the different spectral energy distribution of the natural reference sounds.

Experiment

Vowel sounds and speakers: Voiced sounds of the Standard German vowels /ɛ/ and /a/ of the Zurich Corpus produced by women and men in nonstyle mode with low or medium vocal effort in V context at calculated $f_o \leq 250$ Hz were taken as the basis of investigation, for which the listening test conducted when creating the corpus provided a 100% recognition rate (matching vowel intention).

Inspection of sound spectra and sound selection: Natural sounds of the two vowels and their resynthesised replicas applying two levels of f_o in synthesis – calculated f_o of the natural reference sound and an approximately one-octave higher f_o level – were inspected, and two sound samples per vowel were compiled: A sample with vowel quality maintained for the resynthesised sounds applying both lower and higher f_o levels, and a sample with vowel quality shifts in an open–close direction for the resynthesised sounds applying the higher f_o level (Klatt synthesis; author's estimate). On this basis, subsequently, for each of the two vowels and a further limited f_o range < 160 Hz for the natural sounds, an exemplary documentation of the phenomenon in question was created in the form of a comparison of two natural sounds and their resynthesised replicas with maintained vowel quality for the replicas independent of f_o variation, and two natural reference sounds and their resynthesised replicas with vowel quality shifts for the replicas dependent on f_o variation. For this reduced sound sample (four sounds per vowel and eight sounds in total), vowel recognition of the resynthesised sounds was further investigated in a listening test involving the five standard listeners (see below).

Natural sounds produced in a lower frequency range of f_o were investigated for two reasons: For these sounds, the estimation of *F*-patterns (to which resynthesis was related) and spectral envelopes was much less problematic than for sounds produced at middle and higher f_o levels, and in the previous experiments reported in this treatise, no systematic vowel quality shifts were observed due to a one-octave increase in f_o for sounds of these vowels produced in a lower frequency range of f_o.

Estimation of *F*-patterns: *F*-patterns were estimated according to the standard acoustic analysis of the Zurich Corpus and a visual cross-check of the calculated formant frequency values based on the spectrum, spectrogram and formant tracks. If discontinuous formant tracks occurred, LPC parameters were altered until continuous tracks were obtained.

Additional note: The inspection of sound spectra and *F*-patterns, the first examination of resynthesised sounds and the subsequent sound selection was once again focused on finding examples best suited to documenting and illustrating the phenomenon of the fine structure of spectral energy distribution having an impact on the relation of the lower vowel spectrum to f_o in the context of the present treatise.

Klatt resynthesis experiment for the eight selected sounds: As indicated, using the KlattSyn tool (default parameters, 1 sec. sound duration, including a 0.05 sec. fade in/fade out), every single natural sound was resynthesised based on its estimated *F*-pattern but applying two f_o levels, its calculated f_o and an approximately one-octave higher f_o level of 300 Hz. As a result, a sample of 16 resynthesised sounds was created for a listening test.

Listening test: The vowel recognition of the resynthesised sounds was tested in a listening test according to the standard procedure of the Zurich Corpus and involving the five standard listeners (forced choice, excluding vowel boundaries but including schwa), with an additional test specification: Each prompt consisted of the original natural sound followed by one of the resynthesised sounds (separated by a 0.5 sec. pause), and the listeners were asked to label the vowel quality of the second sound.

Results

Table 1 in the chapter appendix lists the selected exemplary sounds and estimated *F*-patterns and shows the vowel recognition results for the resynthesised replicas. In the entire sound sample of the Zurich Corpus, numerous sounds of /ɛ/ produced at f_o levels ≤ 250 Hz were found for both resynthesis conditions, that is, maintained or shifted vowel quality as a result of an approximate one-octave increase in f_o. However, only a limited number of sounds of /a/ were found for the second resynthesis condition (occurring vowel quality shifts in an open–close direction in resynthesis). For the exemplary sounds of both vowels /ɛ/ and /a/ presented here, according to the vowel recognition results, no marked shifts were found for two of the sounds when resynthesised with f_o variation (see Series 1a and 2a) but shifts in an open–close direction were found for the other two sounds (see Series 1b and 2b).

Discussion

For all presented natural reference sounds, their f_o levels were within a narrow frequency range of 128–157 Hz, and f_o variation in resynthesis was comparable (approximately one octave in a similar frequency range). Thus, the observed difference in perceived vowel quality for the resynthesised sounds with increasing f_o, that is, a vowel quality shift for two sounds but no marked shift for the other two sounds of the same vowel, cannot be attributed to different frequency levels and ranges of f_o of the natural sounds and of f_o variation for their resynthesised replicas. Further, seven of the eight presented natural sounds were produced with medium vocal effort. Consequently, vocal effort variation cannot be considered the main explanation of the vowel recognition differences found for the resynthesised sounds. Therefore, the fine structure of spectral energy distribution (here in terms of the individual course of the spectral envelope and estimated LPC curve) in its turn has an impact on the relation of the (lower) vowel spectrum to f_o and, consequently, on occurring formant pattern and spectral shape ambiguity of vowel sounds. This phenomenon is exemplified here.

Note that, comparing sound pairs of Series 1a and 1b, and 2a and 2b, the differences in the fine structure of spectral energy distribution of the sounds related to marked differences of estimated F_1. Note also that, in this exemplary documentation, occurring examples of sounds of open-mid and open vowels are presented for which a one-octave f_o variation in resynthesis of approximately 150–300 Hz triggered a vowel quality shift in an open–close direction. Here, this frequency range of

f_o variation is assumed to be the lowest range for occurring formant pattern and spectral shape ambiguity (for corresponding examples, see also Chapter M3.3).

Chapter appendix

Table 1. Compilation of sounds of /ɛ/ and /a/ produced by speakers at similar f_o levels: Illustration of different recognised vowel qualities for resynthesised sounds of a vowel, increasing f_o by approximately one octave in resynthesis. Columns 1–6 = sounds (V = intended and recognised vowel quality of the natural reference sounds; S/L = sound pairs and sound links; SP = speaker ID in the Zurich Corpus; SG = speaker group, where w = women and m = men; VE = vocal effort; fo = calculated f_o of the natural reference sounds, in Hz). Columns 7–10 = estimated *F*-patterns (F(i) = formant frequencies; Par = LPC parameter setting for maximum formant number). Columns 11–14 = vowel recognition results for the resynthesised sounds (fo org = f_o of the natural reference sound applied in resynthesis; fo 300 Hz = increased f_o of 300 Hz applied in resynthesis; maintained = vowel quality of the natural reference sound maintained in resynthesis according to the labelling majority; open–close = vowel quality shift direction in resynthesis according to the labelling majority. Colour code: Blue = maintained vowel quality in resynthesis; red = vowel quality shift in resynthesis. For a cross-examination of the resynthesis, use the KlattSyn tool in the corpus, taking into account the parameter settings for LPC analysis indicated in Column 10.
[M-07-07-T01]

Table 1. Compilation of sounds of /ɛ/ and /a/ produced by speakers at similar fo levels: Illustration of different recognised vowel qualities for resynthesised sounds of a vowel, increasing fo by approximately one octave in resynthesis. [M-07-07-T01]

Sounds							F-patterns (Hz)				Vowel recognition for resynthesis			
											Summary		Details	
V	S/L		SP	SG	VE	fo (Hz)	F1	F2	F3	Par	fo org	fo 300 Hz	fo org	fo 300 Hz
ɛ	1	a ↗	1068	w	med	147	804	2670	3518	P5	maintained	maintained	ɛ ɛ ɛ ɛ ɛ	ɛ ɛ ɛ ɛ ɛ
			1052	w	med	148	828	2180	3143	P5	maintained	maintained	ɛ ɛ ɛ ɛ ɛ	ɛ ɛ ɛ ɛ ɛ
		b ↗	1076	m	med	128	584	2212	2814	P6	maintained	open–close	ɛ ɛ ɛ ɛ ɛ	ɛ e e e e
			1001	w	low	152	599	2029	2873	P5	maintained	open–close	ɛ ɛ ɛ ɛ ɛ	ɛ ɛ e e e
a	2	a ↗	1032	w	med	134	817	1313	–	P5	maintained	maintained	a a a a a	a a a a a
			1004	w	med	153	826	1336	–	P5	maintained	maintained	a a a a a	a a a a a
		b ↗	1059	w	med	131	568	1180	–	P5	maintained	open–close	a a a ɔ ɔ	o o o ɔ ɔ
			1006	w	med	157	627	1242	–	P5	maintained	open–close	a a a a a	o o o ɔ ɔ

M7.8 The Role of Vocal Effort Variation With Respect to the Relation of the Lower Vowel Spectrum to f_o

Introduction

In the context of the impact of the spectral fine structure on the relation of the lower vowel spectrum to f_o, finally, vocal effort variation for natural sounds of a vowel must also be taken into consideration (for corresponding previous indications, see Chapters M5.4 and M7.3). Therefore, and to exemplify this further impact in the present context of the treatise and its line of argument, a corresponding documentary study was conducted: Natural sounds of the two vowels /e/ and /o/ were selected from the Zurich Corpus for which both the lower spectral energy and the recognised vowel quality of their resynthesised replicas strongly related to a low or high vocal effort.

Experiment

Vowel sounds and speakers, and sound selection: Based on the sample presented in Chapter M5.4 (see Tables 1 and 2 in that chapter) and including four additional sounds from the Zurich Corpus, for each of the two vowels /e/ and /o/ and each of the two speaker groups, four sounds were compiled and arranged into two sound pairs. The first pair consisted of lower f_o and high vocal effort for the first sound and higher f_o and low vocal effort for the second sound; inversely, the second pair consisted of lower f_o and low vocal effort for the first sound and higher f_o and high vocal effort for the second sound. Lower calculated f_o levels were in the frequency range of 141–170 Hz for the sounds of men and 206–263 Hz for the sounds of women; higher f_o levels were in the frequency range of 319–351 Hz for the sounds of men and 429–441 Hz for the sounds of women; thus, the f_o difference between the two sounds of a sound pair was approximately one octave. The recognition rate for all selected natural sounds was 100% according to the standard listening test conducted when creating the Zurich Corpus, matching vowel intention.

Estimation of *F*-patterns, spectral comparison: For all sounds, *F*-patterns were estimated according to the standard acoustic analysis of the Zurich Corpus, with speaker group-specific default LPC parameters (P6 for men, P5 for women, P4 for children) and including a visual crosscheck of the calculated formant frequency values based on the spectrum, the spectrogram and the formant tracks. If discontinuous formant tracks occurred, LPC parameters were altered until continuous tracks were obtained (see Table 1 in the chapter appendix, indication for LPC parameters).

The entire estimated *F*-patterns were used for resynthesis (see below). However, the spectral comparison was limited to F_1, given an identifiable lowest spectral peak: The first formant is commonly considered as a main indicator of vowel-related spectral characteristics < 1.5 kHz, but, as shown, the lower vowel spectrum is strongly related to f_o.

Formant frequency estimation for the sounds produced at higher f_o levels has to be considered in the context of the methodological estimation problem for these f_o levels. Therefore, spectral differences were further related to directly comparing the sound spectra in question (for a crosscheck, view the sound pairs using the links in Table 1).

Crosschecking the effect of an increase in f_o on vowel quality in Klatt resynthesis: For all sounds produced at lower f_o levels < 300 Hz (for which LPC analysis is less problematic than for higher f_o levels), based on the estimated *F*-patterns and using the KlattSyn tool (default parameters, 1 sec. sound duration including a 0.05 sec. fade in/fade out), resynthesis was conducted with a step-by-step f_o increase from the f_o level of the natural reference sound to the higher level of f_o of the opposing sound of a sound pair. According to the author's estimate, an open–close vowel quality shift was triggered by a considerably smaller increase in f_o for sounds produced with a low vocal effort than for sounds produced with a high vocal effort. To demonstrate this phenomenon, an upper f_o limit of 250 Hz (sounds of men) or 330 Hz (sounds of women) for f_o variation in resynthesis was set as a default for the present study.

Klatt resynthesis experiment: In this context, using the KlattSyn tool (default parameters, 1 sec. sound duration including a 0.05 sec. fade in/fade out), every single natural sound produced at a lower f_o < 200 Hz (sounds of men) or < 300 Hz (sounds of women) was resynthesised based on its estimated *F*-pattern but applying two f_o levels, calculated f_o of the natural reference sound and increased f_o to 250 Hz (sounds of the men) or 330 Hz (sounds of the women), respectively. As a result, a sample of 16 resynthesised sounds was created for a listening test.

Listening test: The vowel recognition of the resynthesised sounds was tested in a listening test according to the standard procedure of the Zurich Corpus and involving the five standard listeners (forced choice, excluding vowel boundaries but including schwa), with an additional test specification: Each prompt consisted of the original natural sound followed by one of the resynthesised sounds (separated by a 0.5 sec. pause), and the listeners were asked to label the vowel quality of the second sound.

Results

Table 1 in the chapter appendix shows the selected sound pairs and estimated F_1 as well as the vowel recognition results for the resynthesised replicas. (Note that estimated F_1 for sounds produced at higher f_o levels are given in parentheses to point towards the methodological estimation problem.)

The pairwise spectral comparison of the natural sounds produced with either a high vocal effort at lower f_o or a low vocal effort at higher f_o showed marginal spectral envelope differences in the $F1$ frequency region, with estimated F_1 differences of < 50 Hz. Conversely, very pronounced spectral envelope differences were found in this frequency region for the pairwise comparison of the sounds produced with a low vocal effort at lower f_o and a high vocal effort at higher f_o, with estimated F_1 differences of 332–483 Hz.

Concerning vowel recognition for the resynthesised replicas applying calculated f_o of the natural reference sounds, according to the labelling majority, the vowel quality was maintained for all sounds independently of vocal effort and f_o levels. Concerning vowel recognition for the resynthesised replicas applying increased f_o, according to the labelling majority, f_o variation had no marked effect on the replicas of the sounds produced with a high vocal effort at lower f_o. Conversely, the effect was pronounced for the replicas of the sounds produced with a low vocal effort at lower f_o.

Discussion

In the previous chapter, sounds of /ɛ/ and /a/ were presented, for which the effect of f_o variation in resynthesis depended on the fine structure of spectral energy distribution. In these examples, vocal effort variation was highly limited: All sounds except one were produced with medium vocal effort, and the exception concerned a sound produced with low effort. For the present sound compilation, however, the spectral fine structure of the natural sounds and its impact on recognised vowel quality in resynthesis was directly related to a very pronounced low–high vocal effort variation in sound production: Concerning the spectral characteristics < 1.5 kHz, comparing an approximate one-octave difference for the natural sound pairs produced with a high vocal effort at a lower f_o level and those produced with a low vocal effort at a higher f_o level, only marginal differences were found for the frequency range of $F1$ despite the f_o difference. Consequently, it could be expected that the resynthesis of the natural sounds produced with a high vocal effort, including a limited f_o variation, would not result in a marked vowel quality shift. This expectation was confirmed in the vowel recognition experiment. In contrast, comparing an approximate one-octave difference for the natural sound pairs produced with a high vocal effort at a higher f_o level and those produced with a low vocal effort at a lower f_o level, very pronounced spectral differences < 1.5 kHz were found for the frequency range of $F1$. In consequence, it could be expected that the resynthesis of the natural sounds produced with a lower vocal effort, including a limited f_o variation, would result in a marked vowel quality shift. This second expectation was also confirmed in the vowel recognition experiment.

As discussed in Chapter M5.4, for vowel sounds produced at comparable f_o levels, vocal effort-related differences in the estimated P_1/F_1 often substantially surpassed 100 Hz, a frequency difference that approximates or equals F_1 differences of two adjacent vowel qualities as given in formant statistics. Here, because of the interaction of vocal effort and f_o level variation, P_1/F_1 differences comparing natural sounds produced with a low vocal effort at a lower f_o level and natural sounds produced with a high vocal effort at a higher f_o level were found in the range of 332–483 Hz, that is, equalling F_1 differences of two non-adjacent vowel qualities as given in formant statistics. This observation highlights, in its turn, the empirical contradiction to the thesis of formants as the primary acoustic representation of vowel quality.

In these terms, the observation of pronounced differences in the spectral fine structure of the sounds of a vowel caused by a marked vocal effort variation thus adds to the aspects that need to be considered when assessing the relation of the lower vowel spectrum to f_o. However, further research is needed to clarify the interrelation of vowel qualities, f_o ranges and level variations and spectral fine structure variation of sounds.

Register changes were not investigated and are not discussed in detail here, as holds true for the entire treatise. However, the empirical results presented (including the results of vowel resynthesis) and conclusions drawn in this treatise are not relativised by the general characteristics reported for register changes in the literature (for an overview of these characteristics, see Lee et al., 2021). Moreover, register classification during sound production requires special scrutiny to be exerted by both the investigator and the speaker, an undertaking that is barely possible within the scope of creating such a large-scale sound sample as presented by the Zurich Corpus.

Chapter appendix

Table 1. Compilation of sound pairs of /e/ or /o/ produced by men or women, the two sounds compared produced with either a lower or a higher vocal effort, including f_o variation: Illustration of the impact of vocal effort variation on the relation of the lower vowel spectrum to f_o. Columns 1–6 = sounds (V = intended and recognised vowel quality of the natural reference sounds; S/L = sound series and sound links; SP = speaker ID in the Zurich Corpus; SG = speaker group, where m = men and w = women; VE = vocal effort; fo = calculated f_o of the natural reference sounds, in Hz). Columns 7–9 = estimated first formant frequencies and related differences (F1 = estimated F_1, in Hz; ΔF1 = F_1 differences, in Hz; Par = LPC parameter setting for maximum formant number; estimated F_1 for sounds produced at higher f_o levels are given in parentheses to point towards the methodological estimation problem). Columns 10–13 = vowel recognition results for the resynthesised sounds (fo org = f_o of the natural reference sound applied in resynthesis; fo incr = increased f_o applied in resynthesis, with f_o = 250 Hz for the sounds of the men and f_o = 330 Hz for the sounds of the women; maintained = vowel quality of the natural reference sound maintained in resynthesis according to the labelling majority; open–close = shift direction comparing the vowel qualities of the original reference sound and its resynthesis at increased f_o, according to the labelling majority). Colour code: Blue = marginal differences in estimated F_1 for the comparison of the two sounds of a pair, associated with maintained vowel quality for the resynthesised replica of the sound with a higher vocal effort applying increased f_o in resynthesis; red = marked differences in estimated F_1, associated with an open–close vowel quality shift for the resynthesised replica of the sound with a lower vocal effort applying increased f_o in resynthesis. For a cross-examination of the resynthesis, use the KlattSyn tool in the corpus, taking into account the parameter settings for the LPC analysis indicated in Column 9.
[M-07-08-T01]

Table 1. Compilation of sound pairs of /e/ or /o/ produced by men or women, the two sounds compared produced with either a lower or a higher vocal effort, including fo variation: Illustration of the impact of vocal effort variation on the relation of the lower vowel spectrum to fo. [M-07-08-T01]

Sounds							Formant (Hz)			Vowel recongition for resynthesis			
										Summary		Details	
V	S/L		SP	SG	VE	fo (Hz)	F1	ΔF1	Par	fo org	fo incr	fo org	fo incr
e	1	a ↗	1049	m	high	170	437	21	P6	maintained	maintained	ɛ e e e e	e e e e e
			1063	m	low	351	(416)		P5	–	–	–	–
		b ↗	1030	m	low	141	316	332	P5	maintained	open–close	e e e e e	e i i i i
			1103	m	high	336	(648)		P6	–	–	–	–
	2	a ↗	1001	w	high	260	509	39	P5	maintained	maintained	e e e e e	e e e e e
			1031	w	low	431	(470)		P5	–	–	–	–
		b ↗	1102	w	low	206	343	483	P5	maintained	open–close	ɛ e e e e	i i i i i
			1006	w	high	429	(826)		P5	–	–	–	–
o	3	a ↗	1030	m	high	155	428	34	P6	maintained	maintained	o o o ɔ ɔ	o o o o ɔ
			1051	m	low	319	(394)		P6	–	–	–	–
		b ↗	1002	m	low	161	277	401	P6	maintained	open–close	o o o o u	u u u u u
			1077	m	high	341	(678)		P6	–	–	–	–
	4	a ↗	1088	w	high	263	527	49	P5	maintained	maintained	o o o o o	ɔ ɔ o o o
			1046	w	low	441	(478)		P5	–	–	–	–
		b ↗	1048	w	low	253	382	466	P5	maintained	open–close	o o o u u	u u u u u
			1006	w	high	432	(848)		P5	–	–	–	–

M8 Vowel Recognition of Filtered Vowel Sounds

M8.1 Low-Pass Filtering of Vowel Sounds and Related Vowel Recognition

Introduction

Studies on the effect of low-pass (LP) sound filtering and filter-like sound manipulation (such as formant level and spectral tilt variation) on vowel recognition have shown that, often, a front–back vowel quality shift (for sounds of front vowels) or an open–close shift (for sounds of back vowels and /a/) is associated with LP filtering and stepwise decreasing cutoff frequencies (CFs) or lowering the level of the second formant (L_{F2}) or steepening spectral tilt or varying the high- to low-frequency amplitude ratio. Inversely, raising the L_{F2} or flattening spectral tilt has, in some cases, been shown to result in a back–front vowel quality shift. (For the early related study of Stumpf, 1926, see the first appendix to this chapter; for an early study on LP filtering a comprehensive set of sustained vowel sounds with stepwise increasing CFs, see Lehiste and Peterson, 1959; for the effect of LP filtering sounds of front vowels, see Delattre et al., 1952; Dubno and Dorman, 1987; Shriberg, 1992; for LP filtering of sounds of back vowels and /a/, discussed in detail in Chapter M7.3, see Carpenter and Morton, 1962, and Morton and Carpenter, 1962; for an overview on formant level and spectral tilt variation, see Kiefte and Kluender, 2005; 2008, and Kiefte et al., 2010; see also Ito et al., 2001, for *F*2 suppression and for the variation of high- to low-frequency amplitude ratio.)

The fact that vowel recognition is not generally impaired as such but that, rather, the sound filtering and sound manipulation mentioned resulted in specific vowel quality shifts is viewed here as a possible further key phenomenon of vowel sounds in terms of the vowel being a kind of foreground–background phenomenon (see the excursus on vowel quality and harmonic spectrum in Part II). From this perspective, following the approach of Lehiste and Peterson (1959) and including variation of the age and gender of the speakers and of phonation type and the f_o level of voiced sounds, an LP filtering experiment was conducted.

Experiment

Vowel sounds and speakers, and sound selection: Based on the inspection of the Zurich Corpus, for each of the eight long Standard German vowels /i, y, e, ø, ɛ, a, o, u/ and for utterances of three speakers, a man, a woman and a child, a sample of sounds (V context) was compiled according to the following selection criteria per vowel and speaker: One whispered sound, one breathy sound and voiced sounds produced in nonstyle mode with a medium vocal effort at intended f_o levels of 131–262–523 Hz (man, three sounds) and 262–523 Hz (woman and child, two sounds each). (For the rationale of the f_o levels, see below, the LP sound filtering paragraph.) Note that, according to the standard procedure of the Zurich Corpus, vowel sounds with breathy phonation were produced by the speakers spontaneously, with speaker-related levels of intended f_o. Since creaky sounds were indicated to correspond to voiced sounds produced at lower f_o levels of the speakers' vocal range, creaky phonation was not investigated to limit the number of sounds. (Note that the spectral energy distribution for voiced and creaky sounds was generally found to be very similar and that synthesis based on creaky-related *F*-patterns with a voiced source at f_o of 131 Hz for the man and 220 Hz for the woman had no effect on vowel recognition; see Chapters M5.1 and M5.3.)

Based on the results of the standard listening test conducted when creating the corpus, for every single speaker and each of the above production parameter configurations, a sound with the highest occurring recognition rate was selected. Except for five sounds, the recognition rate for the selected sounds was 100%, matching vowel intention. The exceptions concerned two breathy sounds (80% and 60% recognition rate) and three voiced sounds (two sounds with 80% and one sound with 60% recognition rate; see the sounds marked in Table 1, Column 3). As a result, a sample of 104 natural reference sounds in total was created (40 sounds of the man, 32 sounds of each the woman and the child).

LP sound filtering: LP filtering was applied to all sounds with CFs of 2640–2370–2100–1840–1570–1310–1050–790–530 Hz using the Praat filter functionality (see the Praat manual, Sound: Filter [pass Hann band], with smoothing = 100 Hz). These CF frequencies corresponded to (i) approximate multiples in whole numbers of the high f_o level of the voiced sounds, the CFs given according to the musical C-major scale and rounded to the nearest tenth (from high to low = 2640–2100–1570–1050–530 Hz) and (ii) intermediate levels in between, rounded to the nearest tenth (from high to low = 2370–1840–1310–790). Accordingly, statistical F_2 of the sounds of long Standard German vowels were LP filtered at a CF of 1840 Hz for the investigated sounds of /i, e/ (taking the harmonic structure of all spectra into account), a CF of 1570 Hz for sounds of /ɛ/, a CF of 1310 Hz for sounds of /y, ø/, a CF of 1050 Hz for sounds of /a/ and a CF of 530 Hz for sounds of /o, u/ (see e.g. statistical F_2 given by Pätzold and Simpson, 1997; see also Maurer et al., 1992). As a result, a sample of 936 filtered sounds was created (360 sounds of the man, 288 sounds of each the woman and the child).

Listening test: Vowel recognition of the sounds was investigated in a listening test according to the standard procedure of the Zurich Corpus (including the vowel schwa) and involving the five standard listeners of the corpus. The entire sample of filtered sounds was divided into five subsets, separating the voiced sounds at f_o of 131 Hz, 262 Hz and 523 Hz, and further separating whispered and breathy sounds. Subsets were tested separately, with a pause of 15 minutes at a minimum in between the tests.

Results

Table 1 in the third chapter appendix lists the sample of the unfiltered natural reference sounds, including sound links, and shows the results of the vowel recognition test for the LP-filtered sounds. (For replication of LP filtering, refer to the SpecFilt tool in the corpus.) Table 2 summarises the analysis of occurring vowel quality shift directions and individual quality shifts. Below, the results are discussed for clear vowel quality shifts only, that is, for shifts from one quality to another, labelled by the majority of the listeners.

Filtered sounds of close and close-mid front unrounded vowels /i, e/: According to the general results given in Table 2, LP filtering with stepwise decreasing CFs resulted in a vowel quality shift in a front–back direction for all sounds of /i, e/. For most sounds, this general shift was preceded by an unrounded–rounded shift. For two sounds of /i/ and one sound of /e/, preceding close–open front shifts occurred, and for one sound of /i/ and six sounds of /e/, succeeding back open–close shifts occurred. Within the limit of these general shift directions, single vowel quality shifts varied among the sounds both between and within speakers with regard to phonation, f_o levels, CF levels and occurring series of recognised vowel qualities (see Tables 1 and 2).

Filtered sounds of close and close-mid front rounded vowels /y, ø/: As was the case for sounds of /i, e/, LP filtering caused a vowel quality shift in a front–back direction for all sounds of the vowels /y, ø/

(see Table 2). For four sounds of /y/, preceding close–open front shifts occurred, and for one sound of /y/ and six sounds of /ø/, succeeding open–close back shifts occurred. Within the limit of these general shift directions, comparable to the sounds of /i, e/, single vowel quality shifts varied, but the variation was very limited. Two single deviations from the general shift directions discussed occurred for two whispered sounds (see Table 1, sounds 1 and 10 of /ø/, highlighted in grey). Such interfering single deviations may indicate an additional acoustic effect of filtering vowel sounds.

Filtered sounds of the open-mid front vowel /ɛ/: LP filtering caused a vowel quality shift in a front–back direction and a subsequent shift in a back open–close direction for all sounds of the vowel /ɛ/. Within the limit of this general shift direction, comparable to the sounds of /i, e/, single vowel quality shifts varied (see Tables 1 and 2).

Filtered sounds of /a/: LP filtering caused a vowel quality shift in an open–close direction for all sounds of the vowel /a/. Within the limit of this general shift direction, single vowel quality shifts markedly varied among the sounds (see Tables 1 and 2). However, note that the indication /a/ in the text refers to the vowel area /a–ɑ/, that is, including all allophones of /a/ or /ɑ/ (for details, see the Introduction). This experimental condition may in part explain the variations mentioned.

Filtered sounds of the close-mid vowel /o/: LP filtering caused a vowel quality shift in an open–close direction for six of the 13 sounds of the vowel /o/. Besides, for the original breathy sound of the man (the recognition rate of the original sound was 60% only) and the whispered sound of the woman, the vowel quality of some filtered sounds deviated from the initial recognition of /o/ in that they displayed a close–open direction (see Table 1, recognition results marked in grey). This result may be an effect of the limited recognition of the original breathy sound or of artefacts of sound filtering.

Filtered sounds of the close vowel /u/: LP filtering caused no general vowel quality shift. However, for the original whispered sounds of all speakers, the original breathy sound of the woman and the voiced sounds at an f_o of 262 Hz of the child and an f_o of 523 Hz of the woman, the filtered sounds in part deviated from the initial recognition of /u/ (see Table 1, recognition results marked in grey). This result may support the occurrence of artefacts caused by LP filtering.

Discussion

Our diverse experiences regarding the recognition and acoustic representation of vowel quality, and our reading of the literature, have given rise to the general assumption that the vowel sound is a kind of foreground–background phenomenon. The idea as such is explained in detail in the excursus on vowel quality and harmonic spectrum in Part II. Here, only the effect of LP filtering on vowel recognition as one aspect of this phenomenon shall be discussed.

The main goal of the present experiment was to investigate in more detail the indications given in the literature as well as previous observations made within our own team that, if vowel sounds are LP filtered or manipulated LP filter-like, the recognised vowel quality may not generally be impaired or corrupted, but it may shift. In terms of an extended and more systematic approach and corresponding experiment, vowel recognition of LP-filtered voiced, breathy and whispered sounds with stepwise decreasing CFs was tested, with the sounds investigated being produced by three speakers different in age or gender and with different phonation types and at different f_o levels for the voiced sounds. On this basis, two main results were obtained.

Firstly, for all sounds of front vowels, LP filtering caused a general vowel quality shift in a front–back direction. For most sounds of the close front vowels, this shift was preceded by an unrounded–rounded shift, and for some sounds of the front vowels (above all for sounds of the close-mid front vowels), an additional open back to close back shift was observed. For all sounds of /a/, LP filtering caused shifts in an open–close direction. The same held true for half of the sounds of /o/.

Secondly, in the course and within the limits of these general shift directions, depending on vowel qualities, multiple single vowel quality shifts occurred for LP filtering of single natural reference sounds, including some variation regarding the order of single quality shifts for different reference sounds of a vowel. This variation was likely related to phonation types, f_o levels of the sounds and the course of the harmonic envelope. (Note in this context that the order of vowel qualities for shifts involving the recognition of the close-mid front rounded vowel /ø/ and the open-mid front unrounded vowel /ɛ/ was generally /ɛ/ before /ø/, that is, it was inverse to the close–open direction commonly given in the literature; see Table 1, extended online version, details of the vowel recognition results, sounds 2, 10 and 12 of /e/. We have often experienced that the vowel order close-mid front rounded and subsequent open-mid front unrounded does not correspond to observations of spectral characteristics and to results of vowel recognition tests. This aspect has to be considered for future research.)

The general vowel quality shift in a front–back direction was already indicated by the results of previous studies (see the introduction to this chapter and the below appendices I and II). The preceding unrounded–rounded shift for several filtered sounds of close front vowels and the occurrence of multiple single vowel quality shifts for LPC filtering of single sounds with stepwise decreasing CFs, as well as the variation of their succession for different reference sounds of a vowel, are important further indications (see also below).

The general phenomenon that the vowel quality of the sounds investigated was not lost but changed when higher spectral frequency ranges were deleted supported the notion that vowel sounds are describable in terms of a relation of lower spectral similarity (background) and subsequent spectral difference (foreground; see the Preliminaries, p. 81): If the effect of LP filtering is not looked at in terms of stepwise decreasing CFs but in terms of stepwise increasing CFs, sounds produced as different vowels can be perceived as similar in vowel quality up to a given CF level (related to the qualities in question as well as to the spectral energy distribution of the individual sounds compared for the frequency range up to this level), and the sounds only differ if higher spectral frequencies of the natural reference sounds above that CF level are included. As said, and as is laid out in more detail in the excursus mentioned above, this notion is considered here as being at the core of the vowel phenomenon. In the context of the present experiment, special attention should be given to the evidence (supporting earlier indications given in the literature) that vowel recognition is not based on model patterns of spectral peaks and does not directly (in an unmediated manner) relate to sound production.

With respect to the present experiment, most importantly, if front unrounded vowels were LP filtered with stepwise decreasing CFs below the frequency range of the second spectral peak of the original sounds (or the frequency range of the related statistical F_2), in many cases, the filtered sounds were at first still recognised as front vowels – but rounded – and then, subsequently, as back vowels (for exemplary illustration, see Part II, Chapter 8.1, Figures 1 and 2). Consequently, a pattern of two spectral peaks is not a precondition of vowel recognition, and the vowel quality of sounds with only one manifest peak in the spectrum can be clearly recognised. This finding may be less surprising for filtered sounds recognised as back vowels since early vowel synthesis studies have already indicated recognisable sounds, above all, of back vowels and /a/ based on only one synthesis filter below 1.5 kHz (see e.g. Delattre et al., 1952). But, as the above exemplary illustration shows, filtered sounds recognised as front rounded vowels may not present a spectral peak above 1 kHz, and no existing concept of general patterns of spectral peaks related to vowel recognition can account for this type of spectral manifestation. Note in this context that an account of accurate vowel recognition for sounds of front vowels for which only *F*1 is well specified, but the sound spectra manifest a broad energy region with no marked peaks in the higher frequency region, was already given by Dubno and Dorman (1987). The same holds true for occurring front–back shifts found for LP filtering of front vowels, also demonstrated by these authors. However, as was indicated by the results of the present experiment, front–back shifts are not always direct even if CF is set in between F_1 and F_2 of the natural sounds, since some sounds of front unrounded vowels shifted to front rounded vowels after LP filtering of the second spectral peak of the natural sounds. This observation is consistent with the results of Ito et al. (2001), who showed that sounds of front vowels can be recognised even when F2 is suppressed as much as possible without changing the original spectral shape. They concluded that "[…] vowel quality, especially its place of articulation (front/back), can be perceived even if *F*2 information is not available. There would thus be cues for place of articulation instead of the second formant frequency. The amplitude ratio of high- to low-frequency components of the spectrum might be a good candidate for one of these cues." Finally, all this corresponds to the observation of sounds of front vowels manifesting flat spectral energy distribution > 1 kHz (see Chapters M7.3 and M7.4). However, in our view, no direct (unmediated) relation can be established between vowel recognition and vowel sound production: Evidently, the actual resonance characteristics of vowel sound production – and with it, the actual articulator positions – cannot, in general, be recognised on the basis of a radiated vowel sound, as was demonstrated in a paradigmatic manner by the observable, recognised shifts from front unrounded to front rounded to back vowels in LP filtering.

The variation regarding individual vowel quality shifts occurring within the limits of the general shift directions indicated that f_o, as well as the individual course of the spectral envelope and/or the individual configuration and frequency distance of the harmonics of the unfiltered reference sounds, has to be taken into account when considering the effect of LP filtering or LP filter-like sound manipulations on vowel recognition. The same holds true for two additional aspects: Firstly, depending on vowel qualities and CFs, the consensus of the listeners on vowel quality labelling of LP-filtered sounds varied. Secondly, filtered sounds for which vowel recognition deviated either from general shift directions (rare in number) or from the vowel quality of original sounds of /o, u/ indicated that possible artefacts of LP filtering might occur.

In the general context of discussing the foreground–background character of the vowel sound and in the specific context of discussing LP sound filtering, the observation of Shriberg (1992, relating to the results of her investigation) is worth noting: When sounds of back and front vowels were LP filtered with a CF of 1 kHz, as was to be expected, almost all listeners recognised the sounds as back vowels. But when HP-filtered noise with a CF of 1.3 kHz was added to the LP-filtered voiced sounds, some front–back shifts of sounds produced as front vowels were reversed in that the front quality of the vowels was recognised and “restored”, and some sounds produced as back vowels were recognised as front vowels. This observation, in turn, supports the foreground–background character of the vowel sound.

Besides, note also that an estimation of *F*-patterns for the filtered sounds often lacked a methodological substantiation.

Finally, with regard to a future investigation of the foreground–background character of the vowel sound, this investigation should include a more extensive within-speaker variation of production style, f_o level and vocal effort of the produced sounds.

Chapter appendix I

In an early comprehensive study, Stumpf (1926) investigated the harmonic spectrum of voiced sounds for the Standard German vowels /i, y, e, ø, ɛ, a, o, u/ produced by female and male singers (adults and children) at different f_o levels from 65 to 523 Hz (the main levels investigated being 65–93–131–185–262–370–523 Hz according to the musical C-major scale as referred to in this treatise; see Stumpf, 1926, p. 54). In his study, he analysed formants in terms of dominant harmonics with relative energy maxima in the sound spectrum using an interference apparatus: “The device consisted in a huge constellation of tubes […]. It led sound through a labyrinth made up of tubes of various lengths. When a sound component with a particular wave length passed through a tube of the same length, it was cancelled out by its own mirror image in the tube. In this way, particular sound components could be subtracted from a compound sound.” (Kursell, 2019; for an extensive description of the method, see Stumpf, 1926, pp. 36–53.) Investigating the vowel spectrum of a natural sound, using this apparatus, Stumpf deleted all harmonics except for the first one and then, step-by-step, increased the number of harmonics of the natural reference sound (“Aufbaureihe”). In the course of successively increasing the number of harmonics, he assessed the upper harmonic number needed to recognise the vowel quality of the natural reference sound. On this basis, he derived the vowel quality-related formant frequency of the sound. As a control method, he investigated natural vowel sounds in an inverse way, that is, step-by-step deleting harmonics from the highest one to the first one only (“Abbaureihe”). Moreover, he also investigated the effect of deleting only harmonics in the middle of a sound spectrum.

Looking at his recognition results within the present context of LP filtering (step-by-step deleting harmonics from the highest one to the first one only; “Abbaureihe”), in sum and ignoring sounds produced at f_o of 65 and 523 Hz, he reported general vowel quality shift directions of /i/–/y/–/u/, /e/–/ø/–/o/–/u/ and /ɛ/–/a/–/ɔ/–/o/–/u/. However, some differences were found for different fundamental frequencies of the sounds. (For details, see Stumpf, 1926, pp. 57–59.) Remarkably, these observations are largely consistent with the findings reported in this chapter: Roughly speaking, according to Stumpf, LP filtering causes vowel quality shifts and shift directions from front unrounded to front rounded, from front to back and from back open to back closed. On his part, he considered particularly important the fact that the vowel character is not lost when deleting higher harmonics but that, instead, the vowel quality changes and that the quality shift directions are predictable. Furthermore, he also referred to the phenomenon of front–back confusions or double-vowel recognition in the transition from a front to a back vowel. (For these indications, see Stumpf, 1926, p. 61 and p. 342.) Thus, Stumpf had already been attentive to the phenomenon of the vowel sound being (according to our terminology) a kind of perceptual and acoustic foreground–background phenomenon.

Chapter appendix II

As said, we consider the phenomena encountered in LP filtering of primary importance for the understanding of the acoustics and the perception of the vowel. In this context, special reference to the study of Ito et al. (2001) in the form of longer quotes shall be made in this second appendix, for two reasons: They indicate that front vowels can be perceived even if the spectrum does not manifest a peak in the frequency range commonly assumed to be related to statistical F_2 and that the amplitude ratio of the high- to low-frequency range can be crucial for vowel recognition.

We do not further comment on or discuss these quotes. However, the entire line of argument and investigation in the present treatise, as well as the entire context of the results reported here, give reason for a different interpretation of the vowel recognition results for filtered sounds compared with the conclusions of Ito et al. (for our interpretation, see the excursus on vowel quality and harmonic spectrum as well as Chapter 10 in Part II).

Quotes: “The formant hypothesis of vowel perception, where the lowest two or three formant frequencies are essential cues for vowel quality perception, is widely accepted. There has, however, been some controversy suggesting that formant frequencies are not sufficient and that the whole spectral shape is necessary for perception. Three psychophysical experiments were performed to study this question. In the first experiment, the first or second formant peak of stimuli was suppressed as much as possible while still maintaining the original spectral shape. […] In the second experiment, *F*2-suppressed stimuli, whose amplitude ratios of high- to low-frequency components were systemically changed, were used. […] In the third experiment, the full-formant stimuli, whose amplitude ratios were changed from the original and whose F2’s were kept constant, were used.”

“In our first experiment we examined the response for the stimuli with the first or second formant suppressed. If formant frequencies are exclusive cues for vowel quality perception, this suppression should inevitably change the perception of vowel quality. However, in contrast with this prediction, the response for the suppressed stimuli was not greatly changed from the control. These results cannot easily be explained by the formant hypothesis based on a local peak-picking mechanism. It strongly implies to us that formant frequencies are not exclusive cues for vowel perception.”

“In addition to formant frequencies, we demonstrated that the amplitude ratio of high- to low-frequency components might also be a crucial cue for vowel perception. The result of our second experiment showed that changes in this ratio could induce a change in perceived vowel quality, especially its place of articulation (front/back). This cannot be explained by COG theory because the frequency separation in our experiment was always greater than 3.5 bark.”

“Beddor and Hawkins (1990) concluded that spectral shape was the primary cue for vowel perception only when the lower spectral prominence was weak, while it was secondary when the prominence was sufficient. In experiment 2, there was no prominent peak in the frequency region of the stimuli where *F*2 was expected to be in natural speech. This seems to agree well with the results in the limited conditions as pointed out by Beddor and Hawkins. The results of our third experiment, however, showed that the amplitude ratio was still effective for vowel perception even when the prominence of the *F*2 peak was sufficient. That is, the effect of the amplitude ratio on the perception of place of articulation was found to be equal to or greater than that of *F*2."

"It should be noted that our hypothesis has at least three problems. The first is the definition of the amplitude ratio. […] The second problem of the amplitude ratio hypothesis is that it cannot explain the relation between spectral shape and vowel height. For F1-suppressed stimuli in experiment 1, some recognition rates of vowels changed depending on the suppression of F1. Another cue other than the amplitude ratio seems to be required to explain this result. The third problem is a context effect. Since vowel perception is known to be context dependent, there is a possibility that subjects adapted their criteria to each type of stimuli in our experiments."

"The results of our experiments indicate that vowels can be perceived by the whole spectral shape of stimuli even if their formant frequency is not available. The simplest model for these results is storing the spectral shapes of all vowels as the templates, comparing input with them, and determining the output vowel based on the similarity between templates and input. This type of model has been adopted often in recent commercially available systems for speech perception. However, this is obviously not proper as a cognitive model for human beings because it cannot explain our excellent ability at speaker adaptation with limited storage capacity. This is the reason why an elegant compressed representation of whole spectrum shape is required. We propose the amplitude ratio as one of candidates for the compressed representation. This cannot be psychophysically verified using a cascade-type synthesizer, which was used in several studies (e.g., Fahey and Diehl, 1996), because the synthesizer cannot produce the formant frequencies and their amplitudes independently. This might be one of the reasons why the effect of amplitude ratio has not been clarified."

"Conclusions: (1) Formant frequencies are not exclusive cues for vowel perception. All three of our experiments support this conclusion. In experiment 1, a suppression of the first or second formant peak did not cause a great difference in distribution patterns of recognition rates when whole spectral shape was not largely changed […]. In experiments

2 and 3, a modification of whole spectral shape did induce a change of perceived vowel even when all formant frequencies were constant […].

“(2) The amplitude ratio of high- to low-frequency components is no less effective for place of articulation than the second formant frequency. […] As seen in experiment 2, this change of the amplitude ratio of high- to low-frequency components itself caused the change of perceived vowel in its place of articulation […]. This effect was also verified in experiment 3, in which the amplitude ratio of high- to low-frequency components competed against the second formant frequency […].”

On this matter, see also the quotation from Carpenter and Morton (1962) given in Chapter M7.3.

Chapter appendix III

Table 1. LP filtering of sounds of the long Standard German vowels: Vowel recognition results. Columns 1–5 = unfiltered natural reference sounds (SP = speakers and speaker group, where m = man, w = woman, c = child; S = order number of a sound in a vowel-related series; P = phonation type, where v = voiced, b = breathy, w = whispered; fo = intended f_o, in Hz; note that the f_o level of breathy sounds is given according to the scale of the intended levels of the voiced sounds; V/L = intended and recognised vowel quality of the unfiltered natural reference sounds, and sound links in the order of sound listing). Columns 6–14 = recognised vowel qualities of the LP-filtered sounds per CF (summary; vowel recognition rate ≥ 60%). Extended online table: Columns 15 ff. = details of the vowel recognition results (labelling of the five listeners). Colour code: Dark blue = recognised vowel quality matching the quality of the unfiltered reference sound; light blue = recognised front qualities differing from the quality of the unfiltered reference sound; light red = recognised open, open-mid and close-mid back qualities differing from the quality of the unfiltered reference sound; dark red = recognised close back quality differing from the quality of the unfiltered reference sound; grey = recognised vowel quality shifts in close–open or rounded–unrounded direction for single sounds of /ø/, /o/ and /u/; no colour = vowel confusion. Natural unfiltered sounds with a vowel recognition rate below 100% are marked as follows (see Column 3): “1” = 80% recognition rate, “2” = 60% recognition rate. Note that, in Columns 15 ff., ns = not specified (no vowel quality recognised), and txt = free comment.
[M-08-01-T01]

Table 2. LP filtering of sounds of the long Standard German vowels: General vowel quality shift directions and single quality shifts. Analysis of the results given in Table 1; results are given in relation to the intended and recognised vowel quality of the unfiltered natural reference sounds. Columns 1 and 2 = natural unfiltered reference sound (V = intended and recognised vowel quality; N = number of all sounds of a vowel investigated). Columns 3–7 = occurring vowel quality shifts and shift directions in LP filtering in reference to the quality of the unfiltered natural reference sounds, and related number of the reference sounds (shifts all = occurring general shifts in one or several shift directions as given in the table; c–o front = close to open front shift direction; ur–r front = unrounded to rounded front shift direction; front–back = front to back shift direction; o–c back = open to close back shift direction). Columns 8 ff. = single vowel quality shifts and related number of natural reference sounds (N) for which the shifts occurred. Colour code, see Table 1, except grey = main shift directions for sounds of a vowel. Note that, for filtered sounds of /e/, the order of vowel qualities in shifts involving the recognition of the close-mid front rounded vowel /ø/ and the open-mid front unrounded vowel /ɛ/ was /ɛ/ before /ø/ (see the qualities marked with “*”; see also the text). For the results of /u/, see the text.
[M-08-01-T02]

Table 1. LP filtering of sounds of the long Standard German vowels: Vowel recognition results. [M-08-01-T01]
Extended online table: ↗

Sounds					Vowel recognition per CF (in Hz)								
SP	S	P	fo (Hz)	V/L	2640	2370	2100	1840	1570	1310	1050	790	530
man	1	w	–	/i/ ↗	i		y					u	u
	2	v	131		i	i	y	y	y	u	u	u	
	3	b	220		i	y	y	y	y	u	u	u	u
	4	v	262		i	y	y	y	y	u	u	u	
	5	v	523		y	y	y	y	y	u	u	u	u
woman	6	w	–			y	y	y				u	u
	7	b	262		e		u	u	u	u	u	u	u
	8	v	262		i	y	y	y	y	u	u	u	u
	9	v	523			y	y	y		u	u	u	u
child	10	w	–			y	y				o	u	
	11	b	262			y	y		u	u	u	u	u
	12	v	262		e					u	u	u	u
	13	v	523					u	u	u	u	u	u
man	1	w	–	/e/ ↗	e	e	ø	ø	ø	o	o	o	o
	2	v	131		e	e	e	ø	ø	ø	o	o	o
	3	b	220		e		ø	ø	ø		o		o
	4	v	262			e		ø	ø		o	o	u
	5	v	523		e	e	ø	ø	ø	u	u	u	u
woman	6	w	–		e		ø	ø		o	o	o	u
	7	b	262		e	e	ø	ø		o	o	o	u
	8	v	262		e		ø	ø	ø	o	o	o	o
	9	v	523		e	ø	ø	ø	ø	o	o	u	u
child	10	w	–			ɛ	ø	ø			ɔ	o	u
	11	b	262		e	ø	ø	ø	ø	o	o	o	
	12	v	262		e	e	ø	ø	ø	o	o	o	
	13	v	523		e		ø	ø		o	o	u	u

Sounds					Vowel recognition per CF (in Hz)								
SP	S	P	fo (Hz)	V/L	2640	2370	2100	1840	1570	1310	1050	790	530
man	1	w	–	/y/ ↗	y	y	y	y		u	u	u	u
	2	v	131		y	y	y	y	y	u	u	u	u
	3	b	220		y	y	y	y		u	u	u	u
	4	v	262		y	y	y	y	y		u	u	u
	5	v	523		y	y	y	y	ø	u	u	u	u
woman	6	w	–		y	y	y	ø	ø	u		u	u
	7	b	262		y	y	y	y	y	u	u	u	
	8	v^2	262		y	y	y	y	y	u	u	u	u
	9	v	523		y	y	y	y	y	u	u	u	u
child	10	w	–			e		ø		o	o		u
	11	b	262		y	y	y	y	u	u	u	u	u
	12	v	262		y	y	y	y	ø	u	u	u	u
	13	v	523		y	y	y	y		u	u	u	u
man	1	w	–	/ø/ ↗	ø	ø		y	ø		u	o	u
	2	v	131		ø	ø	ø	ø	ø	ø	o	o	o
	3	b	220		ø	ø	ø	ø			o	u	o
	4	v	262		ø	ø	ø	ø	ø	o	o	o	o
	5	v	523		ø	ø	ø	ø	ø	o	u	u	
woman	6	w	–		ø	ø	ø	ø		o		o	u
	7	b	262		ø	ø	ø	ø	ø	o	o	o	u
	8	v	262		ø	ø	ø	ø	ø		o	o	
	9	v	523		ø	ø	ø	ø			u	u	u
child	10	w	–		e		ø	ø	ø	o	o	o	u
	11	b	262		ø	ø	ø	ø	ø	o	o	o	
	12	v	262		ø	ø	ø	ø	ø	o	o	o	
	13	v	523		ø	ø	ø	ø	o	o	o	u	u

Table 1 (continuation). [M-08-01-T01]

Sounds					Vowel recognition per CF (in Hz)								
SP	S	P	fo (Hz)	V/L	2640	2370	2100	1840	1570	1310	1050	790	530
man	1	w	–		ɛ	ɛ	ɛ		a		ɔ	ɔ	u
	2	v	131		ɛ	ɛ	ɛ	ɛ	ɛ	a	a	ɔ	o
	3	b	220		ɛ	ɛ	ɛ		a	a	ɔ		u
	4	v	262		ɛ	ɛ	ɛ	ɛ	a	ɔ	ɔ		
	5	v	523		ɛ	ɛ			ɔ	ɔ	o		u
woman	6	w	–		ɛ	ɛ	ɛ	a	a	a	ɔ	u	u
	7	b	262	/ɛ/	ɛ	ɛ	ɛ		a	ɔ	ɔ	o	u
	8	v	262	↗	ɛ	ɛ	ɛ		ɔ	ɔ	ɔ	o	u
	9	v	523		ɛ	ɛ	ɛ	a	a		ɔ	u	u
child	10	w	–		ɛ	ɛ	ɛ	a	a	a	ɔ	u	u
	11	b	262		ɛ	ɛ	ɛ	a	a		ɔ		u
	12	v	262		ɛ	ɛ	ɛ	a	a	a	ɔ		u
	13	v	523		ɛ	ɛ		a	a	a	o	u	u
man	1	w	–		a	a	a	a	a	a	ɔ	o	u
	2	v	131		a	a	a	a	a	a	a	a	o
	3	b	220		a	a	a	a	a	a	ɔ	u	u
	4	v	262		a	a	a	a	a	a		ɔ	o
	5	v	523				a	a	a	ɔ	o	u	u
woman	6	w	–		a	a	a	a	a	a		u	u
	7	b	262	/a/	a	a	a	a	a	a	ɔ	u	u
	8	v	262	↗	a	a	a	a	a	a	a	ɔ	u
	9	v	523		a	a	a	a	a	ɔ		u	u
child	10	w	–		a	a	a	a	a	a		u	u
	11	b	262		a	a	a	a	a	a		o	u
	12	v	262		a	a	a	a	a	a	a	ɔ	u
	13	v	523		a	a	a	a	a	a	ɔ		

Sounds					Vowel recognition per CF (in Hz)								
SP	S	P	fo (Hz)	V/L	2640	2370	2100	1840	1570	1310	1050	790	530
man	1	w	–						o	o		o	
	2	v	131		o	o	o	o	o	o	o	o	o
	3	b^2	220					a	ɔ	ɔ		o	
	4	v	262		o	o	o	o	o	o	o	o	
	5	v^1	523		o	ɔ	o	o	o	o	o	u	u
woman	6	w	–		ɔ	ɔ				ɔ	ɔ	ɔ	u
	7	b	262	/o/		o	o	o	o	o	o	o	u
	8	v	262	↗	o		o	o	o	o	o	o	o
	9	v	523		o	o	o	o	o	o	o		u
child	10	w	–			o	o		o	o	o	o	u
	11	b	262		o	o	o	o	o	o	o	o	u
	12	v	262		o	o	o	o	o	o	o	o	
	13	v	523						o		o		u
man	1	w	–		ɛ	ɛ				ɔ		o	u
	2	v	131		u	u	u	u	u	u	u	u	u
	3	b	220		u	u		u		u	u	u	u
	4	v	262		u	u	u	u	u	u	u	u	u
	5	v	523		u	u	u	u	u	u	u	u	u
woman	6	w	–		u		u	u			o	o	u
	7	b^1	262	/u/		u	u				o	o	u
	8	v	262	↗	u	u	u	u	u	u	u	u	u
	9	v^1	523		o	o			u			u	u
child	10	w	–		o				o			o	u
	11	b	262		u	u	u	u	u	u	u	u	u
	12	v	262				o	o	o	o	o	o	
	13	v	523		u	u		u	u	u	u	u	u

Table 2. LP filtering of sounds of the long Standard German vowels: General vowel quality shift directions and single quality shifts. [M-08-01-T02]

Sounds		Schift directions					Shifts (single)									
V	N	Shifts	c–o	ur–r	front–	o–c	i	y	e	ø	ɛ	a	ɔ	o	u	N
		(all)	(front)	(front)	back	(back)										
i	13	13	1	10	13	1	i	y							u	9
							i	y						o	u	1
							i		e						u	2
							i								u	1
e	13	13	1	13	13	6			e	ø				o	u	5
									e	ø				o		6
									e	ø					u	1
									e	ø*	ɛ*		ɔ		u	1
i, e	26	26	2	23	26	7										26
y	13	13	4	0	13	1		y	e	ø				o	u	1
								y		ø					u	3
								y							u	9
ø	13	13	–	–	13	6				ø				o	u	6
										ø					u	1
										ø				o		6
y, ø	26	26	4	0	26	7										26
ɛ	13	13	–	–	13	13					ɛ	a	ɔ	o	u	1
											ɛ	a	ɔ		u	7
											ɛ	a		o	u	1
											ɛ	a	ɔ	o		1
											ɛ	a	ɔ			1
											ɛ		ɔ	o	u	2
ɛ	13	13			13	13										13
a	13	13	–	–	–	13						a	ɔ	o	u	2
												a	ɔ		u	5
												a		o	u	1
												a			u	2
												a	ɔ	o		1
												a	ɔ			1
												a		o		1
a	13					13										13
o	13	6	–	–	–	6								o	u	6
														o		5
												a	ɔ	o		1
													ɔ		u	1
o	13					6										13
u	13	0	–	–	–	–					(ɛ)		(ɔ)	(o)	u	12
														o		1
u	13															13

M8.2 High-Pass Filtering of Vowel Sounds and Related Vowel Recognition

Introduction

As discussed and demonstrated in Chapter M6.4 concerning the suppression of *H*1 and *H*1–*H*2, this kind of vowel sound filtering did not generally impair or corrupt vowel recognition: For speaker group-specific levels of f_o, as given in formant statistics, there was either no marked shift in recognised vowel quality or a shift in a close–open direction.

Concerning the general question of HP filtering of vowel sounds, in an early study, Lehiste and Peterson (1959) investigated the recognition of filtered sustained sounds of two sets of vowels, a set of 18 IPA vowels and a set of ten English vowels and two diphthongs (presumably American English vowels and diphthongs), the sounds produced by the second author (Peterson) at f_o of c. 140 Hz. CF levels of HP filtering were set to 0–550–1050–2100–3500–4800 Hz. As the results showed, vowel recognition depended strongly on individual vowel qualities. To give three examples concerning the results for the sounds of the second set of vowels: (i) When HP-filtered with CFs of 0–550–1050–2100 Hz, the recognition rates matching vowel intention (converted to percentage of accurate recognition) for the sound of /i/ were 100–80–60–100%, with vowel substitutions (inaccurate identifications) for the diphthong /ei/ or the monophthong /ɛ/. Thus, the sound of /i/ was accurately recognised when the estimated *F*1 was deleted, and spectral energy was manifest in the higher frequency range of the estimated *F*2–*F*3 range only. (Note also the temporary decrease of the recognition rate for CFs of 550–1050 Hz and the subsequent recovery of the rate for CF of 2100 Hz.) The study also reported results of LP filtering of the sound of /i/ (see Chapter M8.1) and, remarkably, the vowel quality of /i/ proved also to be maintained when only the estimated *F*1–*F*2 were represented in the sound spectrum but *F*3 was deleted. (ii) When HP-filtered with the above CFs of 0–550–1050–2100 Hz, the recognition rates matching vowel intention for the sound of /ɛ/ were 100–100–40–10%, with multiple vowel substitutions (inaccurate identifications). Thus, deleting *F*1 caused irretrievable vowel confusion. (iii) When HP-filtered with the same CFs of 0–550–1050–2100 Hz, the recognition rates for the sound of /ɑ/ were 90–90–70–0%, with vowel substitutions for the CFs of 550–1050 Hz mostly being /ɔ/ or /ou/ and for the CF of 2100 being one of the front vowels /ɛ/ or /ɪ/ or /i/. Thus, deleting *F*1 (CFs of 550–1050 Hz) caused a somewhat limited vowel confusion. Aside from these vowel quality-related recognition patterns, the HP filtering of sounds of the back vowels /u, ɔ, ɑ/ and of the diphthong /ou/ (indication according to Lehiste and Peterson) with a CF of 2100 Hz resulted in recognised front qualities (≥ 90% recognition rate for back–front shifts).

In the present context, three indications of this early study deserve special attention. Firstly, investigating filtered sounds of a small set of vowels does not allow for a generalisation of results for other vowels since the study showed – similarly to many other studies on other matters – that the recognition results were nonuniform among vowel qualities. This observation is in line with the numerous indications of the nonuniform character of the vowel spectrum. Secondly, the observation that a stepwise increase of the CF level in HP filtering can cause an initial decrease of accurate vowel recognition that is then followed by a recognition regain is also remarkable. This observation supports the foreground–background thesis. Thirdly, the finding of a back–front confusion (substitution) when filtering the frequency range < 2100 Hz of sounds of back vowels in its turn supports the foreground–background thesis. At the same time, the back–front confusions observed are remarkable considering the fact that the spectral representation of the assumed *R*1 and *R*2 of sound production was lacking: From the perspective of the prevailing acoustic theory, how can it be understood that a vowel quality other than the one intended by the speaker can be recognised when listening to a sound if the entire assumed vowel-related frequency range of its production is deleted?

However, when interpreting the results of this early study, some relativisations have to be made: On the one hand, the relation of CF levels and the frequency ranges of vowel-specific spectral energy for different vowel qualities was not considered, and above 1050 Hz, the CF frequency intervals were large. This may in part explain the nonuniform results among vowel qualities. On the other hand, no variation in the speakers, phonation type, f_o levels and vocal effort was investigated. This limits the generalisation of the results for other settings of production parameters.

In a more recent study investigating American English vowel sounds produced by a woman in /bVd/ syllable context, Liu and Eddins (2008) tested the effects of HP filter-like sound manipulation on vowel identification by progressively HP filtering vowel sounds in the spectral modulation domain, applying cutoff frequencies of 0.0, 0.5, 1.0, 1.5, and 2.0 cycles/octave. In general, according to the conclusion of the authors, vowel qualities were first confused with adjacent counterparts, and vowel confusion increased with increasing CF levels. However, marked differences in the recognition results related to vowel backness occurred, with a nonuniform decrease of accurate recognition among vowels: Filtered sounds of the back vowels /u, o/ were strongly affected, but accurate vowel recognition > 70% was maintained for the sounds of the front vowels /i, e, ɛ/ up to a filter condition of 2 cycles/octave. (Note that recognition details are given by the authors for filter conditions of 0–1–2 cycles/octave only.) Further, within the three groups of back, central and front vowels, the decrease in accurate vowel recognition rate depended on the individual vowel qualities. To give three examples for front vowels, the recognition rates for the filter conditions of 0–1–2 cycles/octave were 100–94–91% for /i/, 99–95–76% for /e/ and 89–93–71% for /ɛ/. For the same filter conditions, to give three examples for back vowels, the recognition rates were 93–21–12% for /u/, 95–45–25% for /o/ and 98–74–57% for /ɔ/. However, as with the Lehiste and Peterson (1959) study, these results were again obtained for sounds of a single speaker only, with no further within- or between-speaker variation of production parameters.

As was the case for the LP-filtered vowel sounds, the indication given by the results of these HP filter studies that vowel recognition as such was not generally impaired or corrupted but that, rather, the HP sound filtering resulted in specific vowel quality shifts is viewed here as a further key phenomenon regarding the vowel being a kind of foreground–background phenomenon. From this perspective, extending the experiment discussed in Chapter M6.4 and following the approach of Lehiste and Peterson (1959) but including a variation of age and gender of the speakers and of phonation type as well as variation of f_o for voiced sounds, an HP filter experiment was conducted based on the three sound samples of a man, a woman and a child described in the previous chapter, with the sound samples enlarged.

Experiment

Vowel sounds and speakers, and sound selection: The sound sample of the long Standard German vowels of the previous experiment was again used and enlarged by additional voiced sounds of the Zurich Corpus produced by the three speakers in nonstyle mode with medium vocal effort in V context at intended f_o levels of 220–330–440–659 Hz (all speakers). Sound selection of the additional sounds accorded to the procedure as described in the previous experiment. As a result, for each speaker and each vowel, the speaker-specific subsamples created consisted of one whispered and one breathy sound and of voiced sounds produced at intended f_o of 131–220–262–330–440–523–659 Hz or 220–262–330–440–523–659 Hz, respectively (nine sounds per vowel and a total of 72 sounds for the man and eight sounds per vowel and a total of 64 sounds each for the woman and the child). Thus, a sample of 200 natural sounds was investigated. (See below for more information on the reasons for the sample extension. Note again that, according to the standard procedure of the Zurich Corpus, the speakers spontaneously produced vowel sounds with breathy phonation at speaker-related levels of f_o.) Except for 11 sounds, the recognition rate for the selected sounds was 100% according to the standard listening test conducted when creating the Zurich Corpus, matching vowel intention. The exceptions concerned two breathy sounds (80% and 60% recognition rate) and nine voiced sounds (five sounds with 80%, three sounds with 60% and one sound with 40% recognition rate; see the sounds marked in Table 1, Column 2).

HP sound filtering: HP filtering was applied to all selected natural sounds with CFs of 440–660–990–1320 Hz using the Praat filter functionality (see the Praat manual, Sound: Filter [pass Hann band], with smoothing = 100 Hz). As a result, a sample of 800 filtered sounds (288 sounds related to the original reference sounds of the man and 256 sounds related to the original reference sounds of each the woman and the child) was created for the vowel recognition test.

Note that all sounds were filtered with all four CFs to maintain the ratio of sounds per vowel in the listening test, although filtering a sound with a CF below the level of f_o of sound production did not cause a vowel-relevant change in the spectral characteristics of that sound. However, the vowel recognition results below are given for sounds filtered with a CF substantially above the f_o of the natural reference sounds only, with intended f_o of 659 Hz and CF of 660 Hz considered equal frequency levels.

Notably, for the close vowels investigated, a CF of 440 Hz surpassed the average F_1 of the sounds produced at lower f_o as given in formant statistics (see Fant, 1959; Maurer et al., 1992; Pätzold and Simpson, 1997; values for men and women). Similarly, for the sounds of the close-mid vowels, a CF of 660 Hz also surpassed the statistical average F_1. For filtered sounds with a CF of 990 or 1320 Hz, spectral energy was completely lacking for the entire frequency range of assumed F_1 as given in formant statistics for different vowel qualities. On this basis, concerning vowel recognition, the effect of deleting spectral energy in the frequency region of $F1$ of a vowel could be investigated, in combination with the relation of f_o of the sounds and CFs applied. The same held true for $F2$ of back vowels.

Listening test: Vowel recognition of the sounds was investigated in a listening test according to the standard procedure of the Zurich Corpus (including, in addition, the vowel schwa) and involving the five standard listeners of the corpus. The entire sample of filtered sounds was divided into four subsets, whispered sounds, sounds produced at intended f_o of 131–220 Hz and breathy sounds with f_o in this frequency range, sounds produced at intended f_o of 262–330 Hz and breathy sounds with f_o in this frequency range, and sounds produced at intended f_o of 440–523–660 Hz. Subsets were tested separately, with a pause of 15 minutes at a minimum between the listening tests. The subdivision was applied to minimise the possible effect of pitch differences for the vowel recognition task.

Results

Table 1 in the chapter appendix lists the sample of the unfiltered natural reference sounds, including sound links, and shows the vowel recognition results for the HP-filtered sounds. The recognition results are given for the sounds that were HP-filtered with a CF substantially above the f_o of the natural reference sounds. (For replication of HP filtering, refer to the SpecFilt tool in the online corpus; for a simplified summary of occurring vowel quality shift directions and individual quality shifts, see also Part II, the corresponding table in Chapter 8.2.) Table 2 presents exemplary illustrations of the main findings. Below, the results are discussed for clear vowel quality shifts only, that is, for shifts from one quality to another that were labelled by the majority of the listeners. Exceptions to this are some remarks on vowel boundary identifications. For the first account of the shifts found for the sounds of /i, y/, the references to columns and sound numbers in Table 1 are given in a systematic manner. To avoid numerous repetitions and for better readability, references to columns and sound numbers are given in the subsequent accounts for specific aspects only.

Filtered sounds of the close front vowels /i, y/: For the sounds produced at f_o up to 330 Hz, HP filtering with a CF of 440 Hz caused a vowel quality shift in a close–open direction for 11 of the 16 sounds of /i/ and 15 of the 16 sounds of /y/, with the shifts consisting of /i/ to /e/ and of /y/ to /ø/ or /e/ (see Table 1, Columns 5 and 6). In consequence, when filtered, the recognised vowel quality of the majority of the natural sounds produced at f_o up to 330 Hz (below the CF of 440 Hz applied) shifted in a close–open direction and differed from the quality of the unfiltered sounds produced at f_o of 440–523–659 Hz (equal to or above the CF applied).

Comparable shifts were found for the sounds produced at f_o up to 523 Hz that were HP-filtered with a CF of 660 Hz: The filtering resulted in a vowel quality shift in a close–open direction for 15 of the 22 sounds of /i/ and 13 of the 22 sounds of /y/, with the shifts consisting of /i/ to /e/ or /ɛ/ and of /y/ to /ø/ or /ɛ/ or /ə/ (see Table 1, Column 7). Again, when filtered, the recognised vowel quality of the majority of the natural sounds produced at f_o up to 523 Hz (below the CF of 660 Hz applied) shifted and differed from the quality of the unfiltered sounds produced at an intended f_o of 659 Hz (equal to the CF applied). However, for two sounds of /y/, HP filtering with a CF of 660 Hz reverted the shift observed for the CF of 440 Hz (see Table 1, sounds 4 and 5).

The effect of HP filtering at a further increased CF of 990 Hz proved to be more dependent on the individual natural sounds: For some sounds, the preceding quality shifts were reverted to the original vowel qualities of the unfiltered sounds, while for other sounds, the preceding close–open shifts remained (see Table 1, Column 8).

With few exceptions, HP filtering with a CF of 1320 Hz caused a reversal of preceding quality shifts for all sounds of both vowels in terms of restoring the intended vowel quality (see Table 1, Column 9).

Notably, no filter effect on vowel quality was found for the whispered sounds of /i/ of the man and the woman.

Filtered sounds of the close-mid front vowels /e, ø/: For the sounds produced at f_o up to 330 Hz, conversely to close vowels, HP filtering with a CF of 440 Hz did not cause a vowel quality shift except for the whispered sound of /ø/ of the child, recognised as /e/.

For the sounds produced at f_o up to 523 Hz, HP filtering with a CF of 660 Hz caused a vowel quality shift in a close–open direction for ten of the 22 sounds of /e/ but only for three of the 22 sounds of /ø/, with the shifts consisting of /e/ or /ø/ to /ɛ/. In consequence, comparable to the sounds of /i/ but for higher levels of CF, the recognised vowel quality of some of the natural sounds produced at f_o up to 523 Hz (below the CF of 660 Hz applied) shifted when filtered and differed from the quality of the unfiltered sounds produced at f_o of 659 Hz (equal to the CF of 660 Hz applied).

HP filtering of sounds with a CF of 990 Hz caused a vowel quality shift in a close–open direction for 14 of the 25 sounds of /e/ and for seven of the 25 sounds of /ø/, with the shifts consisting of /e/ to /ɛ/ and of /ø/ to /ɛ/ or /a/. However, for one sound of /e/ and two sounds of /ø/, reverted or inverted open–close shifts were observed.

Besides increased vowel confusion, HP filtering with a CF of 1320 Hz caused either reverted/inverted open–close shifts or confirmed preceding close–open shifts. The exceptions were one sound of /e/ and two sounds of /ø/, for which the intended vowel quality was maintained for all four CFs applied.

Filtered sounds of the open-mid front vowel /ɛ/: Besides a few vowel boundary recognitions, HP filtering of the sounds with CFs of 440–660–990 Hz had no pronounced effect on the recognised vowel quality. Exceptions were two sounds for which HP filtering with a CF of 440 Hz caused an open–close shift to /e/, which was reverted when filtering with CFs of 660–990 Hz.

HP filtering with a CF of 1320 Hz caused an open–close shift for three of the 25 sounds, and for another nine sounds, vowel boundary recognition or vowel confusion was found.

Filtered sounds of /a/: Besides three cases of vowel boundary recognitions, five cases of the recognition of /ɔ/ and one case of vowel confusion, HP filtering of the sounds with CFs of 440–660–990 Hz again had no pronounced effect on the recognised vowel quality.

For five of the nine filtered sounds of /a/ of the man, HP filtering with a CF of 1320 Hz caused a shift to /ɛ/. For the remaining four sounds, vowel confusion was found. However, for the sounds of /a/ of the woman and the child, which were HP-filtered with a CF of 1320 Hz, vowel recognition was maintained, with only one exception.

Filtered sounds of the close-mid back vowel /o/: For the sounds produced at f_o up to 330 Hz and HP-filtered with a CF of 440 Hz, no pronounced effect on vowel quality was found, comparable to the sounds of the close-mid front vowels /e, ø/.

For the sounds produced up to an f_o of 523 Hz, HP filtering with a CF of 660 Hz caused a close–open vowel quality shift to /ɔ/ for 11 of the 22 sounds. Consequently, when filtered, the recognised vowel quality of some natural sounds produced at f_o up to 523 Hz (below the CF of 660 Hz applied) shifted and differed from the quality of the unfiltered sounds produced at f_o of 659 Hz (equal to the CF of 660 Hz applied), comparable to the findings for the sounds of /e/.

A marked increase in vowel confusion occurred for HP-filtered sounds with CFs of 990–1320 Hz. Besides, HP filtering with a CF of 990 Hz caused or confirmed close–open shifts to /ɔ/ for three sounds, reverted shifts back to /o/ for one sound, and inverted open–close shifts to /u/ for two sounds. For four sounds produced at intended f_o of 523–659 Hz, the intended vowel quality was maintained. Further, HP filtering with a CF of 1320 Hz caused a back–front shift for five of the nine sounds of the man and one of the eight sounds of the woman. For two sounds of the woman and four sounds of the child, sound filtering caused a shift to either /u/ or /a/. None of the sounds were recognised as /o/.

Filtered sounds of the close back vowel /u/: For all natural sounds produced at intended f_o up to 330 Hz, HP filtering with CFs of 440–660–990 Hz either had no pronounced effect on recognised vowel quality or caused an initial vowel quality shift in a close–open direction to /o/ and subsequently a reversed shift back to /u/.

For the natural sounds produced at higher f_o levels, HP filtering with CFs of 660–990 Hz caused a vowel quality shift in a close–open direction for 12 sounds and vowel confusion for two of the 15 sounds. For one sound produced at an f_o of 523 Hz, the vowel quality was maintained for the HP-filtered sound with a CF of 660 Hz.

When the sounds were HP-filtered with a CF of 1320 Hz, a marked increase in vowel confusion occurred. Besides, for seven of the 25 sounds, back–front shifts occurred, and one sound was recognised as /a/.

Discussion

The main goal of the present experiment was to extend the previous investigation discussed in Chapter M8.1 and to address vowel quality recognition for sounds with filtered spectral energy in the lower frequency range. In parallel to the preceding LP filter experiment, here, sounds produced by three speakers different in age or gender and produced with different phonation types and at different f_o levels (for the voiced sounds) were HP-filtered at stepwise increasing CFs < 1.5 kHz. Vowel quality was tested by means of a vowel recognition test. On this basis, the below main results were obtained. Note that these main results have to be understood within the limits of varying vowel recognition rates, vowel boundary recognition and marked changes in sound timbre for the filtered sounds.

Firstly, HP filtering of the frequency region of $F1$, generally assumed to be vowel-related, in many cases did not result in a sound for which vowel quality was lost: According to the labelling majority of the vowel recognition test, for most of the HP-filtered sounds with CFs up to 990 Hz, the intended vowel quality either was maintained or shifted to another quality. If it shifted, an initial close–open shift direction was found for most sounds of /i, y, e/ and for some of the sounds of /ø, o, u/, and an initial open–close shift direction was found for a few sounds of /ɛ/

and /a/. (For exemplary illustrations, see Table 2, Series 1 and 2, vowel recognition results for CFs < 1kHz.)

Secondly, for the majority of the sounds of the close front vowels /i, y/ and some of the sounds of the close back vowel /u/ produced at intended f_o of up to 330 Hz and HP-filtered with a CF of 440 Hz, the vowel quality shifted and differed from the quality of the unfiltered sounds produced at intended f_o of 440–523–659 Hz (equal to or above the CF applied). A similar effect was found for some of the sounds of the close-mid vowels /e, ø, o/ produced at intended f_o of up to 523 Hz and filtered with a CF of 660 Hz when compared with the unfiltered sounds produced at an intended f_o of 659 Hz. Thus, in general terms, the effect of HP filtering of natural close and close-mid vowels depended on the f_o level of the sounds. This finding was to be expected from the many other indications reported that the lower frequency range of the vowel spectrum of natural sounds is related to f_o. It indicated anew that, for natural vowel sounds, no spectral energy distribution in terms of a spectral envelope represents vowel quality independently of f_o. (For exemplary illustration, see Table 2, the sounds of Series 3.)

Thirdly, with increasing CFs, initial close–open shifts for filtered sounds of close and close-mid front vowels were, in many cases, reverted back to the intended vowel qualities of the unfiltered sounds (above all for natural sounds of close front vowels) or even inverted from close-mid to close vowels. Thus, it was again demonstrated that numerous natural sounds of close and close-mid front vowels remained recognisable even if the entire frequency range of statistical *F*1 of all vowels of a language is HP-filtered, that is, energy in the lower frequency range commonly assumed to be vowel quality-specific was not a general precondition for vowel recognition. (For exemplary illustration, see Table 2, the sounds of the front vowels in Series 2, and the sounds of Series 4; also consider the sounds of the front vowels in Series 1.)

Fourthly, HP filtering sounds with a CF of 1320 Hz caused a back–front shift for some sounds of /o, u/ in strong contrast to the assumed vowel-related resonances of vowel production not being represented in the sound spectra. (For exemplary illustration, see Table 2, the sounds of the back vowels in Series 2.)

In conclusion, when sounds were HP-filtered with stepwise increasing CFs, the recognised vowel quality of the filtered sounds of the close and close-mid vowels was either maintained or initially shifted in a close–open direction; subsequently, above all for CFs of 990–1320 Hz, the vowel quality of most sounds of close front vowels and some sounds of close-mid front vowels shifted in the reverse open–close direction, while the quality of some sounds of the close and close-mid back vowels shifted to front vowels. The quality of the filtered sounds of the open-mid front vowel /ɛ/ was mostly maintained with few occurring shifts to adjacent /e/. The quality of the filtered sounds of /a/ produced by the woman and the child was mostly maintained, with a few occurring shifts to adjacent /ɔ/. The quality of the filtered sounds of /a/ produced by the man was mostly maintained up to a CF of 990 Hz, with a subsequent shift to adjacent /ɛ/ when applying a CF of 1320 Hz. Exceptions of maintained vowel quality or of quality shifts which occurred in contrast to these general shift directions were rare. However, numerous vowel confusions (sounds for which no labelling majority was found and for which, in some cases, identifications of more than two adjacent vowel qualities were given) occurred.

Looking at the general shift directions found, if shifts occurred, the effect of HP filtering proved to differ in relation to openness: An initial close–open shift direction was found for close and close-mid vowels and, conversely, an initial open–close shift direction was found for open-mid and open vowels. Further, within the limits of the general shift directions and the role of f_o in this type of filter experiment, CFs and associated vowel quality shifts also varied to some extent for sounds of the same vowel. Hence, once again, the vowel qualities investigated and the individual spectral energy distribution of single sounds of a vowel also have to be accounted for when interpreting and generalising the results.

As discussed in the introduction to this chapter, the early study of Lehiste and Peterson (1959) has already indicated vowel quality-specific effects of HP filtering as well as initial close–open and subsequent reverted open–close shifts for HP-filtered sounds of front vowels with stepwise increasing CFs. Notably, their study has also documented recognisable vowel sounds with HP-filtered spectral energy below 2 kHz associated with back–front shifts for filtered sounds of back vowels. The results of the experiment presented here were in line with these indications.

As an ensemble, the main findings supported the foreground–background thesis: Vowel recognition seems to value a given spectral energy distribution in the higher frequency range in relation to a given spectral energy distribution in the lower frequency range.

In this context, the finding that only the general initial vowel quality shift directions were found to be uniform and predictable but that single initial shifts within the limit of these directions as well as subsequent shifts, above all for CFs of 990–1320 Hz, somewhat varied is an important indication: It supports the notion of the vowel spectrum being non-uniform, and it suggests that no simple average spectral envelope or configuration of harmonics acoustically represents a vowel quality even if the spectral envelope or the harmonic spectrum is related to f_o and/or pitch. The foreground–background character of the vowel sound may thus not only concern a general relation between lower, middle and higher spectral energy, but it may also concern a sound-specific fine structure of this relation. We address this question in the excursus on vowel quality and harmonic spectrum (see Part II).

With regard to future studies of HP filtering natural vowel sounds, the impact of different sound production parameters on vowel recognition such as single speaker characteristics, speaker age and gender, phonation type, f_o of voiced sounds and vocal effort, as well as possible artefacts of sound filtering, have to be investigated in more detail. (In this context, singular cases for which the recognition results are difficult to understand based on existing knowledge, e.g. maintained vowel quality for filtered whispered sounds of front vowels up to CFs of 1320 Hz and maintained vowel quality for filtered whispered and voiced sounds of /u/ up to CFs of 990 Hz, may also be addressed.) Note also that, in the experiment presented here, HP filtering of vowel sounds was investigated for the lower frequency range usually associated with the range of statistical F_1 of all vowels and statistical F_2 of back vowels. The question of vowel recognition for sounds which are HP-filtered with CFs above the range of statistical F_2 of some front vowels is not addressed here. The same holds true for vowel recognition for sounds that are HP-filtered with CFs above the range of statistical F_3 of all vowels (on this matter, see Donai et al., 2016.)

Chapter appendix

Table 1. HP filtering of sounds of the long Standard German vowels: Vowel recognition results. Columns 1–5 = unfiltered natural reference sounds (SP = speakers and speaker group, where m = man, w = woman, c = child; OR = order number of a sound in a vowel-related series; P = phonation type, where v = voiced, b = breathy, w = whispered; fo = intended f_o, in Hz; note that the f_o level of breathy sounds is given according to the scale of the intended levels of the voiced sounds; V/L = intended and recognised vowel quality, and links to the natural reference sounds, in the order of listing). Columns 6–9 = recognised vowel qualities (VR) for the LP-filtered sounds, per CF (summary; vowel recognition rate ≥ 60%). Extended online table: Columns 10 ff. = details of the vowel recognition results (labelling of the five listeners). Colour code: Dark blue = recognised vowel quality matching the quality of the unfiltered natural reference sound; light blue = vowel quality shifts; purple = reverted or inverted vowel quality shifts from a close-mid to a close vowel quality; red = back–front vowel quality shifts. Natural unfiltered sounds with a vowel recognition rate below 100% are marked as follows (see Column 3): “1” = 80% recognition rate, “2” = 60% recognition rate, “3” = 40% recognition rate.
[M-08-02-T01]

Table 2. HP filtering of sounds of the long Standard German vowels: Exemplary illustration of the main findings. Columns 1–4 = unfiltered natural reference sounds (S = sound series; V = intended and recognised vowel quality; SP = speaker and speaker group, where m = man, w = woman, c = child; fo = f_o intended, in Hz). Column 5 = CFs applied (in Hz). Column 6 = vowel recognition of the unfiltered and filtered sounds (VR; vowel recognition rate ≥ 60%) and links to the series of unfiltered and filtered sounds (L). Column 7 = exemplification (illustrated main findings).
[M-08-02-T02]

Table 1. HP filtering of sounds of the long Standard German vowels: Vowel recognition results. [M-08-02-T01] Extended online table: ↗

Sounds					VR per CF (Hz)			
SP	ON	P	fo (Hz)	V/L	440	660	990	1320
Man	1	w	–	i ↗	i	i	i	i
	2	v	131		(e–ε)	ε	i	i
	3	b	220		e			i
	4	v	220		e	ε	i	i
	5	v	262		e		i	i
	6	v	330		e	e	(i–e)	i
	7	v	440			e	i	i
	8	v	523			e	(e–ε)	i
	9	v	659				e	
Woman	10	w	–		i	i	i	i
	11	v	220		e		i	i
	12	b	262		e			i
	13	v	262		e	e	(i–e)	i
	14	v	330		e	e	i	i
	15	v	440			e	ε	
	16	v	523			e	e	i
	17	v	659				i	i
Child	18	w	–		i	e	i	i
	19	v	220		(i–e)	(i–e)	i	i
	20	b	262		e	e	i	i
	21	v	262		e	ε	ε	i
	22	v	330		e	e		i
	23	v	440			e	e	
	24	v	523			e	ε	
	25	v	659					e
Man	1	w	–	y ↗	ø		y	y
	2	v	131		ø		y	y
	3	b	220		ø	ε	y	
	4	v	220		e	y	y	y
	5	v	262		ø	y	y	y
	6	v[1]	330		e		y	y
	7	v	440				ø	y
	8	v	523				ε	y
	9	v	659				y	y
Woman	10	w	–			ε		y
	11	v	220		ø	ə	y	y
	12	b	262		ø	ε	ε	y
	13	v[2]	262		ø			
	14	v	330		ø	ø	ø	y
	15	v	440			ø	ø	ø
	16	v	523			ø	ø	y
	17	v	659				ø	y
Child	18	w	–		e	e		
	19	v	220		ø	ø	y	y
	20	b	262		ø	ø		y
	21	v	262		ø			y
	22	v	330		ø	ø		ø
	23	v	440			ø	ε	
	24	v	523			ø	ø	y
	25	v	659					y
Man	1	w	–	e ↗	e	e		
	2	v	131		e	ε	e	i
	3	b	220		e	ε		
	4	v	220		e	ε		e
	5	v	262		e	ε		e
	6	v	330		e	e	e	e
	7	v	440			ε	ε	
	8	v	523			e	ε	
	9	v[1]	659				ε	
Woman	10	w	–		e	ε	ε	
	11	v	220		e			i
	12	b	262		e	ε	ε	e
	13	v	262		e	ε	(e–ε)	e
	14	v	330		e	e	ε	
	15	v	440			e	ε	e
	16	v	523			e	e	
	17	v	659				(e–ε)	ε
Child	18	w	–		e	e		
	19	v	220		e	ε	ε	i
	20	b	262		e	e	ε	e
	21	v	262		e	ε	ε	
	22	v	330		e	e	ε	
	23	v	440			e	ε	ε
	24	v	523			e	ε	ε
	25	v	659				ε	ε
Man	1	w	–	ø ↗	ø	ø	y	y
	2	v	131		ø	ε	ø	ø
	3	b	220		ø	ε		y
	4	v	220		ø	ø	(y–ø)	y
	5	v	262		ø	ε		y
	6	v	330		ø	(ø–ε)		
	7	v	440				ε	ø
	8	v	523			ø	ε	
	9	v	659				ε	ε
Woman	10	w	–		ø		ø	
	11	v	220		ø	ø	ø	ø
	12	b	262		ø	ø	ø	y
	13	v	262		ø		ø	
	14	v	330		ø	ø	ø	ø
	15	v	440			ø		
	16	v	523			ø	a	y
	17	v	659				ε	
Child	18	w	–		e		ε	
	19	v	220		ø	ø		
	20	b	262		ø	ø		y
	21	v	262		ø	ø		
	22	v	330		ø	ø	a	a
	23	v	440			ø		
	24	v	523			ø		
	25	v	659				ø	

Table 1 (continuation). [M-08-02-T01]

Sounds					VR per CF (Hz)			
SP	ON	P	fo (Hz)	V/L	440	660	990	1320
Man	1	w	–		(e–ɛ)	ɛ	ɛ	ɛ
	2	v	131		ɛ	ɛ		e
	3	b	220		ɛ	ɛ	ɛ	ɛ
	4	v	220		ɛ	ɛ	ɛ	
	5	v	262		ɛ	ɛ	ɛ	
	6	v	330		ɛ	ɛ	ɛ	e
	7	v	440			ɛ	ɛ	ɛ
	8	v	523			ɛ	ɛ	
	9	v	659				ɛ	
Woman	10	w	–		ɛ	ɛ	ɛ	ɛ
	11	v	220		e	ɛ	ɛ	
	12	b	262			ɛ	ɛ	ɛ
	13	v	262	ɛ	e	ɛ	ɛ	e
	14	v	330	↗		ɛ	ɛ	
	15	v	440			ɛ	ɛ	
	16	v	523			ɛ	ɛ	ɛ
	17	v	659				ɛ	ɛ
Child	18	w	–		ɛ	ɛ	ɛ	ɛ
	19	v	220		ɛ	ɛ	ɛ	
	20	b	262		ɛ	ɛ	ɛ	ɛ
	21	v	262		ɛ	ɛ	ɛ	ɛ
	22	v	330		ɛ	ɛ	ɛ	ɛ
	23	v^2	440				ɛ	ɛ
	24	v	523			ɛ	ɛ	(ɛ–a)
	25	v	659				ɛ	ɛ
Man	1	w	–		a	a	a	ɛ
	2	v	131		a	a	a	
	3	b	220		a	a	a	ɛ
	4	v	220		a	a	a	
	5	v	262		a	a	a	
	6	v	330		ɔ	a	a	ɛ
	7	v	440			(a–ɔ)	a	ɛ
	8	v	523			a	a	
	9	v	659				a	ɛ
Woman	10	w	–		a	a	a	a
	11	v	220		a	a	a	
	12	b	262		a	a	a	a
	13	v	262	a	a	a	a	a
	14	v	330	↗	ɔ	a	a	a
	15	v	440			ɔ	a	a
	16	v	523				a	a
	17	v	659				a	a
Child	18	w	–		a	(a–ɔ)	a	a
	19	v	220		a	(a–ɔ)	a	a
	20	b	262		a	a	a	a
	21	v	262		a	a	a	a
	22	v	330		a	a	a	a
	23	v	440			ɔ	a	a
	24	v	523			ɔ	a	a
	25	v	659				a	a

Sounds					VR per CF (Hz)			
SP	ON	P	fo (Hz)	V/L	440	660	990	1320
Man	1	w	–		o	ɔ		y
	2	v	131		o	ɔ		
	3	b^2	220		o	ɔ		y
	4	v	220		o	ɔ		y
	5	v	262		o	ɔ		i
	6	v	330		o			
	7	v	440			o		
	8	v^1	523			ɔ	o	
	9	v^2	659					ɛ
Woman	10	w	–		o	o		
	11	v	220		o	ɔ		
	12	b	262		o	ɔ		
	13	v	262	o	o	ɔ	ɔ	
	14	v	330	↗	o	o		y
	15	v	440			o	ɔ	
	16	v	523			o	o	u
	17	v^1	659				o	a
Child	18	w	–		o		u	
	19	v	220		o	ɔ	u	u
	20	b	262		o	o		
	21	v	262		o	ɔ		
	22	v	330		o	o	ɔ	a
	23	v	440			o		a
	24	v	523			o	o	a
	25	v^3	659				o	
Man	1	w	–		u	u	u	
	2	v	131		o	u	u	y
	3	b	220		o	u	u	
	4	v	220		u	u	u	y
	5	v	262		o	u	u	y
	6	v	330		u	u	u	y
	7	v	440			o	a	
	8	v	523			o		
	9	v	659					
Woman	10	w	–		u	u	u	
	11	v	220		u	u	u	y
	12	b^1	262		o	u	u	
	13	v	262	u	(o–u)	u	u	
	14	v	330	↗	(o–u)		u	
	15	v	440			o	ɔ	
	16	v^1	523			o	o	
	17	v	659				o	
Child	18	w	–		u	u	u	y
	19	v	220		u	u	u	
	20	b	262		(o–u)	u	u	y
	21	v	262		o	u	u	
	22	v	330		u	o	u	
	23	v	440			o	ɔ	a
	24	v	523			u	o	
	25	v	659				o	

Table 2. HP filtering of sounds of the long Standard German vowels: Exemplary illustration of the main findings. [M-08-02-T02]

Sounds				HP Filtering			Exemplification
S	V	SP	fo (Hz)	CFs (Hz)	VR	L	
1	i	m	w	440–660–990–1320	i–i–i–i–i	↗	Maintained vowel quality.
	e	m	330		e–e–e–e–e	↗	
	ɛ	c	262		ɛ–ɛ–ɛ–ɛ–ɛ	↗	
	a	c	330		a–a–a–a–a	↗	
2	i	m	131	440–660–990–1320	i–(e-ɛ)–ɛ–i–i	↗	Initial close–open shift; subsequent reverted open–close shift (sounds of front vowels) or back–front shift (sounds of back vowels).
	e	w	262		e–e–ɛ–(e-ɛ)–e	↗	
	o	m	262	440–660–1320	o–o–ɔ–i	↗	
	u	m	262	440–1320	u–o–y	↗	
3	i	m	131	440	i–(e-ɛ)	↗	General relation of the lower vowel spectrum and fo, and specific relation of HP filtering of sounds of close and close-mid vowels and fo of the natural sounds.
	i	m	440	–	i	↗	
	e	w	262	660	e–ɛ	↗	
	e	w	660	–	e	↗	
	u	m	131	440	u–o	↗	
	u	m	440	–	u	↗	
4	i	c/w/m	var	1320	i	↗	Recognisable sounds of close front vowels with the entire frequency range of statistical *F*1 of all vowels being HP filtered.
	y	c/w/m	var	1320	y	↗	

M9 Resonance Characteristics of Vowel Sound Production and Their Detectability in the Acoustic Analysis of Radiated Sound

M9.1 Questioning the Direct Relation Between Resonances of Vowel Sound Production and Estimated Resonance Characteristics of Radiated Sound

Introduction

According to the prevailing theory, vowel sounds generally mirror the resonance pattern of sound production in a direct way. However, objections brought forward in the literature (above all concerning limitations of formant and spectral shape estimation) and the many phenomena discussed in this treatise give reason to question such a direct mirroring. The following considerations summarise some of these objections and phenomena. (Note in this context the controversial debate in the literature on the relation between perception and production of speech; to give two examples of reflections that run counter to a direct relation, see the text entitled "Speech perception is hearing sounds, not tongues" by Ohala, 1996, and Pardo and Remez, 2021, arguing that perception and production of speech are "coordinated but neither reciprocal nor recruited for each other's function".)

Conceptual considerations: From a physical perspective, the effect of a resonance pattern is quasi-independent of the source sound it transforms. Thus, within a purely physical concept of resonances, the observation that vowel-related spectral characteristics in general and the spectral envelope in particular (if its estimation is methodologically substantiated) of natural vowel sounds relate to the f_o of sound production – and that the spectral envelope is, therefore, an ambiguous representation of vowel quality – is hard to comprehend.

Further experimental findings presented in this treatise, indicating that the observed relation of the spectral envelope of natural vowel sounds to the f_o of sound production is to be explained by the fact that the vowel spectrum is related to pitch (or to its alternative), accentuate the above statement: Pitch is not an acoustic characteristic and, therefore, the relation of the vowel spectrum to pitch cannot be understood within a primarily physical model such as the source–filter model of sound production.

The temptation to assume that, for natural sounds of a single speaker, the resonances of the vocal tract are directly adapted (related) to the pitch and f_o of the source sound within the process of sound production is confronted with several counterarguments. Three main objections are made here by referring to numerous sound examples and experimental results presented in this treatise: Firstly, one has to reflect on the fact that the idea of resonance adaptation to f_o and pitch is a defensive reaction to remain within the prevailing source–filter model while, at the same time, introducing a relation of source and filter which is alien to this very model. (Note that the f_o and pitch relation in question here is not comparable to a limited source–filter interaction as discussed in the literature, neither in its general character nor in its extent and non-systematic spectral manifestation; for a summary of the debate on source–filter interaction, see Titze and Palaparthi, 2016; on this matter, see also de Cheveigné and Kawahara, 1999.) Secondly, referring to the finding of formant pattern and spectral shape ambiguity, the vocal tract configuration of a sound of /e/ produced at f_o of c. 200 Hz would be, in numerous cases, similar to a configuration of a sound of /i/ produced at f_o of c. 400 Hz, and vice versa, which has no plausibility in our everyday articulation of speech. The same holds true for most sounds of close-mid and close vowels and also for some of the sounds of open-mid vowels if f_o variation is extended. Thirdly, the observable extent of variation and nonuniformity of vowel-related spectral characteristics is considered here as being too complex with respect to a direct and unmediated adaptation of the vocal tract to f_o and pitch. In our documentation and experimentation, for example, the spectral envelope for most sounds of /a/ was not indicated to vary markedly when f_o was changed from 200 to 400 Hz, much in contrast to the above example of /e/ and /i/. In consequence, the production process would not only have to relate the resonance configuration of the vocal tract to f_o and pitch but also to a specific vowel quality. Furthermore, to give another example, if the estimated filter curve for some sounds of /o/ produced at f_o of 200 Hz is indicated to resemble the estimated filter curve of sounds of /u/ produced at f_o of 400 Hz, this may not be true for other sounds of /o/ due to the impact of the spectral fine structure on vowel recognition (for an exemplary illustration, see e.g. Figure 1 in the appendix to Chapter M6.10; on the general matter of the role of the spectral fine structure in the acoustic representation of vowel quality, see Chapters M7.7 and M7.8). Hence, the production process would not only have to relate resonance configurations of the vocal tract to pitch/f_o and a specific vowel quality but also to a specific configuration of harmonics and their levels of radiated sound. Finally, to give a last example, if the vocal tract configuration of a sound of /e/ produced at an f_o level of c. 200 Hz were similar to the configuration of a sound of /i/ produced at f_o of c. 400 Hz, on the contrary, the configuration of a sound of /e/ produced at c. 100 Hz would still be the same as the configuration of a sound of that vowel produced at c. 200 Hz, according to the indications of the sound examples documented and the experimental results provided in this treatise. In consequence, the production process would also have to take into account the specific frequency ranges and levels of f_o. In short, we conclude that the complexity level of adaptation that articulation would have to undergo to embrace both the observable variation extent and nonuniformity of vowel quality-specific spectral characteristics is so high that no speculation on a systematic source–filter interaction and filter adaptation to pitch and f_o within the existing concept of the prevailing source–filter theory should be asserted until a thorough experimental investigation of vowel sound production is made that includes an extensive variation of production parameters (see also below, the study of Maurer et al., 1993). In this context, note also further aspects of the nonuniform relation between the sound spectrum and recognised vowel quality, discussed in Chapter M7, such as different occurring numbers of spectral peaks for sounds of a vowel, inversions of spectral peak structures, flat or sloping vowel-related spectral portions and very pronounced spectral variations because of different vocal efforts.

In conclusion, the vowel–pitch relation cannot be understood within a concept according to which spectral characteristics of vowel sounds mirror vowel quality-specific configurations of vocal tract resonances in a direct, unmediated way.

A note on the concept of "formant tuning": Some studies investigating the relation between resonances of the vocal tract, different f_o levels and spectral characteristics of the radiated vowel sounds indeed seem to confirm an adaptation of vocal tract resonances to f_o, above all for European classical singing. This adaptation was termed "resonance tuning" or "formant tuning" (on the matter, see Sundberg, 1987, pp. 124–129; Joliveau et al., 2004a, b; Wolfe et al., 2009, 2020) and was described as a technique to avoid f_o surpassing the frequency of the first vocal tract resonance, with the benefit of "obtaining greater power for a given effort, and also perhaps avoiding the effects on the vocal timbre of having a fundamental whose amplitude varied strongly from note to note or vowel to vowel" (Joliveau et al., 2004b), the drawback being a loss in vowel intelligibility. (However, see Echternach et al., 2015, for sounds at very high f_o for which f_o variation is not associated with an adaptation of the vocal tract; see also Vos et al., 2018, for a discussion of different tuning strategies being related to perceived naturalness, vowel quality and the relation between formants and f_o.)

Yet, until now, no empirical evidence has been provided that such “tuning” occurs in everyday speech for a range of 200–500 Hz, for which formant pattern and spectral shape ambiguity occurs. Moreover, the “tuning” is assumed to concern the sounds of a vowel with f_o surpassing R_1. But as shown in Chapter M3, the spectra of many sounds of close-mid vowels produced at f_o of c. 200–220 Hz exhibited a spectral peak in the frequency range of 400–440 Hz. Increasing f_o by one octave in Klatt resynthesis as well as in spectral shape resynthesis resulted in a close-mid–close vowel quality shift. However, f_o then directly matched with the first spectral peak frequency and the first resonance of resynthesis, that is, f_o did not surpass the assumed R_1. The same held true for many sounds of close vowels produced at f_o of c. 400–440 with a spectral peak at c. 400–440 Hz, for which decreasing f_o by one octave caused a close–close-mid vowel quality shift although f_o then was an octave below the first resonance of resynthesis. (For exemplary illustration, see Chapter M3.2, Table 2, Series 5 and 1; also consider the opposing sounds of /e/ and /i/ in Chapters M6.9 and M6.10.) Sounds of this kind are not addressed by the above discussion of “resonance tuning”, but they represent the core question of formant pattern and spectral shape ambiguity.

A preliminary study of the relation between vocal tract configuration and spectral characteristics of radiated sound: In an earlier study, we examined the relation between vocal tract configuration and vowel sound production using electromagnetic articulography (Maurer et al., 1993). In this study, three types of utterances of two speakers (a woman and a man) were investigated: Type 1 = spontaneous utterances of words including the five long Standard German vowels /i, e, a, o, u/; Type 2 = production of sustained sounds of these five vowels at f_o of spontaneous speech level (V context), moving the articulators as far as possible during the vocalisation but maintaining vowel quality; Type 3 = repeating these vocalisations of sustained vowel sounds with articulator movements but at different levels of f_o. The positions of the tongue blade, tongue tip, upper and lower lips and mandible were measured during the sound production, a spectral analysis of the sounds was conducted, and the vowel quality of words and isolated vowel sounds was tested in a listening test. For spontaneous words, the sounds of each vowel were found to correspond to different positions of the articulators, as was to be expected from prevailing theory. However, comparing the possible extent of vocal tract movement trajectories associated with the vowel sounds produced as isolated sounds (Types 2 and 3), we found that (i) large movements of the articulators could be performed without substantially affecting the spectrum of the radiated sound or the recognised vowel quality, that is, two very different articulator positions were found to represent the same vowel sound, (ii) similar positions were found for sounds of a particular vowel quality produced at different f_o with different spectral peaks, (iii) partly overlapping vocal tract trajectories were found for sounds of different vowels.

In conclusion, there was no empirical evidence found for a direct and unmediated relation between vocal tract resonance configurations and estimated *F*- and *H*-patterns and spectral envelopes of the radiated vowel sounds.

Methodological considerations: In his treatise entitled “Arguments against formants in the auditory representation of speech”, Bladon (1982) wrote: “It is a familiar but inadequately emphasized fact that formants, in the sense of underlying poles in the complex frequency domain, are notoriously elusive in any physical representation of the speech wave. […] Next, even where physical formants are determinable, there may be a poor correlation with the hypothetical auditory ‘formant’-like percept. […] problems of auditory ‘formant’ determination arise at even moderately high fundamental frequencies. […] In conclusion with this section, it emerges quite clearly that, except in some ideal cases, a measured acoustic formant may differ in rather complex ways from a hypothetical auditory ‘formant’ percept.” (Note that the term formant is given in quotation marks in his text.)

Taking up this reflection, further aspects that have already been mentioned should again be taken into account when considering the persisting lack of methodological substantiation not only for formant pattern estimation but, at the same time and to the same extent, also for spectral shape estimation:

– As discussed and referenced in the Preliminaries (Chapters 6 and M6; see also Ladefoged, 1967; Hillenbrand et al., 1995), formant patterns are generally estimated by means of an interactive measurement procedure involving general phonetic knowledge, the selection of software, analytical skill on the part of the examiner, context information (above all age/size and gender of the speaker), visual crosschecks of calculated values based on the sound spectrum and spectrogram, sometimes related to changes of parameter settings and recalculation of the patterns, and manual interpolations of calculated formant tracks. Therefore, “[…] current methods of formant analysis presuppose that researchers have the necessary analytical skills, that is, a knowledge of the existing phonetic principles and rules of interpretation as well as extensive first-hand experience in conducting this type of analysis. This requires prior training because such an analysis involves contextual knowledge, the ability to visually compare numerical values with a corresponding sound spectrogram, together with the ability to interpret the latter visually, and also the skills to vary filter settings interactively and to perform the repetition of numerical analysis. Consequently, methods of formant analysis are not completely objectifiable. If they were, then researchers would play no part as individuals in such research." (Preliminaries, p. 48) If methods of formant analysis were completely objectifiable, formant measurement could be automated. But even though LPC analysis has replaced spectrographic measurement, the same interactive procedure and inherent circularity remain in the method of formant pattern estimation, i.e., the "[…] necessity of having to prejudge the answer before examining the acoustic data". (Ladefoged, 1967, p. 86)

– However, even accepting the lack of automatic calculation of *F*-patterns and the circularity in the *F*-pattern estimation method, prevailing measurement procedures further lose methodological substantiation with increasing f_o levels due to a spectral "undersampling" of the (assumed) resonance curve of sound production, and if f_o of the vowel sounds surpasses c. 300 Hz, substantiation is lacking. The same holds true for the spectral envelope. (On this matter, see de Cheveigné and Kawahara, 1999; Hillenbrand and Houde, 2003; see also Ladefoged, 1967, pp. 80–81; Monsen and Engebretson, 1983; Fereirra, 2007; Swerdlin et al., 2010; Preliminaries, Chapters 6 and M6.) Note that some scholars give even lower critical f_o frequency levels: "[…] formant frequencies are hard to determine when fundamental frequency is higher than about half of the frequency of the first formant" (Sundberg, 1987, pp. 124-125). In the literature, this "undersampling" phenomenon is often understood as being associated with an inevitable impairment of vowel recognition with increasing f_o (for an overview, see Diehl et al., 1996; in this context, see also Joliveau et al., 2004b, who mention that "vowel identifiability is inevitably compromised once f_o exceeds *R*1"), an assumption which is experimentally contradicted by clearly recognisable vowel sounds produced far above an f_o level of 300 Hz, as documented in Chapter M2.

- In addition, and independent of higher levels of f_o, the nonuniform spectral representation of vowel quality and the examination of synthesised sounds with no spectral peak structure add to the methodological problems (see also below).

Contrary to the expectation of Ladefoged (1967, p. 86), the general methodological problem of formant pattern estimation based on the analysis of radiated sound could not be surmounted by developing better analysis procedures, and there are no indications that this problem can ever be resolved. In these terms, we conclude that there is no methodological basis for directly relating the acoustic characteristics of every recognisable vowel sound (as a radiated sound) to a resonance pattern of sound production. Thus, for a substantial number of vowel sounds, their actual resonance or filter characteristics during sound production may not be acoustically detectable in the radiated sound in a direct way. Why, then, should sound perception and recognition directly relate to the resonance or filter characteristics of sound production?

In conclusion and to repeat, there is neither a methodological basis for formant measurement for all recognisable vowel sounds nor evidence of formants being a perceptual cue. Thus, both aspects again counter the thesis that a radiated sound mirrors a specific vocal tract resonance configuration in a direct, unmediated way.

Considerations concerning vowel sound filtering: As shown for LP- and HP-filtered sounds in the previous main chapter, vowel recognition is often not based on model patterns of spectral peaks and does not stand in direct relation to the actual resonance pattern of sound production commonly assumed as being vowel-related.

Considerations concerning vowel synthesis not relating to a spectral peak structure or a spectral envelope: Recognisable vowel sounds can be synthesised with either a lack of any spectral peak structure and/or a lack of a spectral fine structure allowing for the assessment of a spectral envelope. Thus, as stated, a resonance or filter pattern usually assumed to be characteristic of vowel sounds is not a prerequisite for sound production and vowel recognition.

To sum up: (i) The finding that vowel recognition relates to pitch (or to an alternative sound characteristic) and the observable variation extent and nonuniformity of vowel-related spectral characteristics, (ii) the lack of methodological substantiation of formant and spectral shape estimation for all recognisable vowel sounds, (iii) the lack of evidence of formants being a perceptual cue for vowel recognition, (iv) the finding that vowel recognition for filtered sounds differed from intended vowel qualities of the unfiltered sounds and that the acoustic characteristics of some of these filtered sounds did not relate to expected resonances of sound production, and (v) the finding that synthesised vowel sounds produced outside the framework of the prevailing source–filter model are recognisable all stand against the understanding of specific vocal tract resonance configurations always being directly and imperatively mirrored in the radiated vowel sound. This leads to the assumption that the vocal tract resonance configuration is reflected in the produced vowel sound in a mediated way, a topic that has to be addressed and clarified in future research.

Although there is a long-standing and controversial debate on the relation between production and perception, an objection to understanding the vowel spectrum as directly mirroring the resonance pattern of sound production still seems provocative and difficult to accept. We will return to this matter in the next chapter. In this chapter, the discussion is limited to the exposition of the above counterarguments and to the examination and documentation of an additional spectral aspect that emerged in the course of analysing the spectra of natural vowel sounds and creating the documentation provided in Chapters M2.2 and M2.3 (vowel sounds produced at high f_o levels), M5.1 (breathy vowel sounds) and M7.3 (flat or sloping vowel-related spectral portions). While investigating such voiced and breathy sounds, we often observed noise and noise peaks manifest in the spectra parallel to the harmonic series. The noise peaks could be interpreted as a direct indication of a resonance characteristic of sound production. However, in numerous cases, one or several of these noise peaks did not correspond to the course and the peaks of the harmonic spectrum.

To document this observation, a corresponding study was conducted: Three samples of natural vowel sounds were created for which the peaks of the spectral noise were at least in part not in accordance with corresponding peaks or relative energy maxima of the harmonic series: Sounds produced with breathy phonation or with voiced phonation and low vocal effort, sounds which manifested a flat or sloping harmonic envelope in the sound spectrum, and voiced sounds produced at high f_o levels. These three samples represented three different f_o ranges of sound production and concerned three different aspects of the vowel spectrum.

Note that, for this experiment, interpreting occurring noise peaks as being direct indications of resonances of sound production is a hypothetical approach, and if resonances of production are referred to in this chapter, then it is only in this hypothetical sense. The reflection and documentation put forward here aim only to serve as a basis for future experimental designs and clarification.

Experiment

Creation of three sound samples, sound selection criteria: Sounds of the Zurich Corpus were inspected for which the spectra manifested a contrast between the peak structure of noise and the peak structure of the harmonic spectrum (the relative energy maxima within the harmonic configuration or the frequencies of the harmonics) for a frequency range or a part of it that is usually assumed to be vowel-related. Based on the inspection, three samples of sounds produced in V context were compiled, with all selected sounds being fully recognised in the standard listening test conducted when creating the corpus (100% vowel recognition rate matching vowel intention). Further production parameters not explicitly given below were disregarded.

The first sample consisted of sounds of the eight long Standard German vowels produced with breathy phonation or voiced phonation and low vocal effort at intended f_o ranging from 98–392 Hz. Sounds with these characteristics were investigated because they often exhibit a dominant first harmonic and subsequent sloping and/or flat parts of the harmonic spectrum, either for the entire vowel-related frequency range or a substantial part of that range. Therefore, if the noise spectrum manifested a more differentiated peak structure substantially above *H*1, a corresponding contrast could be demonstrated for the harmonic spectrum not directly mirroring the resonances indicated by the manifest noise.

The second sample consisted of sounds of the eight long Standard German vowels produced at intended f_o in the range of 440–587 Hz, manifesting a flat or sloping harmonic spectrum for the entire vowel-related frequency range or for a part of it, with peaks of noise occurring in that frequency range. Sounds with these characteristics were investigated for a similar reason to that mentioned above: If the spectrum manifested noise peaks but no corresponding peaks in the harmonic spectrum, the contrast in question could be demonstrated again.

The third sample consisted of sounds of the corner vowels /u, a, i/ produced at intended f_o of 784–880 Hz, whose spectra showed either a noise peak on a frequency level substantially below the frequency level of *H*1 or noise peaks in between *H*1 and *H*2 and/or *H*2 and *H*3. Sounds with these characteristics were investigated because of the high frequency level of *H*1 and the large frequency distance between the harmonics in the spectrum: If the spectrum exhibited a noise peak on a frequency level substantially below the level of *H*1 or a noise peak in between lower harmonics separated by a large frequency distance, the contrast in question could be demonstrated anew.

Analysis of different types of contrasts between peak structures of noise and harmonics: During sound selection, occurring contrasts between peak structures of noise and harmonics were analysed, described accordingly and classified in terms of different types of spectral contrasts (see below).

Results

Table 1 shows the three sound compilations, including sound links, and lists the occurring types of spectral contrasts between the peak structures of noise and the vowel-related harmonic spectrum.

For the first compilation of breathy and voiced sounds produced with low vocal effort, four types of incongruent noise peaks and energy maxima of harmonics were observed:

A = Noise in the spectrum of sounds of back vowels and /a/ indicated two lower resonances < 1.5 kHz related to sound production; the harmonic spectrum did not exhibit a corresponding distinct spectral double-peak structure with corresponding frequency levels.

B = Noise in the spectrum of sounds of /a/ indicated a resonance in the frequency range of c. 1–1.5 kHz related to sound production (frequency range of statistical F_2); the harmonic spectrum did not exhibit a corresponding distinct second peak at a corresponding frequency level.

C = Noise in the spectrum of sounds of front vowels indicated a lower resonance < 1 kHz related to sound production; the harmonic spectrum did not exhibit a corresponding distinct peak at the corresponding frequency level.

D = Noise in the spectrum of sounds of front vowels indicated resonances > 1 kHz related to sound production in a frequency range usually considered vowel-related; the harmonic spectrum did not exhibit a corresponding distinct and marked peak structure with corresponding frequency levels of pronounced relative energy maxima of the harmonics.

For the second sample of sounds with flat or sloping harmonic spectra, three types of incongruent noise peaks and energy maxima of harmonics were observed:

E = Noise in the spectrum of sounds of /u/ indicated two lower resonances < 1.5 kHz related to sound production; the frequency level of *H*1 was equal to the frequency level of the first resonance, or it occurred in between the two lower resonances, or it was equal to the frequency level of the second resonance; *H*2 manifested a markedly lower level than *H*1 and was substantially above the frequency of the second indicated resonance.

F = Noise in the spectrum of sounds of /o/ indicated two resonances < 1.5 kHz related to sound production; the frequency distance between the first two or three harmonics was large, and one or both indicated lower resonances occurred in between the frequency levels of the lower harmonics.

G = Noise in the spectrum of sounds of a vowel (all vowels except /u/) indicated two or three peaks related to sound production, the peaks being in a frequency range or in parts of that range usually assumed as vowel-related; the harmonic spectrum in this frequency range or a part of it was either flat or sloping.

For the third of sounds produced at high f_o levels, three further types of incongruent noise peaks and energy maxima of harmonics were observed:

H = Noise in the spectrum of sounds of /u/ indicated two lower resonances < 1.5 kHz related to sound production; the frequency level of *H*1 was near or equal to the frequency level of the second indicated resonance and, therefore, the first indicated resonance was not represented in the harmonic spectrum (compare with type E).

I = Noise in the spectrum of sounds of /a/ indicated two resonances < 1.5 kHz related to sound production (in some cases close in frequencies); the frequency distance between *H*1 and *H*2 was large, and one or both of the resonances occurred in between the frequency levels of the lower harmonics (compare with type F).

J = Noise in the spectrum of sounds of /i/ indicated one lower resonance in the range of c. 450–550 Hz related to sound production; the frequency of *H*1 was equal to or above 750 Hz; therefore, the first indicated resonance was not represented in the harmonic spectrum.

Discussion

As a general observational result, numerous sound spectra were found and are documented here for which the peak structure of noise stood in considerable contrast to relative spectral maxima of harmonics or to their frequencies, for a frequency range usually considered vowel-related. If the noise peaks of the documented sounds indeed indicated the actual resonances of sound production, this would support the thesis that the harmonic spectrum does not, *in general,* mirror the resonances of sound production in a direct (unmediated) way. Consequently, numerous cases of vowel sounds with marked differences between vowel-related *R*-patterns of sound production and estimated *F*-patterns and/or patterns of spectral energy maxima would have to be expected to occur. This observation and reflection is transferred into a synthesis experiment in the next chapter.

Chapter appendix

Table 1. One compilation of voiced and breathy and two compilations of voiced vowel sounds produced by children, women and men: Illustration of observed types of spectral contrasts between the peak structure of noise and the harmonic spectrum. Column 1 = sound sample. Columns 2–6 = sounds (S/L = sounds series and sound links; V = intended and recognised vowel quality of the selected sounds; P = phonation type, where v = voiced, b = breathy; fo = range of intended f_o, in Hz; N = number of sounds of a series). Columns 7–10 = occurring types of spectral contrasts between the peak structure of noise and the harmonic spectrum (for details, see text).
[M-09-01-T01]

Table 1. One compilation of voiced and breathy and two compilations of voiced vowel sounds produced by children, women and men: Illustration of observed types of spectral contrasts between the peak structure of noise and the harmonic spectrum. [M-09-01-T01]

Sample	Sounds					Spectral contrast□			
	S/L		P	fo (Hz)	N	A	B	C	D
1: Breathy / voiced sounds low vocal effort	1 ↗	u	v, b	131–392	13	x			
	2 ↗	o	v, b	147–392	10	x			
	3 ↗	a	v, b	147–392	24	x	x		
	4 ↗	ä	v, b	175–392	21			x	x
	5 ↗	ö	v, b	110–392	25			x	x
	6 ↗	e	v, b	98–247	12			x	
	7 ↗	y	v	247–392	13				x
	8 ↗	i	v	330–392	5				x
						E	**F**	**G**	**–**
2: Flat / sloping harmonic spectra / spectral parts	9 ↗	u	v	523–587	6	x			
	10 ↗	o	v	440–587	14		x	x	
	11 ↗	a	v	440–587	17			x	
	12 ↗	ä	v	523–587	4			x	
	13 ↗	ö	v	440–587	17			x	
	14 ↗	e	v	440–587	16			x	
	15 ↗	y	v	440–523	26			x	
	16 ↗	i	v	440–587	10			x	
						H	**I**	**J**	**–**
3: High fo level	17 ↗	u	v	784–880	15	x			
	18 ↗	a	v	784–880	7		x		
	19 ↗	i	v	784	10			x	

M9.2 Resonance Patterns of Sound Production That Differ From Estimated Formant Patterns and Characteristics of the Harmonic Spectrum of Radiated Sounds

Introduction

The observation that the spectra of natural voiced and breathy vowel sounds indicated noise peaks that sometimes did not correspond to the characteristics of the harmonic spectrum led to the question of whether vowel sounds can be produced by means of a vowel synthesis based on resonance or filter patterns that cannot be detected in the acoustic analysis of radiated sounds.

In the context of the methodological problems of *F*-pattern and spectral shape estimation, in the literature, a possible contrast between a resonance or filter pattern of sound production and its detection from the radiated sound is discussed above all concerning f_o levels of voiced sounds and the resulting sampling of the resonance or filter curve (see the previous chapter). As de Cheveigné and Kawahara (1999) state: “The timbre and identity of a sustained vowel are determined by the shape of the vocal tract transfer function, particularly the positions of the first two or three formants. However, the listener has no access to this shape, but only to the waveform or auditory representations derived from it.” With increasing f_o, the resonance or filter curve is progressively undersampled, and sampling is poor for f_o levels above c. 300 Hz. In consequence, as mentioned repeatedly, formant estimation for sounds at middle or higher levels of f_o in terms of detecting resonance characteristics of sound production is often not methodologically substantiated.

However, for middle and higher levels of f_o, the sampling problem is not uniform but depends on whether or not the harmonics of a sound spectrum match the resonance frequencies of sound production. To give an example for sounds of the vowel /a/: If two sounds are produced at equal f_o of 500 Hz but with two different *R*-patterns of 1150–1350–3000 Hz and 1000–1500–3000 Hz, respectively, estimated F_1 will markedly differ from R_1 of sound production for the first sound but not for the second, even if the frequency distance between the harmonics is the same (see also below, Table 1 in the appendix to this chapter, Series 7). This is a consequence of H_1–H_2 matching R_1–R_2 for the second but not for the first sound.

Following this reflection and further developing the experimental design, the question of detectability of resonance or filter characteristics of radiated sounds as characteristics of sound production was addressed in a vowel synthesis model experiment: An attempt was made to synthesise two voiced-like sounds at an equal f_o level but based on two different *R*-patterns in such a way that

- the two radiated sounds manifested similar harmonic spectra and similar estimated *F*-patterns;
- for one sound, part of the measured *F*-pattern markedly deviated from the *R*-pattern of synthesis;
- for the other sound, the entire measured *F*-pattern corresponded with the *R*-pattern of synthesis;
- for both sounds, the harmonic spectra and estimated *F*-patterns were comparable to spectra of natural vowel sounds as documented in the Zurich Corpus.

If two sounds can be synthesised with two different *R*-patterns in such a way that the resulting harmonic spectra and measured *F*-patterns of the radiated sounds are similar, then cases occur for which resonances of production cannot be unambiguously detected on the basis of the radiated and perceived sounds. If, in addition, sound pairs of this kind are recognised as vowel sounds, then cases of *R*-patterns of sound production that are undetectable in the radiated sounds are relevant for an acoustic theory of the vowel.

Experiment

f_o ranges of synthesised sounds and vowel qualities investigated: When we began, in a pre-study, to investigate synthesised replicas of natural breathy and voiced vowel sounds produced with a low vocal effort at lower or middle f_o levels, the replication of the observations reported in the previous chapter proved to be difficult using the Klatt synthesis technique. Therefore, the design of the present experiment was based on the observations reported for sounds produced at f_o $\geq$ 400 Hz, limiting the upper f_o in synthesis to 700 Hz. Furthermore, when attempting a reproduction of sounds at $f_o \geq$ 400 Hz using a Klatt synthesiser, a synthetic replication of the types of incongruent noise peaks and energy maxima of harmonics related to one or two resonances $\leq$ 1.5 kHz proved to be far more feasible than the replication of the types related to vowel-related resonances including frequencies > 1.5 kHz. Therefore, the experiment addressed sounds of /u, o, a/ only.

Creation of model pairs of different *R*-patterns for the synthesis of sounds with comparable harmonic spectra and estimated *F*-patterns: Thus, as a first step, vowel synthesis based on various configurations of pairs of R_1–R_2 patterns and f_o levels in the range of 400–700 Hz was investigated by the author by means of a trial-and-error approach, attempting configurations according to the further developed experimental design described above. Thereby, R_3 was always set to 3000 Hz. Since middle and higher f_o levels were investigated, R_4–R_5 were set to 4200–5400 Hz (see Chapter M3.1, synthesis parameters for women). Bandwidths of lower resonances ≤ 1.5 kHz and spectral tilt were set individually for each *R*-pattern to bring the harmonic spectra and estimated formant frequencies of a sound pair close to each other. Bandwidths of higher resonances were set to 100 Hz. Based on the experiences of this first investigation, in a second step, eight exemplary pairs of configurations of *R*-patterns and f_o levels fulfilling the above conditions of the further developed experimental design were created for the final sound synthesis, acoustic analysis and vowel recognition test (see Table 1 in the chapter appendix, Columns 1–10).

Sound synthesis: Based on the created eight pairs of *R*-patterns and related f_o levels, 1 sec. steady-state sounds (no f_o contour) including a fade-in/fade-out of 0.05 sec. were synthesised using the KlattSyn tool (Klatt synthesis, cascade mode; default parameters, with the below exceptions of sound-specific parameters for glottal source, flutter, breathiness and aspiration). For every single configuration of an f_o level and an *R*-pattern, three sounds with sound-specific parameters were produced. Non-default parameter settings for sound 1 (hereafter voiced-like sound) were: glottal source = impulsive; flutter = 0; breathiness = -50; aspiration = -50 (levels of breathiness and aspiration were lowered to improve the sound quality of the voiced-like sounds of comparison). Non-default parameter settings for sound 2 (hereafter whispered-like sound) were: glottal source = white noise; flutter = 0; breathiness = -25; aspiration = -25. Non-default parameter settings for sound 3 (hereafter illustrating sound) were: glottal source = impulsive; flutter = 0; breathiness = -25; aspiration = -25. Tilt was set according to a given *R*-pattern (see Table 1). Sounds 1 and 2 were investigated concerning acoustic analysis and vowel recognition. Sound 3 and its spectrum only served as a graphic illustration of the contrast between resonance characteristics of sound production (in most cases visible based on the noise in the spectrum related to breathiness and aspiration) and the harmonic spectrum of the sound produced. As a result, for the eight pairs of *R*-patterns and related f_o levels, a total sample of 48 synthesised sounds was created, of which 32 sounds (16 voiced-like sounds, see Table 1, and 16 whispered-like sounds, see Table 2) were compiled for acoustic analysis and a vowel recognition test, and 16 additional sounds were used for documentary purposes only.

When designing the experiment, whispered-like sounds were included for two reasons: Firstly, in most cases, a synthesis with a noise source results in a sound that approximately reflects the resonances of production in the sound spectrum in terms of noise peaks at frequencies corresponding to the *R*-pattern of synthesis. Consequently, the spectral similarity or dissimilarity of two synthesised sounds based on a single *R*-pattern but with two different sources, periodic and noise, can be demonstrated graphically. Secondly, the vowel recognition for both types of sounds can be tested and compared with each other.

***F*-pattern estimation:** Acoustic analysis was conducted for the synthesised sounds according to the standard procedure of the Zurich Corpus, including a crosscheck of the calculated *F*-patterns based on sound spectra, spectrograms and formant tracks. In the crosscheck, calculated values were kept that related to one of the three parameter settings of LPC analysis for the maximum number of formants providing the best match of F_1–F_2 with spectrum, spectrogram and formant tracks (6, 5 or 4 formants at a maximum for the frequency range of 0–5.5 kHz; see Column 19 in Tables 1 and 2). Values associated with large formant bandwidths > 450 Hz were disregarded.

Listening test: Vowel recognition of the sounds was tested in a listening test according to the standard procedure of the Zurich Corpus (forced choice, all long Standard German vowels and schwa, no vowel boundaries) and involving the five standard listeners. The test was divided into two subtests separating the two source types, and every single sound was presented twice within the corresponding subtest.

Results

For the synthesised voice- and whispered-like sounds separately, Tables 1 and 2 in the chapter appendix list the eight pairs of *R*-patterns and related f_o levels investigated in vowel synthesis and show the results of the acoustic analysis of the synthesised sounds and their pairwise spectral comparison. Table 3 shows the results of the vowel recognition test. The links to the voiced-like sound pairs are given in Table 1, the links to the whispered-like sound pairs are given in Table 2, and the links to all sounds of a pair of *R*-patterns, including the sounds for documentary purposes, are given in Table 3.

Results of spectral comparison of the voiced-like sounds (Table 1): According to the results of the acoustic analysis of the voiced-like sound pairs, for all eight pairs investigated, the first sound “a” of a pair showed estimated F_1 or F_2 or F_1–F_2 markedly deviating from R_1 or R_2 or R_1–R_2 of synthesis (see Columns 5, 7, 11 and 12, and 15 and 16, values marked in red), the differences between F_1 and R_1 being in the range of 131–328 Hz and the differences between F_2 and R_2 being in the range of 94–251 Hz (see Columns 12 and 16). Conversely, the second sound “b” of a pair showed corresponding F_1–F_2 and R_1–R_2, the differences between F_1 and R_1 being in the range of 0–33 Hz and the differences between F_2 and R_2 being in the range of 4–44 H (see Columns 13 and 17, values marked in dark blue). Finally, a comparison of the estimated *F*-patterns of both sounds “a” and “b” of a pair showed that these patterns also matched, the differences for F_1 being in the range of 6–22 Hz and the differences for F_2 being in the range of 0–13 Hz (see Columns 13 and 14, and 17 and 18, with the values in Columns 14 and 18 marked in light blue). Likewise, the harmonic spectra of the two voiced-like sounds of a pair corresponded with each other (for visual verification, see the sound links in Column 1). Thus, for the first sound “a” of a pair, R_1 or R_2 or R_1–R_2 of production was not identified by its estimated *F*-pattern, but for the second sound “b”, the vowel-related R_1–R_2 pattern of production was identified by the *F*-pattern. Moreover, the *F*-pattern of the first sound resembled the vowel-related *R*- and *F*-patterns of the second sound of a pair.

Besides, the comparison of R_1–R_2 of sound production and H_1–H_2 or H_2–H_3 of the synthesised sounds (see Columns 20–23) showed that vowel sounds can be produced which manifest large frequency distances between the resonance frequencies of production and the frequencies of the harmonics in the spectrum of the radiated sounds as well as, in some cases, also manifesting H_1 markedly above R_1 or H_2 markedly above R_2.

Results of spectral comparison of the whispered-like sounds (Table 2): When crosschecking the calculated *F*-patterns of the whispered-like sounds with noise peaks manifest in the spectrum and the spectrogram, we observed insufficient matches of F_1 and R_1 for all cases of sound production with the lowest resonance ≤ 500 Hz and a bandwidth = 100 Hz. However, the calculated values were kept, but they have to be interpreted as rough approximations. (Notably, the LPC measurement of whispered-like sounds of this type of synthesis turned out to be a further methodological issue.)

Given this measurement limitation, with the exception of Series 1, the results nevertheless indicated that, for the two sounds of a pair, the similarity between an R_1–R_2 pattern of sound production and the related estimated F_1–F_2 of the produced sound (either “a” or “b”) was more pronounced than the similarity between the two estimated F_1–F_2 of a pair: Concerning different resonances for a sound pair (see Columns 5 and 7, values marked in grey), for Series 2–8, values for |F1a - R1a| and |F1b - R1b| were found in the range of 7–91 Hz (see Columns 12 and 13, values marked in grey), while the values for |F1a - F1b| were found in the range of 118–279 Hz (see Column 14, values marked in red). Likewise, for Series 2–4 and 6–8, values for |F2a - R2a| and |F2b - R2b| were found in the range of 2–58 Hz (see Columns 16 and 17, values marked in grey), while the values for |F2a - F2b| were found in the range of 66–207 Hz (see Column 18, values marked in red). For Series 1 only, the estimated *F*-pattern of sound “a” showed somewhat more resemblance to the *R*-pattern and *F*-pattern of sound “b” than to its proper *R*-pattern of sound production (see values marked in purple), yet within a very limited frequency range of the differences in question. Thus, in contrast to the finding for the voiced-like sounds, the difference of the *R*-patterns of sound production with noise as the source was mostly mirrored in a corresponding difference of the estimated *F*-patterns when comparing two sounds of a pair.

Vowel recognition results (Table 3): According to the labelling majority of the vowel recognition test, the recognised vowel qualities of the voiced-like sounds of a pair corresponded to each other and to the qualities of the natural voiced sounds imitated in their spectral characteristics for the Series 1–4 and 7 and 8 (compare Columns 2 and 11). The same held true for the remaining two Series 5 and 6 concerning a vowel boundary of the natural voiced sounds imitated. (Note that the labelling majority for the vowel boundary of /ɔ–o/ was interpreted for the sounds of Series 5, and the labelling majority for the vowel boundary of /o–u/ was interpreted for the sounds of Series 6.) In contrast, for all sounds of the two back vowels investigated and the corresponding pairs of configurations of *R*-patterns, the vowel quality of at least one whispered-like sound differed from that of the related voiced-like sound (see Column 14, indications marked in red). Moreover, for five of the six pairs of whispered-like sounds recognised as back vowels, the vowel qualities also differed according to the two different *R*-patterns of sound production compared (see Series 2–6). Besides, for the sounds recognised as back vowels, the recognition rate somewhat varied among the pairs of *R*-patterns and source characteristics applied in synthesis, and listener-specific recognition inconsistencies occurred.

Discussion

The results of the acoustic analysis showed that two voiced-like sounds manifesting similar harmonic spectra and similar estimated *F*-patterns could be synthesised based on two different *R*-patterns, with dissimilar vowel-related *R*-pattern of sound production and estimated *F*-pattern of radiated sound for one sound of a pair but similar *R*- and *F*-patterns for the other sound. For the synthesised sound pairs, vowel-related *R*-pattern differences were in the range of 150–350 Hz for R_1 (compare values in Table 1, Column 5) and 100–220 Hz for R_2 (compare values in Table 1, Column 7; equal R_1 and R_2 disregarded). These ranges equalled or exceeded the differences for estimated average F_1 and F_2 for sounds of two adjacent back vowels, as they are often given in formant statistics. (For sounds of /o/ and /u/ produced by women or men, for example, statistical average F_1 or F_2 differences are given as < 130 Hz by Fant, 1959, Pätzold and Simpson, 1997, and Hillenbrand et al., 1995.)

The vowel recognition results showed that, according to the labelling majority, the two voiced-like sounds of a pair of *R*-pattern and f_o level configurations were recognised as the same vowel, which confirmed both their spectral similarity and their attempted similarity with spectra of natural sounds but which was in contrast to their dissimilar *R*-patterns of sound production. On the contrary, for five of the six whispered-like sound pairs based on *R*-pattern configurations related to back vowel qualities, the sounds of a pair were recognised as different vowels in parallel to their differences in *R*-pattern of sound production and the corresponding spectral differences. Again, the recognition results for the whispered-like sounds recognised as /a/ deviated from the results of the sounds recognised as back vowels, that is, the recognition results were nonuniform with respect to vowel quality.

The comparison of *R*-patterns and *F*-patterns has to be relativised because of the methodological problem of *F*-pattern estimation for middle and higher f_o levels. Moreover, it may seem inconsistent to refer to the estimation problem and claim that vowel sounds are not, *in general,* characterised by spectral peaks, and then conduct an experiment in which *F*-patterns are estimated for sounds at middle and higher f_o levels and interpret the results thereof. However, the purpose of conducting this experiment was to demonstrate in an exemplary manner that, within a source–filter model, resonances of production may or may not be detected in the acoustic analysis of the radiated sounds. Further, and most importantly, the harmonic spectra of the two voiced-like sounds of a pair as well as their recognised vowel quality were found to be similar, in contrast to the two different *R*-patterns of sound production and the related synthesised whispered-like sounds.

In these terms, exemplary cases of voiced-like sound pairs of a vowel could be synthesised and are demonstrated here for which the vowel-related production resonances of one sound could be detected based on the radiated and perceived sound but could not be detected for the other sound, with equal f_o levels and quasi-equal harmonic spectra and estimated *F*-patterns for the two sounds. Thus, two different *R*-patterns can result in voiced or voiced-like sounds of a vowel with similar harmonic spectra and similar estimated *F*-patterns. This conclusion is further strengthened by the finding that, in the parallel synthesis of whispered-like sound pairs, *R*-pattern differences of production were mostly detected by the estimated *F*-patterns of the radiated sounds, and vowel recognition differed according to different *R*- and *F*-patterns for five of the six sound pairs related to the back vowels /o, u/. Notably, these kinds of occurring undetectable resonances of sound production and the corresponding conclusion that vowel recognition does not directly rely on formant patterns support two of the three major arguments of Bladon (1982) against formants in the auditory representation of speech, labelled as the determinacy and the perceptual adequacy objections.

In the present experiment, the main finding was obtained for different configurations of *R*-patterns for which R_1 was below or equal to or above H_1 and R_2 was below or equal to or above H_2 (see Table 1, Columns 20–23). Also, the f_o level of the voiced-like sounds was equal for each sound pair, and for some voiced sounds "b" of the pairs, R_1–R_2 and harmonic frequencies matched (see Table 1, Series 4 and 6–8, sounds 7b and 8b being the same). Therefore, the finding cannot be explained solely by a general "undersampling" of the resonance curve of production due to middle or high f_o levels.

As stated, although there is a long-standing and controversial debate on the relation between production and perception, scepticism with regard to the understanding of the vowel spectrum as directly mirroring the resonance pattern of sound production still seems difficult to accept. However, the question of that relation inevitably arises when considering the vowel–pitch relation, the nonuniform spectral representation of vowel quality, the lack of a methodological substantiation of *F*-pattern and spectral shape estimation and, as demonstrated, sounds for which resonance characteristics of production cannot unambiguously be detected in the acoustic analysis of radiated sound.

The experiment presented here may serve as a model for future examinations of the relation between the resonances of vowel sound production, the acoustic characteristics of radiated sound and vowel quality recognition. However, future experiments on the matter need new synthesis techniques that allow for a more natural-like quality of the produced sounds in general and for synthesis related to any harmonic envelope or harmonic configuration and any levels and ranges of f_o that are observable in natural vowel sounds in particular. Also, noise-extraction techniques may be used to investigate the perceptual difference between noise- and voiced-related parts of natural vowel sounds. Furthermore, with improved synthesis techniques, it may be possible to produce single sounds based on single *R*-patterns for which two sounds of two different vowels, a voiced-like and a whispered-like sound, can be recognised.

Finally, in the context of the present experimentation, we observed marked pitch level differences for some of the whispered-like sounds. To give an example, the author's estimates for the comparison of the three pairs of whispered-like sounds of Series 2, 4 and 5 of the sample investigated are shown in Table 4 (general level difference, approximate difference in semitones, approximate levels according to the musical C-major scale and corresponding frequency levels in Hz), indicating a pitch range of 175–294 Hz for these sounds. (Notably, for the sounds of Series 2 and 5 and the first sound of Series 4, the estimated pitch level corresponds to approximately half of the first resonance frequency of synthesis. However, this is not the case for the second sound of Series 4.)

Chapter appendix

Table 1. Synthesised voiced-like vowel sound pairs based on dissimilar *R*-patterns: Synthesis parameters investigated and results of acoustic analysis. Columns 1–10 = sounds and sound synthesis (S/L = sound pairs and sound links; V = vowel quality of the natural vowel sounds imitated in synthesis; P = phonation type imitated, source characteristic of synthesis, where v = voiced-like; fo = f_o of synthesis, in Hz; R(i) and B(i) = resonances and bandwidths of synthesis, in Hz; Tilt = tilt in dB; Ra≠Rb = "a" versus "b" difference of R_1 or R_2 or R_1–R_2 of a sound pair). Columns 11–19 = comparison of *R*- and *F*-patterns and their respective difference, in Hz, and indication of the LPC parameters used for formant estimation (Par). Columns 20–23 = comparison of $H_{(i)}$ and $R_{(i)}$ and their respective difference, in Hz. Colour code: Purple = differences in R_1 or R_2 or R_1–R_2 of synthesis; red = estimated F_1 or F_2 or F_1–F_2 deviating from R_1 or R_2 or R_1–R_2 of synthesis; dark blue = approximate match of estimated F_1 or F_2 or F_1–F_2 and R_1 or R_2 or R_1–R_2 of synthesis; light blue = approximate match of the *F*-patterns of both sounds "a" and "b" of a pair despite differences in R_1 or R_2 or R_1–R_2 of synthesis.
[M-09-02-T01]

Table 2. Synthesised whispered-like vowel sound pairs based on the dissimilar *R*-patterns shown in Table 1: Synthesis parameters investigated (repetition) and results of acoustic analysis. Columns, see Table 1 (in Column 3, w = whispered-like). Colour code: For all series except Series 1, for the two sounds of a pair, the similarity of an R_1–R_2 pattern of sound production and the related estimated F_1–F_2 of the produced sound (either "a" or "b") was more pronounced than the similarity of the two estimated F_1–F_2 (for details, see text); in order to highlight this finding, the corresponding values in Columns 5, 7, 11–13 and 15–17 are coloured in grey, and the corresponding values in Columns 14 and 18 are coloured in red. Concerning the exception of the sounds of Series 1 (see text), the corresponding values are coloured in green.
[M-09-02-T02]

Table 3. Synthesised voiced-like and whispered-like vowel sounds based on the dissimilar *R*-patterns shown in Table 1: Vowel recognition results. Columns 1–9 = see corresponding columns in Table 1. Note that, in the links, the voiced-like and whispered-like sounds related to a single *R*-pattern and f_o level configuration of synthesis are preceded by a mixed sound thereof (by the sound produced for illustration purposes only; see text) in order to facilitate spectral comparison. Columns 10–12 = vowel recognition results for the voiced-like sounds (P = source characteristic of synthesis, where v = voiced-like; V = vowel recognised according to the labelling majority; M = majority of labelling, with ten identifications per sound at a maximum). Columns 13–15 = vowel recognition results for the whispered-like sounds (w). Columns 16–25 = listener-specific details of vowel recognition. Colour code: Purple = differences in R_1 or R_2 or R_1–R_2 of synthesis (see Table 1); dark blue in Column 11 = matching vowel recognition of both voiced-like sounds of a sound pair and matching vowel recognition with the vowel quality of the natural vowel sounds imitated in synthesis; light blue in Column 11 = matching vowel recognition of both voiced-like sounds of a sound pair within a vowel boundary of the vowel quality of the natural vowel sounds imitated in synthesis; dark blue in Column 14 = matching vowel recognition for *R*-pattern-related whispered-like and voiced-like sounds; light blue in Column 14 = matching vowel recognition for *R*-pattern-related whispered-like and voiced-like sounds concerning the vowel boundary /o-u/ and the vowel quality /o/; dark red in Column 14 = mismatch of vowel recognition for the whispered-like sound in comparison to the *R*-pattern-related voiced-like sound as well as for the opposed whispered-like sound of the pair in question.
[M-09-02-T03]

Table 4. Three selected synthesised whispered-like vowel sound pairs based on the dissimilar *R*-patterns shown in Table 1: Approximate comparison of the pitch levels of the opposing sounds of a pair. Selection of Table 3 (see Series 2, 4 and 5). Pitch levels are given as estimates of the author (see text). For Columns 1–11, compare with Table 3. Columns 12–15 = pitch level comparison (Levels = general level difference of comparison; ST = approximate level differences in semitones; MS = approximate levels according to musical C-major scale notation; Hz = frequency levels in Hz according to the musical C-major scale). Column 16 = sound link (L).
[M-09-02-T04]

Table 1. Synthesised voiced-like vowel sound pairs based on dissimilar R-patterns: Synthesis parameters investigated and results of acoustic analysis. [M-09-02-T01]

Sounds and sound synthesis											Comparison of R- and F-patterns (in Hz), and LPC parameters									Comparison of H(i) and R(i)			
S/L		V	P	fo (Hz)	*R*-pattern (Hz)				Tilt (dB)	Ra≠ Rb	F1	|F1a- R1a|	|F1- R1b|	|F1a- F1b|	F2	|F2a- R2a|	|F2- R2b|	|F2a- F2b|	Par	H1:R1	|H1- R1|	H2:R2	|H2- R2|
					R1	B1	R2	B2															
1 ↗	a	u	v	400	400	100	600	100	0	R2	411	11	11	6	714	114	14	6	4	H1=R1	0	H2>R2	200
	b				400	40	700	100	12		405	–	5		720	–	20		4	H1=R1	0	H2>R2	100
2 ↗	a	u	v	525	350	100	700	100	0	R1–R2	516	166	9	6	951	251	31	13	4	H1>R1	175	H2>R2	350
	b				525	40	920	100	0		522	–	3		964	–	44		6	H1=R1	0	H2>R2	130
3 ↗	a	u	v	700	350	100	700	100	0	R1–R2	678	328	22	13	843	143	3	13	4	H1>R1	350	H2>R2	700
	b				700	100	840	100	0		709	–	9		824	–	16		4	H1=R1	0	H2>R2	560
4 ↗	a	o	v	500	750	50	1100	70	24	R1–R2	541	209	41	11	1006	94	6	0	5	H1<R1	250	H2<R2	100
	b				500	100	1000	30	0		530	–	30		1006	–	6		4	H1=R1	0	H2=R2	0
5 ↗	a	o	v	500	350	50	1000	50	0	R1	515	165	55	22	1007	7	7	1	4	H1>R1	150	H2=R2	0
	b				570	50	1000	50	0		537	–	33		1006	–	6		4	H1<R1	70	H2=R2	0
6 ↗	a	o	v	500	700	100	800	100	0	R1–R2	536	164	36	20	997	197	3	10	4	H1<R1	200	H2>R2	200
	b				500	100	1000	100	0		516	–	16		1013	–	13		4	H1=R1	0	H2=R2	0
																				H2:R1	|H2- R1|	H3:R2	|H3- R2|
7 ↗	a	a	v	500	1150	100	1350	100	0	R1–R2	1019	131	19	19	1494	144	6	2	4	H2<R1	150	H3>R2	150
	b				1000	100	1500	100	0		1000	–	0		1504	–	4		4	H2=R1	0	H3=R2	0
8 ↗	a	a	v	500	1200	100	1300	100	0	R1–R2	1018	182	18	18	1494	194	6	2	4	H2<R1	200	H3>R2	200
	b				1000	100	1500	100	0		1000	–	0		1504	–	4		4	H2=R1	0	H3=R2	0

Table 2. Synthesised whispered-like vowel sound pairs based on the dissimilar R-patterns shown in Table 1: Synthesis parameters investigated (repetition) and results of acoustic analysis. [M-09-02-T02]

Sounds and sound synthesis											Comparison of R- and F-patterns (in Hz), and LPC parameters (P)								
S/L		V	P	fo	*R*-pattern (Hz)				Tilt	Ra≠	F1	|F1a-	|F1-	|F1a-	F2	|F2a-	|F2-	|F2a-	Par
				(Hz)	R1	B1	R2	B2	(dB)	Rb		R1a|	R1b|	F1b|		R2a|	R2b|	F2b|	
1 ↗	a	u	w	–	400	100	600	100	0	R2	469	69	69	44	662	62	38	53	6
	b		w	–	400	40	700	100	12		425	–	25		715	–	15		6
2 ↗	a	u	w	–	350	100	700	100	0	R1–R2	431	81	94	130	736	36	184	207	6
	b		w	–	525	40	920	100	0		561	–	36		943	–	23		6
3 ↗	a	u	w	–	350	100	700	100	0	R1–R2	441	91	259	279	738	38	102	140	6
	b		w	–	700	100	840	100	0		720	–	20		878	–	38		6
4 ↗	a	o	w	–	750	50	1100	70	24	R1–R2	743	7	243	190	1078	22	78	66	6
	b		w	–	500	100	1000	30	0		553	–	53		1012	–	12		6
5 ↗	a	o	w	–	350	50	1000	50	0	R1	401	51	169	221	1011	11	11	10	6
	b		w	–	570	50	1000	50	0		622	–	52		1021	–	21		6
6 ↗	a	o	w	–	700	100	800	100	0	R1–R2	714	14	214	134	858	58	142	166	6
	b		w	–	500	100	1000	100	0		580	–	80		1024	–	24		6
7 ↗	a	a	w	–	1150	100	1350	100	0	R1–R2	1133	17	133	118	1348	2	152	148	6
	b		w	–	1000	100	1500	100	0		1015	–	15		1496	–	4		6
8 ↗	a	a	w	–	1200	100	1300	100	0	R1–R2	1186	14	186	171	1314	14	186	182	6
	b		w	–	1000	100	1500	100	0		1015	–	15		1496	–	4		6

Table 3. Synthesised voiced-like and whispered-like vowel sounds based on the dissimilar R-patterns shown in Table 1: Vowel recognition results. [M-09-02-T03]

Sounds and sound synthesis										Vowel recognition						Vowel recognition (details)									
S/L		V	fo	*R*-pattern (Hz)				Tilt	Ra≠	Voiced-like			Whispered-like			Voiced-like					Whispered-like				
			(Hz)	R1	B1	R2	B2	(dB)	Rb	P	V	M	P	V	M	L1	L2	L3	L4	L5	L1	L2	L3	L4	L5
1 ↗	a	u	400	400	100	600	100	0	R2	v	u	10	w	o	8	u u	u u	u u	u u	u u	u u	o o	o o	o o	o o
	b			400	40	700	100	12			u	10		o	9	u u	u u	u u	u u	u u	o o	o o	o o	o u	o o
2 ↗	a	u	525	350	100	700	100	0	R1–R2	v	u	9	w	u	7	u u	u u	u u	u u	u ö	u u	o o	o u	u u	u u
	b			525	40	920	100	0			u	10		o	10	u u	u u	u u	u u	u u	o o	o o	o o	o o	o o
3 ↗	a	u	700	350	100	700	100	0	R1–R2	v	u	7	w	u	7	u u	u u	u u	o u	e e	u u	o o	o u	u u	u u
	b			700	100	840	100	0			u	7		ɔ	10	u u	u u	o u	o u	e u	ɔ ɔ	ɔ ɔ	ɔ ɔ	ɔ ɔ	ɔ ɔ
4 ↗	a	o	500	750	50	1100	70	24	R1–R2	v	o	6	w	ɔ	8	o o	a o	ɔ o	o o	ɔ ɔ	ɔ ɔ	a a	ɔ ɔ	ɔ ɔ	ɔ ɔ
	b			500	100	1000	30	0			o	7		o	6	o o	ɔ o	o o	o o	u u	o o	u ɔ	o o	u u	o o
5 ↗	a	o	500	350	50	1000	50	0	R1	v	ɔ–o	10	w	u	10	o o	ɔ o	ɔ o	ɔ ɔ	ɔ ɔ	u u	u u	u u	u u	u u
	b			570	50	1000	50	0			ɔ–o	9		ɔ–o	10	o o	ɔ ɔ	o o	u o	ɔ ɔ	o o	ɔ ɔ	ɔ ɔ	o o	o o
6 ↗	a	o	500	700	100	800	100	0	R1–R2	v	o–u	10	w	ɔ	9	o o	o o	o o	o u	u u	ɔ ɔ	ɔ ɔ	ɔ ɔ	ɔ ɔ	a ɔ
	b			500	100	1000	100	0			o–u	8		o	8	o o	u u	o u	u u	ɔ ɔ	o o	ɔ ɔ	o o	o o	o o
7 ↗	a	a	500	1150	100	1350	100	0	R1–R2	v	a	10	w	a	10	a a	a a	a a	a a	a a	a a	a a	a a	a a	a a
	b			1000	100	1500	100	0			a	10		a	10	a a	a a	a a	a a	a a	a a	a a	a a	a a	a a
8 ↗	a	a	500	1200	100	1300	100	0	R1–R2	v	a	10	w	a	10	a a	a a	a a	a a	a a	a a	a a	a a	a a	a a
	b			1000	100	1500	100	0			a	10		a	10	a a	a a	a a	a a	a a	a a	a a	a a	a a	a a

Table 4. Three selected synthesised whispered-like vowel sound pairs based on the dissimilar R-patterns shown in Table 1: Approximate comparison of the pitch levels of the opposing sounds of a pair. [M-09-02-T04]

Sounds and sound synthesis									**Vowel recognition**			**Pitch level comparison**				**L**
S		**V**	**fo**	***R*-pattern (Hz)**				**Tilt**	**Whispered-like**			**Levels**	**ST**	**MS**	**Hz**	
			Hz	**R1**	**B1**	**R2**	**B2**	**dB**	**P**	**V**	**M**					
2	a	u	525	350	100	700	100	0	w	u	7	lower	7	F3	175	↗
	b			525	40	920	100	0		o	10	higher		C4	262	
4	a	o	500	750	50	1100	70	24	w	ɔ	8	lower	6	F#Gb3	185	
	b			500	100	1000	30	0		o	6	higher		C4	262	
5	a	o	500	350	50	1000	50	0	w	u	10	lower	9	F3	175	
	b			570	50	1000	50	0		ɔ–o	6	higher		D4	294	

List of Figures
List of Tables
References

Because of the high number of figures and tables shown, the list of figures and tables are presented online. The corresponding links are given below, followed by the References section.

List of Figures and Tables in the Main Body

The list of the 98 figures and 4 tables in Parts I to III is presented online: ↗

List of Figures and Tables in the Materials

The list of the 4 figures and 77 tables in the Materials is presented online: ↗

References

Aalto, D., Malinen, J., & Vainio, M. (2018). Formants. In *Oxford Research Encyclopedia of Linguistics*. Published online: https://oxfordre.com/linguistics/display/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-419?rskey=bVx2FP.

American National Standards Institute (2013). *American National Standard on Acoustical Terminology*. ANSI/ASA S1.1-2013.

Andersen, H. S., Egsgaard, M. H., Ringsted, H. R., Grøntved, Å. M., Godballe, C., & Printz, T. (2021). Normative voice range profile of the young female voice. *Journal of Voice*, *14*(3), 546–552.

Andreeva, B., Möbius, B., Demenko, G., Zimmerer, F., & Jügler, J. (2015). Linguistic measures of pitch range in Slavic & Germanic languages. In *Proceedings of Interspeech 2015* (pp. 968–972).

Assmann, P. F. (1991). The perception of back vowels: Centre of gravity hypothesis. *The Quarterly Journal of Experimental Psychology*, *43*(3), 423–448.

Assmann, P. F., & Nearey, T. M. (2008). Identification of frequency-shifted vowels. *The Journal of the Acoustical Society of America*, *124*(5), 3203–3212.

Barreda, S., & Nearey, T. M. (2012). The direct and indirect roles of fundamental frequency in vowel perception. *The Journal of the Acoustical Society of America*, *131*(1), 466–477.

Birkholz, P., Gabriel, F., Kürbis, S., & Echternach, M. (2019). How the peak glottal area affects linear predictive coding-based formant estimates of vowels. *The Journal of the Acoustical Society of America*, *146*(1), 223–232.

Bladon, A. (1982). Arguments against formants in the auditory representation of speech. In R. Carlson & B. Granstrom (Eds.), *The representation of speech in the peripheral auditory system* (pp. 95–102). Elsevier.

Bladon, A. (1983). Two-formant models of vowel perception: Shortcomings and enhancement. *Speech Communication*, *2*(4), 305–313.

Bladon, A., & Fant, G. (1978). A two-formant model and the cardinal vowels. *Speech Transmission Laboratory Quarterly Progress and Status Report*, *19*(1), 1–8.

Bladon, A., Henton, C. G., & Pickering, J. B. (1984). Towards an auditory theory of speaker normalization. *Language and Communication, 4*(1), 59–69.

Boersma, P., & Weenink, D. (2020). *Praat: doing phonetics by computer* [Computer program]. Version 6.1.12. Retrieved April 20, 2020, from http://www.praat.org.

Bosker, H. R., Briaire, J. J., Heeren, W. F. L., van Heuven, V. J., & Jongman, S. (2010). Whispered speech as input for cochlear implants. *Linguistics in the Netherlands*, *27*, 1–15.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. MIT Press.

Brumm, H., & Zollinger, S. A. (2011). The evolution of the Lombard effect: 100 years of psychoacoustic research. *Behaviour*, *148*(11–13), 1173–1198.

Bunch, M., & Chapman, J. (2000). Taxonomy of singers used as subjects in scientific research. *Journal of Voice*, *14*(3), 363–369.

Carpenter, A., & Morton, J. (1962). The perception of vowel colour in formantless complex sounds. *Language and Speech*, *5*(4), 205–214.

Chistovich, L. A. (1985). Central auditory processing of peripheral vowel spectra. *The Journal of the Acoustical Society of America*, *77*(3), 789–805.

Chistovich, L. A., & Lublinskaya, V. V. (1979). The 'center of gravity' effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. *Hearing Research*, *1*(3), 185–195.

de Cheveigné, A. (2005). Pitch perception models. In C. J. Plack, R. R. Fay, A. J. Oxenham, & A. N. Popper (Eds.), *Pitch* (pp. 169–233). Springer.

de Cheveigné, A., & Kawahara, H. (1999). Missing-data model of vowel identification. *The Journal of the Acoustical Society of America*, *105*(6), 3497–3508.

Delattre, P., Liberman, A. M., Cooper, F. S., & Gerstman, L. J. (1952). An experimental study of the acoustic determinants of vowel color; observations on one- and two-formant vowels synthesized from spectrographic patterns. *Word*, *8*(3), 195–210.

de Saussure, F. (1916/1995). *Cours de linguistique générale* (édition critique préparée par T. de Mauro). Payot.

de Saussure, F. (1959). *Course in general linguistics* (W. Baskin, Trans.). Philosophical Library. (Original work published 1916)

d'Heureuse, C. (2014). *RMSnormalizer* [program module]. The source-code.biz JAVA DSP collection. Retrieved November 20, 2022, from https://www.source-code.biz/dsp/java/.

d'Heureuse, C. (2018). *SinSyn – Sinusoidal synthesizer* [browser-based web application]. GitHub repository. Retrieved November 20, 2022, from https://github.com/chdh/sin-syn.

d'Heureuse, C. (2019a). *KlattSyn – Klatt formant synthesizer* [browser-based web application]. GitHub repository. Retrieved November 20, 2022, from https://github.com/chdh/klatt-syn.

d'Heureuse, C. (2019b). *HarmSyn – Harmonic analyzer and synthesizer* [browser-based web application]. GitHub repository. Retrieved November 20, 2022, from https://github.com/chdh/harm-syn.

d'Heureuse, C. (2022). *SpecFilt – Spectral filter tool* [browser-based web application]. GitHub repository. Retrieved November 20, 2022, from https://github.com/chdh/spect-filt.

Diehl, R. L., Lindblom, B., Hoemeke, K. A., & Fahey, R. P. (1996). On explaining certain male-female differences in the phonetic realization of vowel categories. *Journal of Phonetics*, *24*(2), 187–208.

Donai, J. J., Motiian, S., & Doretto, G. (2016). Automated classification of vowel category and speaker type in the high-frequency spectrum. *Audiology Research*, *6*(1), 1–5.

Donai, J. J., & Paschall, D. D. (2015). Identification of high-pass filtered male, female, and child vowels: The use of high-frequency cues. *The Journal of the Acoustical Society of America*, *137*(4), 1971–1982.

Dubno, J. R., & Dorman, M. F. (1987). Effects of spectral flattening on vowel identification. *The Journal of the Acoustical Society of America*, *82*(5), 1503–1511.

Echternach, M., Birkholz, P., Traser, L., Flügge, T. V., Kamberger, R., Burk, F., Burdumy, M., & Richter, B. (2015). Articulation and vocal tract acoustics at soprano subject's high fundamental frequencies. *The Journal of the Acoustical Society of America*, *137*(5), 2586–2595.

Eklund, I., & Traunmüller, H. (1997). Comparative study of male and female whispered and phonated versions of the long vowels of Swedish. *Phonetica*, *54*(1), 1–21.

Fahey, R. P., & Diehl, R. L. (1996). The missing fundamental in vowel height perception. *Perception and Psychophysics*, *58*(5), 725–733.

Fant, G. (1959). Acoustic analysis and synthesis of speech with applications to Swedish. *Ericsson Technics*, *15*, 3–108.

Fant, G. (1970). *Acoustic theory of speech production* (2nd ed.). Walter de Gruyter.

Fastl, H. (1971). Über Tonhöhenempfindungen bei Rauschen. *Acustica*, *25*(6), 350–354.

Fastl, H., & Zwicker, E. (2006). *Psychoacoustics: facts and models* (Vol. 22). Springer Science and Business Media. (Kindle version)

Fernald, A., & Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development*, *10*(3), 279–293.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(3), 477–501.

Ferreira, A. J. (2007). Static features in real-time recognition of isolated vowels at high pitch. *The Journal of the Acoustical Society of America*, *122*(4), 2389–2404.

Fletcher, H. (1924). The physical criterion for determining the pitch of a musical tone. *Physical Review*, *23*(3), 427–437.

Fox, R. A., Jacewicz, E., & Chang, C. Y. (2011). Auditory spectral integration in the perception of static vowels. *Journal of Speech, Language, and Hearing Research*, *54*, 1667–1681.

Friedrichs, D., Maurer, D., & Dellwo, V. (2015a). The phonological function of vowels is maintained at fundamental frequencies up to 880 Hz. *The Journal of the Acoustical Society of America*, *138*(1), EL36–EL42.

Friedrichs, D., Maurer, D., Rosen, S., & Dellwo, V. (2017). Vowel recognition at fundamental frequencies up to 1 kHz reveals point vowels as acoustic landmarks. *The Journal of the Acoustical Society of America*, *142*(2), 1025–1033.

Friedrichs, D., Maurer, D., Suter, H., & Dellwo, V. (2015). Vowel identification at high fundamental frequencies in minimal pairs. In *Proceedings of the International Congress of Phonetic Sciences (ICPHs'15)* (s.n., pp. 1–4).

Fujisaki, H., & Kawashima, T. (1968). The roles of pitch and higher formants in the perception of vowels. *IEEE Transactions on Audio and Electroacoustics*, *16*(1), 73–77.

Gelfand, S. A. (2018). *Hearing: An introduction to psychological and physiological acoustics* (6th ed.). CRC Press.

Gramming, P., Sundberg, J., Ternström, S., Leanderson, R., & Perkins, W. H. (1988). Relationship between changes in voice pitch and loudness. *Journal of Voice*, *2*(2), 118–126.

Grieser, D. L., & Kuhl, P. K. (1988). Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. *Developmental Psychology*, *24*(1), 14–20.

Heeren, W. F. (2015). Vocalic correlates of pitch in whispered versus normal speech. *The Journal of the Acoustical Society of America*, *138*(6), 3800–3810.

Higashikawa, M., Nakai, K., Sakakura, A., & Takahashi, H. (1996). Perceived pitch of whispered vowels–relationship with formant frequencies: A preliminary study. *Journal of Voice*, *10*(2), 155–158.

Hillenbrand, J. M., Clark, M. J., & Baer, C. A. (2011). Perception of sinewave vowels. *The Journal of the Acoustical Society of America*, *129*(6), 3991–4000.

Hillenbrand, J., & Gayvert, R. T. (1993). Identification of steady-state vowels synthesized from the Peterson and Barney measurements. *The Journal of the Acoustical Society of America*, *94*(2), 668–674.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5), 3099–3111.

Hillenbrand, J., & Houde, R. A. (1996). Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *Journal of Speech, Language, and Hearing Research*, *39*(2), 311–321.

Hillenbrand, J. M., & Houde, R. A. (2002). Speech synthesis using damped sinusoids. *Journal of Speech, Language, and Hearing Research*, *45*(4), 639–650.

Hillenbrand, J. M., & Houde, R. A. (2003). A narrow band pattern-matching model of vowel perception. *The Journal of the Acoustical Society of America*, *113*(2), 1044–1055.

Hillenbrand, J. M., Houde, R. A., & Gayvert, R. T. (2006). Speech perception based on spectral peaks versus spectral shape. *The Journal of the Acoustical Society of America*, *119*(6), 4041–4054.

Hillenbrand, J. M., & Nearey, T. M. (1999). Identification of resynthesized /hVd/ utterances: Effects of formant contour. *The Journal of the Acoustical Society of America*, *105*(6), 3509–3523.

Hirahara, T., & Kato, H. (1992). The effect of F0 on vowel identification. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception, production and linguistic structure* (pp. 89–112). Ohmsha.

Honingh, A., & Bod, R. (2011). In search of universal properties of musical scales. *Journal of New Music Research*, *40*(1), 81–89.

Houle, N., & Levi, S. V. (2020). Acoustic differences between voiced and whispered speech in gender diverse speakers. *The Journal of the Acoustical Society of America*, *148*(6), 4002–4013.

Houtsma, A. J. M. (1995). Pitch perception. In B. C. J. Moore (Ed.), *Hearing: Handbook of perception and cognition* (2nd ed., pp. 267–295). Academic Press Inc.

Houtsma, A. J. M., & Smurzynski, J. (1990). Pitch identification and discrimination for complex tones with many harmonics. *The Journal of the Acoustical Society of America*, *87*(1), 304–310.

International Phonetic Association (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

International Phonetic Association (2005). The International Phonetic Alphabet (Revised to 2005).

Ito, M., Tsuchida, J., & Yano, M. (2001). On the effectiveness of whole spectral shape for vowel perception. *The Journal of the Acoustical Society of America*, *110*(2), 1141–1149.

Jackson, H. M., & Moore, B. C. (2013). The dominant region for the pitch of complex tones with low fundamental frequencies. *The Journal of the Acoustical Society of America*, *134*(2), 1193–1204.

Jenkins, R. A. (1961). Perception of pitch, timbre, and loudness. *The Journal of the Acoustical Society of America*, *33*(11), 1550–1557.

Johnson, K. (2008). Speaker normalization in speech perception. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 363–389). John Wiley and Sons.

Joliveau, E., Smith, J., & Wolfe, J. (2004a). Tuning of vocal tract resonance by sopranos. *Nature*, 427(6970), 116.

Joliveau, E., Smith, J., & Wolfe, J. (2004b). Vocal tract resonances in singing: The soprano voice. *The Journal of the Acoustical Society of America*, *116*(4), 2434–2439.

Kallail, K. J., & Emanuel, F. W. (1984a). An acoustic comparison of isolated whispered and phonated vowel samples produced by adult male subjects. *Journal of Phonetics*, *12*(2), 175–186.

Kallail, K. J., & Emanuel, F. W. (1984b). Formant-frequency differences between isolated whispered and phonated vowel samples produced by adult female subjects. *Journal of Speech, Language, and Hearing Research*, *27*(2), 245–251.

Kallail, K. J., & Emanuel, F. W. (1985). The identifiability of isolated whispered and phonated vowel samples. *Journal of Phonetics*, *13*(1), 11–17.

Kathiresan, T., Maurer, D., Suter, H., & Dellwo, V. (2018). Formant pattern and spectral shape ambiguity in vowel synthesis: The role of fundamental frequency and formant amplitude. *The Journal of the Acoustical Society of America*, *143*(3), 1919–1920. (Additional materials online: http://www.phones-andphonemes.org/asa/2018b.)

Katz, W. F., & Assmann, P. F. (2001). Identification of children's and adults' vowels: Intrinsic fundamental frequency, fundamental frequency dynamics, and presence of voicing. *Journal of Phonetics*, *29*(1), 23–51.

Keating, P. A., Garellek, M., & Kreiman, J. (2015). Acoustic properties of different kinds of creaky voice. In *Proceedings of the International Congress of Phonetic Sciences (ICPHs'15)* (Vol. 2015, No. 1, pp. 2-7).

Keating, P., & Kuo, G. (2012). Comparison of speaking fundamental frequency in English and Mandarin. *The Journal of the Acoustical Society of America*, *132*(2), 1050–1060.

Kent, R. D., Kent, R. A., & Read, C. (2002). *The acoustic analysis of speech* (2nd ed.). Singular.

Kent, R. D., & Vorperian, H. K. (2018). Static measurements of vowel formant frequencies & bandwidths: A review. *Journal of Communication Disorders*, *74*, 74–97.

Kiefte, M., Enright, T., & Marshall, L. (2010). The role of formant amplitude in the perception of /i/ and /u/. *The Journal of the Acoustical Society of America*, *127*(4), 2611–2621.

Kiefte, M., & Kluender, K. R. (2005). The relative importance of spectral tilt in monophthongs & diphthongs. *The Journal of the Acoustical Society of America*, *117*(3), 1395–1404.

Kiefte, M., & Kluender, K. R. (2008). Absorption of reliable spectral characteristics in auditory perception. *The Journal of the Acoustical Society of America*, *123*(1), 366–376.

Kiefte, M., Nearey, T. M., & Assmann, P. F. (2013). Vowel perception in normal speakers. In K. Pollok (Ed.), *Handbook of vowels and vowel disorders* (pp. 160–185). Psychology Press.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America*, *67*(3), 971–995.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, & perception of voice quality variations among female & male talkers. *The Journal of the Acoustical Society of America*, *87*(2), 820–857.

Koenig, L. L., & Fuchs, S. (2019). Vowel formants in normal and loud speech. *Journal of Speech, Language, and Hearing Research*, *62*(5), 1278–1295.

Konnai, R., Scherer, R. C., Peplinski, A., & Ryan, K. (2017). Whisper and phonation: Aerodynamic comparisons across adduction & loudness. *Journal of Voice*, *31*(6), 773.e11–773.e20.

Konno, H., Kudo, M., Imai, H., & Sugimoto, M. (2016). Whisper to normal speech conversion using pitch estimated from spectrum. *Speech Communication*, *83*, 10–20.

Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production & perception*. John Wiley & Sons.

Kursell, J. (2019). Listening to more than sounds: Carl Stumpf & the experimental recordings of the Berliner Phonogramm-Archiv. *Technology & Culture*, *60*(2), S39–S63.

Ladd, D. R., Turnbull, R., Browne, C., Caldwell-Harris, C., Ganushchak, L., Swoboda, K., Woodfield, V., & Dediu, D. (2013). Patterns of individual differences in the perception of missing-fundamental tones. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(5), 1386–1397.

Ladefoged, P. (1967). *Three areas of experimental phonetics*. Oxford University Press.

Ladefoged, P. (1996). *Elements of acoustic phonetics*. University of Chicago Press.

Ladefoged, P. (2003). *Phonetic data analysis: An introduction to fieldwork and instrumental techniques*. Wiley-Blackwell.

Ladefoged, P. (2005). *Vowels and consonants* (2nd ed.). Blackwell Publishing.

Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages.* Blackwell Publishing.

Lancker, D. V., & Sidtis, J. J. (1992). The identification of affective-prosodic stimuli by left-and right-hemisphere-damaged subjects: All errors are not created equal. *Journal of Speech, Language, and Hearing Research*, *35*(5), 963–970.

Laver, J. (1994). *Principles of phonetics*. Cambridge University Press.

Lee, Y., Oya, M., Kaburagi, T., Hidaka, S., & Nakagawa, T. (2021). Differences among mixed, chest, & falsetto registers: A multiparametric study. *Journal of Voice*, *37*(2), 298.e11–298.e29.

Lehiste, I., & Peterson, G. E. (1959). The identification of filtered vowels. *Phonetica*, *4*(4), 161–177.

Licklider, J. C. R. (1959). Three auditory theories. In S. E. Koch (Ed.), *Psychology: A study of a science: Study I. Conceptual and systematic. Vol. I. Sensory, perceptual, and physiological formulations* (pp. 41–144). McGraw-Hill.

Liénard, J. S., and Di Benedetto, M. G. (1999). Effect of vocal effort on spectral properties of vowels. *The Journal of the Acoustical Society of America*, *106*(1), 411–422.

Liu, C., & Eddins, D. A. (2008). Effects of spectral modulation filtering on vowel identification. *The Journal of the Acoustical Society of America*, *124*(3), 1704–1715.

Maurer, D. (2016). *Acoustics of the vowel: Preliminaries*. Peter Lang LTD International Academic Publishers.

Maurer, D. (2018). Why a phenomenology of vowel sounds is needed. In *Proceedings of the Conference on Phonetics and Phonology in German-Speaking Countries (P&P 13)* (pp. 121–124). (Extended version online: https://www.phones-and-phonemes.org/publications/PuP13-exteded-180212.pdf).

Maurer, D., Cook, N., Landis, T., & d'Heureuse, C. (1991). Are measured differences between the formants of men, women and children due to F0 differences? *Journal of the International Phonetic Association*, *21*(2), 66–79.

Maurer, D., Dellwo, V., Suter, H., & Kathiresan, T. (2017). Formant pattern ambiguity of vowel sounds revisited in synthesis: Changing perceptual vowel quality by only changing fundamental frequency. *The Journal of the Acoustical Society of America*, *141*(5), 3469–3470. (Poster presentation online: https://www.phones-and-phonemes.org/asa/Poster-ASA-2017a-200925.pdf).

Maurer, D., d'Heureuse, C., & Landis, T. (2000). Formant pattern ambiguity of vowel sounds. *International Journal of Neuroscience*, *100*(1–4), 39–76.

Maurer, D., d'Heureuse, C., & Leemann, H. (2020). Does vowel recognition relate to pitch? *The Journal of the Acoustical Society of America*, *148*(4), 2504.

Maurer, D., d'Heureuse, C., Suter, H., & Dellwo, V. (2019). Formant pattern and spectral shape ambiguity of vowel sounds, and related phenomena of vowel acoustics – Exemplary evidence. In *Proceedings of Interspeech 2019* (pp. 2368–2369).

Maurer, D., d'Heureuse, C., Suter, H., Dellwo, V., Friedrichs, D., & Kathiresan, T. (2018). The Zurich Corpus of Vowel and Voice Quality, Version 1.0. In *Proceedings of Interspeech 2018* (pp. 1417–1421).

Maurer, D., d'Heureuse, C., Suter, H., Dellwo, V., Friedrichs, D., & Kathiresan, T. (2024, July 31). *The Zurich Corpus of Vowel and Voice Quality, Version 2.0* [sound archive]. https://zhcorpus.org.

Maurer, D., Gröne, B., Landis, T., Hoch, G., & Schönle, P. W. (1993). Re-examination of the relation between the vocal tract and the vowel sound with electromagnetic articulography (EMA) in vocalizations. *Clinical Linguistics and Phonetics*, *7*(2), 129–143.

Maurer, D., & Landis, T. (1995). FO-dependence, number alteration, and non-systematic behaviour of the formants in German vowels. *International Journal of Neuroscience*, *83*(1–2), 25–44.

Maurer, D., & Landis, T. (1996). Intelligibility and spectral differences in high-pitched vowels. *Folia phoniatrica et logopaedica*, *48*(1), 1–10.

Maurer, D., Mok, P., Friedrichs, D., & Dellwo, V. (2014). Intelligibility of high-pitched vowel sounds in the singing and speaking of a female Cantonese opera singer. In *Proceedings of Interspeech 2014* (pp. 2132–2133). Additional materials online: https://is2014.phones-and-phonemes.org/en/.

Maurer, D., & Suter, H. (2017a). “Flat” vowel spectra revisited in vowel synthesis. *Journal of the Acoustical Society of America*, *141*(5), 3469. Poster presentation online: https://www.phones-and-phonemes.org/asa/Poster-ASA-2017b-170611.pdf. Additional materials online: https://www.phones-and-phonemes.org/asa/Poster-ASA-2017b-Add-170618.pdf).

Maurer, D., & Suter, H. (2017b). Vowel synthesis related to equal-amplitude harmonic series in frequency ranges > 1 kHz combined with single harmonics < 1 kHz, and including variation of fundamental frequency. *Journal of the Acoustical Society of America*, *141*(5), 3469. Materials online [poster presentation]: https://www.phones-and-phonemes.org/asa/Poster-ASA-2017c-170611.pdf. Retrieved November 20, 2022.

Maurer, D., Suter, H., Friedrichs, D., & Dellwo, V. (2015). Gender and age differences in vowel-related formant patterns: What happens if men, women, and children produce vowels on different and on similar F0? *Journal of the Acoustical Society of America*, *137*, 2416.

Maurer, D., Suter, H., Kathiresan, T., & Dellwo, V. (2018). Sinewave vowel sounds: The role of vowel qualities, frequencies and harmonicity of sinusoids, and perceived pitch for vowel recognition. *The Journal of the Acoustical Society of America*, *143*(3), 1920. Poster presentation and additional materials online: https://www.phones-and-phonemes.org/asa/2018a.

McAdams, S., & Giordano, B. L. (2009). The perception of musical timbre. In S. Hallam, I. Cross, & M. Thaut (Eds.), *The Oxford handbook of music psychology* (pp. 72–80). Oxford University Press.

McLoughlin, I. V., Li, J., & Song, Y. (2013). Reconstruction of continuous voiced speech from whispers. In *Proceedings of Interspeech 2013* (pp. 1022–1026).

McLoughlin, I. V., Sharifzadeh, H. R., Tan, S. L., Li, J., & Song, Y. (2015). Reconstruction of phonated speech from whispers using formant-derived plausible pitch modulation. *ACM Transactions on Accessible Computing (TACCESS)*, *6*(4), 1–21.

Melton, J., Bradford, Z., & Lee, J. (2020). Acoustic characteristics of vocal sounds used by professional actors performing classical material without microphones in outdoor theatre. *Journal of Voice*, 733.e23–733.e29.

Ménard, L., Schwartz, J. L., Boë, L. J., Kandel, S., & Vallée, N. (2002). Auditory normalization of French vowels synthesized by an articulatory model simulating growth from birth to adulthood. *The Journal of the Acoustical Society of America*, *111*(4), 1892–1905.

Meyer, J., Meunier, F., Dentel, L., Blanco, N. D. C., & Sèbe, F. (2018). Loud and shouted speech perception at variable distances in a forest. In *Proceedings of Interspeech 2018* (pp. 2285–2289).

Meyer-Eppler, W. (1957). Realization of prosodic features in whispered speech. *The Journal of the Acoustical Society of America*, *29*(1), 104–106.

Miller, R. L. (1953). Auditory tests with synthetic vowels. *The Journal of the Acoustical Society of America*, *25*(1), 114–121.

Monsen, R. B., & Engebretson, A. M. (1983). The accuracy of formant frequency measurements: A comparison of spectrographic analysis and linear prediction. *Journal of Speech, Language, and Hearing Research*, *26*(1), 89–97.

Morton, J., & Carpenter, A. (1962). Judgement of the vowel colour of natural and artificial sounds. *Language and Speech*, *5*(4), 190–204.

Nearey, T. M., & Kiefte, M. (2003). Comparison of several proposed perceptual representations of vowel spectra. In *Proceedings of the International Congress of Phonetic Sciences (ICPHs'15)* (Vol 1, pp. 1005–1008).

Niebuhr, O., Reetz, H., Barnes, J., & Yu, A. C. (2020). Fundamental aspects in the perception of f0. In C. Gussenhoven & A. Chen (Eds.), *The Oxford handbook of language prosody* (pp. 29–42). Oxford University Press.

Ohala, J. J. (1996). Speech perception is hearing sounds, not tongues. *The Journal of the Acoustical Society of America*, *99*(3), 1718–1725.

Pabon, P., & Ternström, S. (2020). Feature maps of the acoustic spectrum of the voice. *Journal of Voice*, *34*(1), 161-e1–161.e26.

Paeschke, A., & Sendlmeier, W. F. (2000). Prosodic characteristics of emotional speech: Measurements of fundamental frequency movements. In *Proceedings of the International Speech Communication (ISCA) Tutorial and Workshop (ITRW) on Speech Emotion* (pp. 75–80).

Pallier, C. (2002). *Computing discriminability and bias with the R software*. Retrieved November 20, 2022, from https://www.pallier.org/pdfs/aprime.pdf.

Pardo, J. S., & Remez, R. E. (2021). On the relation between speech perception and speech production. In J. S. Pardo, R. E. Remez, & D. B. Pisoni (Eds.), *The handbook of speech perception* (pp. 632–655). Wiley.

Pätzold, M., & Simpson, A. P. (1997). Acoustic analysis of German vowels in the Kiel Corpus of Read Speech. *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung Universität Kiel*, *32*, 215–247.

Paul, D. (1981). The spectral envelope estimation vocoder. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *29*(4), 786–794.

Peirce, C. S. (1958–1966). *Collected papers* (Vols. 1–6, C. Hartshorne & P. Weiss, Eds.; Vols. 7–8, A. W. Burks, Ed.). Belknap Press of Harvard University Press.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184.

Pickett, J. M. (1999). *The acoustics of speech communication: Fundamentals, speech perception theory, and technology*. Allyn and Bacon.

Plomp, R. (1967). Pitch of complex tones. *The Journal of the Acoustical Society of America*, *41*(6), 1526–1533.

Plomp, R. (2002). *The intelligent ear: On the nature of sound perception*. Erlbaum.

Potter, R. K., & Steinberg, J. C. (1950). Toward the specification of speech. *The Journal of the Acoustical Society of America*, *22*(6), 807–820.

Probst, L., & Braun, A. (2019). The effects of emotional state on fundamental frequency. In *Proceedings of the International Congress of Phonetic Sciences (ICPHs'19)* (pp. 67–71).

Quinn, S., Oates, J., & Dacakis, G. (2022). The effectiveness of gender affirming voice training for transfeminine clients: A comparison of traditional versus intensive delivery schedules. *Journal of Voice*, S0892-1997(22)00067-4. Advance online publication. https://doi.org/10.1016/j.jvoice.2022.03.001

Rabiner, L. (1968). Speech synthesis by rule: An acoustic domain approach. *Bell System Technical Journal*, *47*(1), 17–37.

Raitio, T., Suni, A., Pohjalainen, J., Airaksinen, M., Vainio, M., & Alku, P. (2013). Analysis and synthesis of shouted speech. In *Proceedings of Interspeech 2013* (pp. 1544–1548).

Reetz, H., & Jongman, A. (2020). *Phonetics: Transcription, production, acoustics, and perception* (2nd ed.). John Wiley and Sons.

Remez, R. E., & Rubin, P. E. (1984). On the perception of intonation from sinusoidal sentences. *Perception and Psychophysics*, *35*(5), 429–440.

Remez, R. E., & Rubin, P. E. (1993). On the intonation of sinusoidal sentences: Contour and pitch height. *The Journal of the Acoustical Society of America*, *94*(4), 1983–1988.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science*, *212*(4497), 947–949.

Rendall, D., Kollias, S., Ney, C., & Lloyd, P. (2005). Pitch (F0) and formant profiles of human vowels and vowel-like baboon grunts: The role of vocalizer body size and voice-acoustic allometry. *The Journal of the Acoustical Society of America*, *117*(2), 944–955.

Robinson, K., & Patterson, R. D. (1995). The stimulus duration required to identify vowels, their octave, and their pitch chroma. *The Journal of the Acoustical Society of America*, *98*(4), 1858–1865.

Rosen, S., & Hui, S. N. C. (2015). Sine-wave and noise-vocoded sine-wave speech in a tone language: Acoustic details matter. *The Journal of the Acoustical Society of America*, *138*(6), 3698–3702.

Sanchez, K., Oates, J., Dacakis, G., & Holmberg, E. B. (2014). Speech and voice range profiles of adults with untrained normal voices: Methodological implications. *Logopedics Phoniatrics Vocology*, *39*(2), 62–71.

Schouten, J. F. (1938). The perception of subjective tones. In *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen*, *41* (1086–1093).

Schouten, J. F. (1940). The residue, a new component in subjective sound analysis. In *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen*, *43* (356–365).

Schouten, J. F., Ritsma, R. J., & Cardozo, B. L. (1962). Pitch of the residue. *The Journal of the Acoustical Society of America*, *34*(9B), 1418–1424.

Seebeck, A. (1843). Ueber die Sirene. *Annalen der Physik*, *136*(12), 449–481.

Shackleton, T. M., & Carlyon, R. P. (1994). The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination. *The Journal of the Acoustical Society of America*, *95*(6), 3529–3540.

Sharifzadeh, H. R., McLoughlin, I. V., & Russell, M. J. (2012). A comprehensive vowel space for whispered speech. *Journal of Voice*, *26*(2), e49–e56.

Shriberg, E. E. (1992). Perceptual restoration of filtered vowels with added noise. *Language and Speech*, *35*(1–2), 127–136.

Siedenburg, K., & McAdams, S. (2017). Four distinctions for the auditory "wastebasket" of timbre. *Frontiers in Psychology*, *8*, 1747.

Signorello, R., Demolin, D., Bernardoni, N. H., Gerratt, B. R., Zhang, Z., & Kreiman, J. (2020). Vocal fundamental frequency and sound pressure level in charismatic speech: A cross-gender and -language study. *Journal of Voice*, *34*(5), 808.e1–808.e13.

Small Jr, A. M., & Daniloff, R. G. (1967). Pitch of noise bands. *The Journal of the Acoustical Society of America*, *41*(2), 506–512.

Smith, D. R., Patterson, R. D., Turner, R., Kawahara, H., & Irino, T. (2005). The processing and perception of size information in speech sounds. *The Journal of the Acoustical Society of America*, *117*(1), 305–318.

Smith, L. A., & Scott, B. L. (1980). Increasing the intelligibility of sung vowels. *The Journal of the Acoustical Society of America*, *67*(5), 1795–1797.

Smith, S. S., Chintanpalli, A., Heinz, M. G., & Sumner, C. J. (2018). Revisiting models of concurrent vowel identification: The critical case of no pitch differences. *Acta Acustica united with Acustica*, *104*(5), 922–925.

Smoorenburg, G. F. (1970). Pitch perception of two-frequency stimuli. *The Journal of the Acoustical Society of America*, *48*(4B), 924–942.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149.

Stumpf, C. (1926). *Die Sprachlaute. Experimentell-phonetische Untersuchungen nebst einem Anhang über Instrumentalklänge*. J. Springer.

Sundberg, J. (1987). *The science of the singing voice*. Northern Illinois University Press.

Sundberg, J. (2013). Perception of singing. In D. Deutsch (Ed.), *The psychology of music* (3rd ed., pp. 69–105). Elsevier.

Swanepoel, R., Oosthuizen, D. J., & Hanekom, J. J. (2012). The relative importance of spectral cues for vowel recognition in severe noise. *The Journal of the Acoustical Society of America*, *132*(4), 2652–2662.

Swerdlin, Y., Smith, J., & Wolfe, J. (2010). The effect of whisper and creak vocal mechanisms on vocal tract resonances. *The Journal of the Acoustical Society of America*, *127*(4), 2590–2598.

Syrdal, A. K. (1985). Aspects of a model of the auditory representation of American English vowels. *Speech Communication*, *4*(1–3), 121–135.

Tartter, V. C. (1989). What's in a whisper? *The Journal of the Acoustical Society of America*, *86*(5), 1678–1683.

Tartter, V. C., & Braun, D. (1994). Hearing smiles and frowns in normal and whisper registers. *The Journal of the Acoustical Society of America*, *96*(4), 2101–2107.

Ternström, S., Bohman, M., & Södersten, M. (2006). Loud speech over noise: Some spectral attributes, with gender differences. *The Journal of the Acoustical Society of America*, *119*(3), 1648–1665.

Titze, I. R. (2000). *Principles of voice production* (2nd ed.). National Center for Voice and Speech.

Titze, I. R. (2008). Nonlinear source–filter coupling in phonation: Theory. *The Journal of the Acoustical Society of America*, *123*(4), 1902–1915.

Titze, I. R., Baken, R. J., Bozeman, K. W., Granqvist, S., Henrich, N., Herbst, C. T., …, & Kreiman, J. (2015). Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization. *The Journal of the Acoustical Society of America*, *137*(5), 3005–3007.

Titze, I. R., & Palaparthi, A. (2016). Sensitivity of source–filter interaction to specific vocal tract shapes. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(12), 2507–2515.

Traunmüller, H. (1981). Perceptual dimension of openness in vowels. *The Journal of the Acoustical Society of America*, *69*(5), 1465–1475.

Traunmüller, H. (1988). Paralinguistic variation and invariance in the characteristic frequencies of vowels. *Phonetica*, *45*(1), 1–29.

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, *88*(1), 97–100.

Traunmüller, H., & Eriksson, A. (1995). The frequency range of the voice fundamental in the speech of male and female adults. *Unpublished manuscript*, *11*.

Traunmüller, H., & Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *The Journal of the Acoustical Society of America*, *107*(6), 3438–3451.

Various Artists (2015). *Burundi – Musiques traditionelles* [Album; CD]. Ocora Radio France.

von Helmholtz, H. (1863). *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. Vieweg.

Vos, R. R., Murphy, D. T., Howard, D. M., & Daffern, H. (2018). The perception of formant tuning in soprano voices. *Journal of Voice*, *32*(1), 126.e1.

Watrous, R. L. (1991). Current status of Peterson–Barney vowel formant data. *The Journal of the Acoustical Society of America*, *89*(5), 2459–2460.

Wightman, F. L. (1981). Pitch perception: An example of auditory pattern recognition. In D. J. Getty & J. H. Howard (Eds.), *Auditory and Visual Pattern Recognition* (pp. 3–25). Erlbaum Press.

Wolfe, J., Garnier, M., Bernardoni, N. H., & Smith, J. (2020). The mechanics and acoustics of the singing voice: Registers, resonances and the source–filter interaction. In F. A. Russo, B. Ilari, & A. J. Cohen (Eds.), *The Routledge companion to interdisciplinary studies in singing, volume I: Development* (pp. 64–78). Routledge.

Wolfe, J., Garnier, M., & Smith, J. (2009). Vocal tract resonances in speech, singing, and playing musical instruments. *HFSP Journal*, *3*(1), 6–23.

Yost, W. A. (2009). Pitch perception. *Attention, Perception, & Psychophysics*, *71*(8), 1701–1715.

Zhang, Y., Nolan, F., & Friedrichs, D. (2022). Perceptual clustering of high-pitched vowels in Chinese Yue Opera. *Speech Communication*, *137*, 60–69.