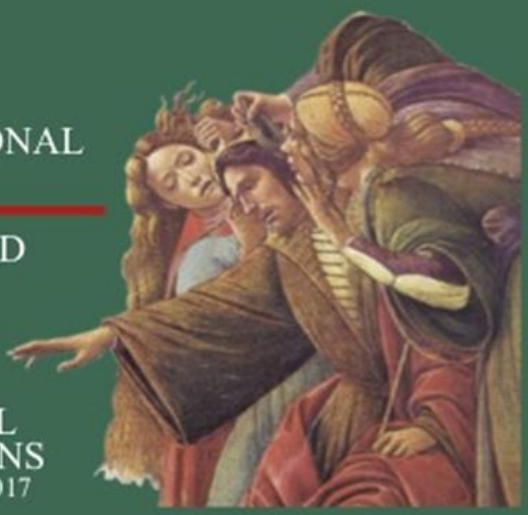
UNIVERSITÀ
DEGLI STUDI
FIRENZE
DINFO
DIPARTIMENTO DI
INGEGNERIA
DELL'INFORMAZIONE

10th
INTERNATIONAL
WORKSHOP

MODELS AND ANALYSIS OF VOCAL EMISSIONS FOR BIOMEDICAL APPLICATIONS

December 13-15, 2017
Firenze, Italy



PROCEEDINGS

FIRENZE
UNIVERSITY
PRESS

MODELS AND ANALYSIS OF VOCAL EMISSIONS FOR BIOMEDICAL APPLICATIONS

10th INTERNATIONAL WORKSHOP

December 13-15, 2017
Firenze, Italy

Edited by
Claudia Manfredi

Firenze University Press
2017

Models and Analysis of Vocal Emissions for Biomedical Applications : 10th International Workshop, december, 13-15, 2017 / edited by Claudia Manfredi. – Firenze : Firenze University Press, 2017.
(Proceedings and report ; 117)

http://digital.casalini.it/9788864536071

ISBN 978-88-6453-606-4 (print)
ISBN 978-88-6453-607-1 (online)

Cover: designed by CdC, Firenze, Italy.

Università degli Studi di Firenze
Firenze University Press
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com
Printed in Italy

MAVEBA 2017

Firenze, Italy

The MAVEBA 2017 Workshop is sponsored by:

Università degli Studi di Firenze

and is supported by:

Firenze University Press

Conservatorio Luigi Cherubini

Fondazione Ente Cassa di Risparmio di Firenze

CONTENTS

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

FOREWORD

A warm welcome to the participants of the tenth edition of the International Workshop MAVEBA!

The MAVEBA Workshop was born in 1999 and is proposed every two years as a multidisciplinary meeting for researchers working in the fields of bioengineering, medicine, psychology, linguistics, singing and related ones, with applications ranging from the infant to the elder.

The goal of MAVEBA is to bring together experts in the areas of human voice to share their knowledge and recent results with anyone who is interested in this multifaceted subject. As evidenced by the list of participants, the scientific community that meets in Firenze on this occasion comes from all over the world, confirming that the study of the human voice, our main means of communication, has no geographical boundaries.

Indeed, the study of the human voice has multiple facets, ranging from the pathologies and malformations of the phonatory apparatus, to the linguistic and phonetic aspects, also related to the emotional state, and to the ability and techniques of singing. MAVEBA is in fact not a purely technological and clinical meeting, as evidenced by the aims and topics dealt with in the various sessions, as artistic aspects are always a relevant part of it.

The 10th Workshop MAVEBA is organized into seven Sections, devoted to the following research subjects:

Session I: VOICE QUALITY ASSESSMENT
Session II: VOICE QUALITY MONITORING
Session III: VOICE AND NEUROCOGNITION
Session IV: VOICE AND SPEECH IN NEUROLOGY
Session V-VII: VOCAL FOLDS DYNAMICS

From the list above, Sessions concern topics that have been debated over the years, but over the years enriched by new scientific findings and technological innovations.

The first Session is about the objective measure of voice quality, that is the measure, assessment and classification of voice irregularities. This topic is closely related to a deep understanding of the dynamics of vocal folds.

Indeed, the last three Sessions deal with vocal folds dynamics, trying to give a rigorous explanation of one of the most complex and varied mechanisms of the human body, not yet fully exploited: our vocal folds are the source of the infinite range of sounds that make each of us unique and unrepeatable.

The subject of the link between voice, neurological disorders and emotional states is of great and increasing interest and is the topic of Sessions III and IV. The neurological problems of premature infants and language development at birth will be addressed, and innovative methods will be proposed for the study of relationships between voice and emotional or pathological states in adults and the elderly.

Another Session addresses the theme of voice quality monitoring, which is becoming increasingly popular thanks to developments in smartphone technology and software for the possibility of remote patient-clinician interaction. An additional Round Table: “Voice and speech processors: ready for global deployment in mobile devices?” is also dedicated to this topic, highlighting the industry's interest in the development of new technologies for the objective analysis of voice.

The subject of all Session is inherently multidisciplinary, promoting increasing collaboration between specialists from different disciplines.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

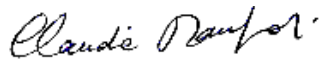
After nearly twenty years, again and as its peculiar feature, the congress venue is in the beautiful city of Firenze that, with its worldwide renowned historical Renaissance heritage, over the years becomes increasingly dynamic, welcoming and rich of events, not only artistic but also scientific and technological such as the MAVEBA Workshop.

The Opening Ceremony is offered in the Aula Magna of the Florentine University with a welcome address by the Rector Prof. Luigi Dei, who is also a keen connoisseur of music and singing. The participants will enjoy the beauty of the Aula Magna of the Rectorate located in the city center, commonly not accessible to the public. Artistic entertainment will be generously offered during the Congress by an actor and a flute player, and at the Luigi Cherubini Music Conservatory, showing the charm and versatility of voice and music. Finally, participants and accompanying persons will visit the Fortepiano Academy, a "hidden treasure" in the so called Diladdarno, perhaps a less known but full of charm district on the left bank of the Arno river. The Fortepiano Academy collects antique pianos and performs restoration in the annexed workshop.

Finally, I wish to thank the anonymous referees who devoted time and expertise in the review of the papers collected in this volume of the MAVEBA Proceedings. I am also very grateful to colleagues for their availability in chairing sessions and Round Tables. Special thanks to Dr. Philippe Dejonckere who coordinated the Round Tables with his usual precision and efficiency.

And, last but not least, I thank my co-workers Alice, Alessandro, Sara, Maria Sole, Gianandrea and my students, who generously devoted time and energy to the organization of this event. I hope that participants will find MAVEBA 2017 scientifically useful in the pleasant Florentine Christmas atmosphere.

Claudia Manfredi



MAVEBA Chair

SESSION I:

VOICE QUALITY ASSESSMENT

ON THE DESIGN OF A VOICE PATHOLOGY ASSESSMENT SYSTEM BASED ON THE GRB SCALE.

J.A. Gómez-García*, L. Moro-Velázquez, J. Mendes-Laureano, J.I. Godino-Llorente
Dpto. Señales, Sistemas y Radiocomunicaciones.
E.T.S. Ingenieros de Telecomunicación.
Universidad Politécnica de Madrid
jorge.gomez.garcia@upm.es; laureano.moro@upm.es; jmendes@ics.upm.es; igodino@ics.upm.es

Abstract: **The purpose of this paper is to present some preliminary results of an automatic system capable of modelling the perceptual abilities of a speech therapist that describes vocal quality in accordance to the GRB scale. The system is trained using three databases that have been evaluated by the same evaluator. 11 spectral, cepstral, modulation spectra and perturbation features are extracted from the input speech. Filter ranking algorithms are utilized to select the most consistent set of features among databases. Decision making is carried out using ordinal classification, to account for the ordering in the GRB scale, and Gaussian regression, to provide continuous decisions about voice quality. Results indicate that the proposed system is proficient when modelling, either by means of the Gaussian regressor and the ordinal classifier, the perceptual abilities of the evaluator. On average the deviations from the actual and predicted label are about half an unit.**

Keywords: **voice pathology assessment, GRBAS, regression, ordinal classification, voice pathology.**

I. INTRODUCTION

The clinical evaluation of voice often relies on an instrumental examination and a perceptual assessment of the speech. The instrumental examination focuses on a primary etiological diagnosis, whereas the perceptual assessment extracts multidimensional information that is not quantifiable instrumentally. Typically, the perceptual examination is performed in concordance to judgment rating scales that evaluate voice quality and provide information about the level of dysphonia present in voice. In this regard, the GRBAS is perhaps the most popular scale. This is composed of five traits ranging from 0 to 3, where 0 is referred to absence of pathology, 1 to light disease, 2 to moderate impairment and 3 to grave disorder. The descriptors define the hoarseness level (G), the roughness (R), breathiness (B), asthenia (A) and strain (S) present in voice. However, due to the unreliability of the A and S parameters, a simplified GRB scale is frequent[1]ly employed [1]. Despite the perceptual judgment scales have been designed to evaluate the most important aspects that are relevant to voice quality analysis, the reliability of the ratings are conditioned by the multidimensional nature of voice, the subjectivity of perception, the experience and background of the evaluators, the intrinsic variability of speech [2], the nonlinear relationship between pathology and measured or perceived voice quality [3], etc. In addition, the discrete nature of the ratings might affect the assessment task as some voices do not fit perfectly into certain categories (say 0 or 1) but in-between them (say 0,3).

Having this in mind, the present paper aims at designing a generalist *Automatic voice quality analysis* (AVQA) system capable of providing an assessment about vocal condition in terms of GRB descriptors. The aim is to model the perceptual capabilities of an evaluator through the analysis of speech material of different sources. The methodology is based on 11 characteristics describing spectral, cepstral, modulation spectra and perturbation aspects of normophonic and pathological voices. This initial set of features is extracted from three databases of sustained vowels which have been previously assessed by the same evaluator following the GRBAS scale. For generalization purposes, three filter ranking algorithms are employed to select the most consistent subset of characteristics capable of predicting G, R and B among the three databases. Decision making is carried out using two types of machines. The first relies on ordinal classification to address the ordinal character of the labels. The second accounts for the continuous nature of the assessment task through a Gaussian regression procedure. In this manner, and despite being trained using discrete GRB ratings, the system outputs a continuous value characterizing the degree of perceived pathology in voice.

II. METHODS

Databases: Three databases are used in this paper: *Hospital Principe de Asturias* (HUPA) [4], *Saarbrucken*

* orcid.org/0000-0002-6060-387X

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

Voice pathology (SVD) [5], and *hospital Gregorio Marañón* (GMar) [6].
GMar contains registers of 95 normophonic and 107 pathological Spanish speakers phonating the vowel /a/. The corpus has been recorded with a sampling frequency of 22050 Hz and 16 bits. SVD contains more than 2000 German speakers phonating different vowels and pronouncing a sentence. Registers are recorded at a sampling frequency of 50 kHz and 16 bits of resolution. Only a subset of 568 normophonic and 970 pathological subjects phonating the vowel /a/ are employed in this paper after having eliminated registers with a low dynamic range or interferences. Finally, HUPA encompasses recordings of the sustained phonation of the vowel /a/ of 366 adult Spanish speakers: 169 pathological and 197 normophonic. The corpus has been recorded with a sampling frequency of 50 kHz and 16 bits of resolution.
The three databases have been assessed by the same speech therapist following the lineages of the GRBAS scale, however the A and S traits are disregarded from further analysis.

Methodology: The methodology followed to design the automatic assessment system is presented in Fig. 1, whereas each one of the constituting blocks are discussed next.

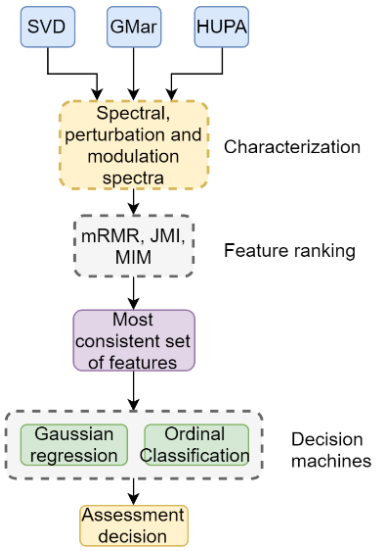


Figure 1: Methodology of the automatic assessment system proposed in this paper

Firstly and to allow comparison of results, the recordings of three databases have been resampled to 20 kHz and max-normalized as follows:

$$s_{norm}(t) = \frac{s(t)}{\max(s(t))}$$

where $s(t)$ is the input signal, and $\max(.)$ is the maximum value in the register. Then, short time analysis has been carried out by means of 40ms Hamming windows overlapped at 50% as in [7].
During the *characterization* stage, 11 features are extracted. These include 3 estimators of turbulent noise, 4 based on spectral/cepstral analysis, and 4 based on modulations spectra as described in [8]. The complete list of features is presented in Table 1.

Table 1. Features employed during the characterization stage.

Set	Features
Perturbation	Harmonics-to-noise ratio
	Normalized Noise Energy
	Glottal-to-Noise Excitation ratio
Spectral/Cepstral	Mel-frequency cepstral coefficients (12:2:20)
	Smoothed Cepstral Peak Prominence
	Low-to-High frequency spectral energy ratio
	Perceptual linear prediction coefficients
Modulation spectra	Modulation spectra homogeneity
	Cumulative Intersection Point
	Rate of Points above Linear Average
	Modulation Spectrum Percentile

After the characterization, a *feature ranking* procedure classifies features in accordance to their contribution in predicting G, B and R. To this end, three filter feature selection algorithms are utilized: *Mutual information maximization* (MIM), *max-relevance min-redundancy* (mRMR) and *joint mutual information* (JMI) [9].
Since the objective is to design a single system capable of generalizing results, a scoring procedure is employed to select the best global set of features among database and ranking algorithms. In this manner, for a certain feature selection technique and database, the scoring procedure rewards the best ranked features with a low score, and penalizes the worst with a large value. These scores are then summed up among the three databases and the three feature selection techniques. At the end, the features with the lowest scores are regarded as the most informative and consistent, and are employed for further testing. Having chosen this reduced set of features, it is now possible to train *decision machines*. It was found empirically that an average of the frames per speaker provides better results than training in a per-frame basis and hence this procedure is followed. Two scenarios are considered for decision making. First, a *Gaussian regressor* is used to predict G, R and B values in a continuous scale (ranging from 0 to 3). The number of Gaussians of the algorithm is varied in the range [2,4,8,16,32]. Then, an ordinal classification is performed through an algorithm called *Proportional Odd Model* (POM) to account for the ordinal nature of the GRB scale.

Finally, 484 registers are selected from the SVD dataset to *evaluate* results. It is worth noting that unlike the training procedure all the frames are employed in the detection task and no averaging is performed. However, a per-file decision is taken at the end.

For the Gaussian regressor two types of errors measure the deviation between the discrete perceptual evaluation given by the evaluator and the continuous decision given by the system: *mean absolute error* (MAE) and *root mean square error* (RMSE). These are calculated as follows:

$$RMSE = \sqrt{\frac{\sum_j (t_j - t_j^*)}{J}}$$

$$MAE = \frac{|t_j - t_j^*|}{J}$$

Where t_j is the actual label, $t_j^*$is the predicted label and J is the total number of trials.

For the ordinal regression, *ordinal mean absolute error* (oMAE) and *ordinal average mean absolute error* (oAMAE) are employed. The latter is used since oAMAE accounts better for errors in the ordering than MAE [10]. These measures are as follow:

$$oMAE = \frac{|\Phi(t_j) - \Phi(t_j^*)|}{J}$$

$$oAMAE = \frac{\sum_k oMAE_k}{K}$$

Where $oMAE_k$ is $oMAE$ calculated for instances of class k, and $\Phi(.)$ is an operator indicating the position of the label in the ordinal rank, i.e., if a certain label can take up values 0, 1, 2, 3 and the predicted label is 2, then position is 3.

III. RESULTS

Firstly, the Pearson's and Kendall's correlation indexes are used to gauge the relationship between G-B, G-R and B-R. The idea is to compare to which degree the different traits are related to each other. These results are presented in Table 2.

Table 2: Correlation between G-B, G-R and B-R

	G-B		G-R		B-R	
	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson
HUPA	0,72	0,73	0,78	0,79	0,68	0,70
SVD	0,73	0,79	0,78	0,82	0,54	0,60
Gmar	0,63	0,67	0,82	0,84	0,50	0,49

As indicated by the feature ranking algorithms, the most consistent results are obtained with just three features: *Cepstral harmonic-to-noise ratio* (CHNR), *Modulation spectra ratio above linear average* (RALA) and *glottal-to-noise excitation ratio* (GNE). These features performed equally well for the three considered traits: G, B and R.

The results of the ordinal classifier trained with this three features are introduced in Table 3. They introduce the error deviation (oAMAE and oMAE) between the actual and the predicted label.

Table 3: Ordinal classifier: oAMAE and oMAE of G, B, and R, calculated for the SVD evaluation partition.

G		B		R	
oMAE	oAMAE	oMAE	oAMAE	oMAE	oAMAE
0,5	0,48	0,54	0,56	0,63	0,64

The results of the Gaussian regression, evaluated using a partition based on the SVD database are presented in Table 4. They introduce the error measures (RMSE and MAE) between the discrete value given by the GRB evaluation and the continuous value predicted by the proposed system.

Table 4: Gaussian regressor: RMSE and MAE of G, B, and R, calculated for the SVD evaluation partition.

# of gaussians	G		B		R	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
2	0,71	0,50	0,74	0,54	0,83	0,62
4	0,69	0,50	0,74	0,54	0,79	0,60
8	0,69	0,50	0,73	0,54	0,78	0,60
16	0,70	0,48	0,73	0,55	0,79	0,60
32	0,70	0,48	0,74	0,54	0,80	0,60

IV. DISCUSSION

As observed from Table 2, there exist a large correlation between all the traits. The correlation between G-B and G-R is expected, as G is a measure of hoarseness, which is typically considered a superclass encompassing both B and R components. However, there is an large correlation between B-R that remains unidentified and that in some cases reaches 0,7. One hypothesis explaining this, might be in the presence of pathologies that affect both B and R similarly. However, this phenomenon deserves a deeper study to gain insight on the reasons behind this behavior. This large correlation might be the reason for which the filter ranking algorithms selected the same set of features no matter if G, B or R were used.

Regarding the ordinal classification, oAMAE is a measure that penalizes errors in the ordering, in such a manner that the farther the predicted label from its target, the larger oAMAE. The error rates given by oAMAE (ranges from 0,48 for the G trait to 0,64 for the R trait) are in the same order of magnitude than those of oMAE (ranges from 0,5 for the G trait to 0,63 for the R

trait), suggesting that in general errors should be located in the neighbor labels.
Regarding the Gaussian regression, it can be noticed that in general errors are of the same order of magnitude as those in the ordinal classification procedure. MAE ranges from 0,48 to 0,60, indicating that on average predicated labels deviate about half an unit from the perceptual evaluation provided by the speech therapist. In a similar way to oAMAE, RMSE penalizes large deviations. In this case, RMSE values are larger than MAE, indicating a certain level of discrepancy in between some of the estimated and the actual labels.
Finally, it is worth to mention, that having used recordings belonging to the same corpus for training and evaluation of results (although after being averaged, concatenated with data of other datasets and post-processed) might inject information into the system about the experimental conditions followed during the recording of the SVD database. The current outcomes thus this outcomes might be regarded as experiments under "favorable conditions".

V. Conclusion

This paper has presented some preliminary results of an automatic assessment system aimed at predicting the GBR scale, considering the ordinal nature of the scale, and the continuous nature of the assessment task. This has been achieved after having modelled the perceptual capabilities of an evaluator in predicting the G, B and R traits. For the sake of generalization, experiments are performed using several types of features, which are extracted from three databases. A feature ranking procedure serves to define the most consistent subset of characteristics -among the datasets and three feature selection algorithms- whereas a Gaussian regressor and an ordinal are employed to provide assessments of vocal quality in terms of the GRB scale. Results indicate that the features providing the most consistent behavior when considering the above-mentioned setup are RALA, GNE and CHNR. Outcomes also indicate that is possible to design a generalist system capable of successfully predicting G, B and R. Moreover, it is also possible to translate the information provided by the discrete GRB scale into a continuous space as observed by the reasonable error values of Table 4.
This work constitutes a preliminary work which models the capabilities of a single evaluator, and as such, it suffers from the subjective factors to which the evaluator is conditioned. Notwithstanding it opens the door to other type of analysis with multiple evaluations and which might generalize even further the results of this type of assessment.
As future work the meaningfulness of the continuous values outputted by the Gaussian regressor are assessed clinically. New tests are also to be performed using other types of characteristics. Likewise, other datasets are to be employed for the sole purpose of testing, and measuring out-of-sample performance.

ACKNOWLEDGMENTS

This research was carried out under grants: *"Ayudas para la realización del doctorado"* (RR01/2011), PRX15/00385, XV *Ayudas Consejo Social-UPM* and *Ayudas EEBB para PDI-UPM*, from Universidad Politécnica de Madrid; and TEC2012-38630-C04-01 from the Spanish Ministry of Education; with special thanks to the Fulbright Foundation.

REFERENCES

[1] Cornelia Moers et al. "Vowel- and Text-Based Cepstral Analysis of Chronic Hoarsenes*". Journal of Voice* , Volume 26 , Issue 4 , 416 - 424
[2] R. Fraile, J. I. Godino-Llorente, N. Sáenz-Lechón, et al., "Characterization of dysphonic voices by means of a filterbank-based spectral analysis: sustained vowels and running speech", *Journal of Voice*, vol. 27, no. 1, pp. 11–23, 2013.
[3] M. Putzer and W. J. Barry, "Instrumental dimensioning of normal and pathological phonation using acoustic measurements", *Clinical linguistics & phonetics*, vol. 22, no. 6, pp. 407–20, 2008.
[4] J. I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, et al., "Acoustic analysis of voice using wpcvox: a comparative study with multi-dimensional voice program" *European Archives of Oto-Rhino-Laryngology*, vol. 265, no. 4, pp. 465–476, Apr. 2008.
[5] Saarbrucken voice database. Available online: http://www.stimmdatenbank.coli.uni-saarland.de/index.php4
[6] L. Moro-Velázquez, J.A Gómez-García, J.I Godino-Llorente,, G. Andrade-Miranda, "Modulation Spectra Morphological Parameters: A New Method to Assess Voice Pathologies according to the GRBAS Scale," *BioMed Research International*, vol. 2015, Article ID 259239, 13 pages, 2015.
[7] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz and G. Castellanos-Domínguez, "Automatic Detection of Pathological Voices Using Complexity Measures, Noise Parameters, and Mel-Cepstral Coefficients," in *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 370-379, Feb. 2011.
[8] L. Moro-Velázquez, J.A Gómez-García & J.I Godino-Llorente. (2016). "Voice Pathology Detection Using Modulation Spectrum Optimized Metrics" *Frontiers in Bioengineering and Biotechnology*, 4,1.
[9] G. Brown, A. Pocock, M.J Zhao & M. Luján "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection". *Journal of machine learning research*, 13, pp. 27-66, Jan 2012.
[10] S. Baccianella, A. Esuli, and F. Sebastiani. "Evaluation Measures for Ordinal Regression". In *Proceedings of the 2009 Ninth International Conference on Intelligent Systems Design and Applications* (ISDA '09). IEEE Computer Society, Washington, DC, USA, 283-287.

EFFECT OF VELOPHARYNGEAL INSUFFICIENCY ON HUMAN VOICE QUALITY

T. Vampola[1], J. Horáček [2]
[1] Dept. of Mechanics, Biomechanics and Mechatronics, CTU in Prague ,Czech Republic
[2] Institute of Thermomechanics, Academy of Sciences of the Czech Republic
Tomas.Vampola@fs.cvut.cz, JaromirH@it.cas.cz

Abstract: **First a complex accurate parametric three-dimensional (3D) finite element (FE) model of the human acoustic cavities of the vocal tract for nasalized vowel [a:] was developed from which a simplified lumped model was created using a special reduction procedure. The simplified lumped model allows numerical simulation of the effects of nasality on the acoustic resonance and antiresonance characteristics of the vocal tract. The model is computationally effective and enables changes of the acoustic cavities continuously within the physiological limits. Usage of the sophisticated 3D FE model of the vocal tract for investigating influence of vocal tract shape modifications on the changes of acoustic resonance properties is time consuming. The accuracy of the results obtained by the reduced model is examined by comparing the results with the full 3D FE model.**
Keywords: **human vocal tract, nasal cavities, bio-acoustics, FE parametric model,**

I. INTRODUCTION

Human voice is produced through self-oscillations of the vocal folds excited by air flowing from the lungs. The vibration of the vocal folds modulates the stream of air producing a primary sound signal in glottis. This signal, which propagates through the supralaryngeal cavities up to the lips and nostrils, is modified by the acoustic resonances of the vocal tract. While the influence of the geometric configuration of the main channel of the vocal tract on the vocal output has been studied rather extensively, the influence of side cavities of human vocal tract, has received less attention. As such, their role for the resulting vocal intensity may be considered negligible or even undesirable, since it contradicts the general goal of enhancing vocal output with the smallest vocal effort. However, the newest studies revealed that besides the undesirable antiresonances there are also new resonances which occur due to the side cavities and that the voice quality can be better when the side branches are present [1], [2]. These specific resonances can contribute to the region of the so-called singer's or actor's/speaker's formant cluster created in the frequency range 2.5 - 4 kHz [3], [4]. Furthermore, the spectral analysis of singers indicates that due to the existence of the side branches the formant structure around 3-5 kHz is more complex than expected.
The effects of nasality or so-called velopharyngeal insufficiency is modeled in the present paper by interconnecting the acoustic cavities of the nasal tract with the vocal tract at the velum (soft palate).

II. METHODS

Sophisticated and accurate 3D FE models of the vocal tract for the vowel [a:] was created from the Computer Tomography (CT) measurement of the subject during phonation, see [2].

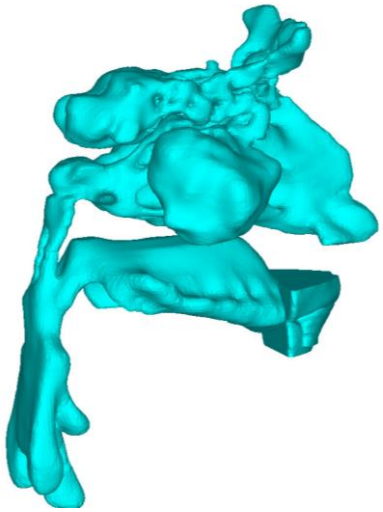


Fig. 1 Volume model of the human vocal tract for vowel [a:] interconnected with the model of the nasal cavities.

The accurate and complete 3D FE model of the acoustic nasal cavities was developed from a detail CT investigation of a patient head of another subject of the same gender, the similar age and size. After segmentation of the CT images we obtained the volume model of the nasal tract which was interconnected with the volume model of the vocal tract, see Fig.1. In the acoustic analyses two types of

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

boundary conditions were modelled (with and without the lips) see Fig. 2.

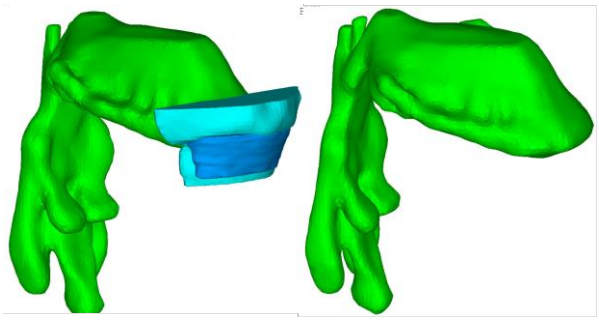


Fig. 2 Volume model of the human vocal tract for vowel [a:] width (left) and without (right) the lips.

The acoustic resonancies of the FE models were excited by a broadband frequency airflow pulse. The pulse excited the model at the glottis level and the acoustic pressure responses were computed at the position of the lips and nose. The pulse has a flat spectrum in the frequency range up to about 10 kHz, see Fig.3

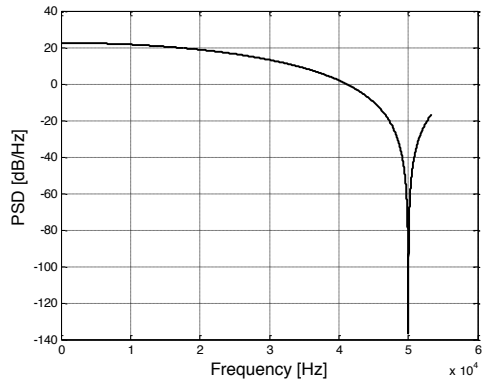


Fig. 3. Exciting pulse in the frequency domain.

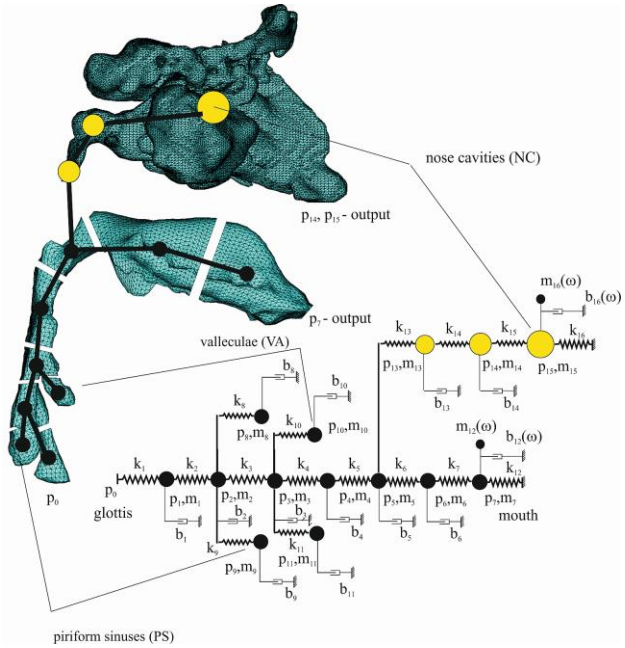
Acoustic energy losses by the sound radiation from the mouth and nose to open atmosphere, belong to main acoustic energy dissipation losses in the vocal tract. The radiation losses were modeled by a circular plate vibrating like a piston in an infinite wall, for which the following frequency dependent acoustic impedance can be derived, see e.g. [5]:

$$Z_a = \frac{c_0 \rho_0}{S}\left(1 - \frac{J_1(2kR)}{kR} + i\frac{H_1(2kR)}{kR}\right), \qquad (1)$$

where c_0 is sound velocity, ρ_0 is air density, R is equivalent radius of a vibrating plate calculated from the cross-section of the vocal tract model at the lips level and at the nostrils, $k = \omega/c_0$ is wave number, ω is angular frequency and i is imaginary unit. The Bessel J_1 and Struve H_1 functions can be calculated using the infinitive series. The acoustic energy losses inside the vocal tract due to, e.g., air viscosity and a material damping of the soft tissues on the boundaries of the acoustic spaces, were incorporated in the model via the boundary admittance coefficient $\mu = r/\rho_0 c_0$, where r is the real component of the specific acoustic impedance (resistance term).

Because of the usage of the complete 3D FE model of the nasalized vowel [a:] for investigating the effects of nasality considering a continuous vocal tract shape modification, and their influence on the changes of acoustic resonance properties of the system is computationally very time consuming, the 3D FE model was reduced to a simplified lumped model, see Fig. 4.

Fig.4 FE model and the simplified lumped model of the human vocal tract for a nasalized vowel **[a:].**



The lumped model of the nasalized vowel [a:] was created including all the dominant parallel cavities (two piriform sinuses, two valleculae and the nasal cavities joint to the main vocal tract at the velum) and their resonance and antiresonance frequencies were tuned in order to correspond to those of the full 3D FE model. The lumped model was developed by using similar reduction procedure as in the paper [2], where only piriform sinuses and valleculae were considered as the side branches of the vocal tract.

III. RESULTS

The influence of the boundary conditions was studied by the modal analysis of the full 3D FE models. Figure 5 shows the character of vibrations for the first eigenfrequencies of the models with and

without the lips. The first acoustic mode shapes are very similar.

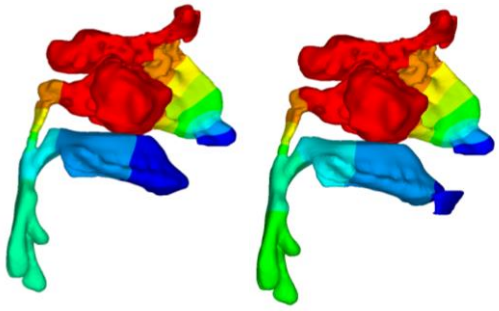


Fig.5 First eigenmodes for the human vocal tract models for a nasalized vowel [a:] with and without the lips.

The sensitivity of FE model on the boundary conditions is presented in Fig.6. The decrease of eigenfrequencies caused by consideration of the lips is the most important for the first two formants where the decrease is of about 7-11 %.

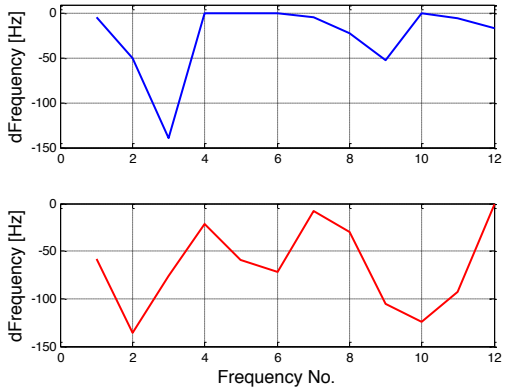


Fig.6 Sensitivity of the human vocal tract model for vowel [a:] on the boundary conditions. Model width nasal tract (top) and without nasal tract (bottom).

Comparison of the acoustic pressure response of the full 3D FE model with the reduced model is shown in Fig. 7. The first seven resonance frequencies and the two antiresonance frequencies up to 4 kHz of the lumped model are in a very good agreement with those obtained from the full 3D FE model. At the frequencies above 4 kHz the two models show different resonances; there are many more resonances in the full 3D model. This can be attributed to the limitation of the reduced model which does not capture the more complicated transversal modes in the higher frequency region.

Influence of connecting the human vocal tract with the nasal tract is demonstrated in Fig. 8. In addition to the three ordinary formants F1-F3 below 3 kHz, there are also two oro-nasal formants, first bellow F1 and second one bellow F3.

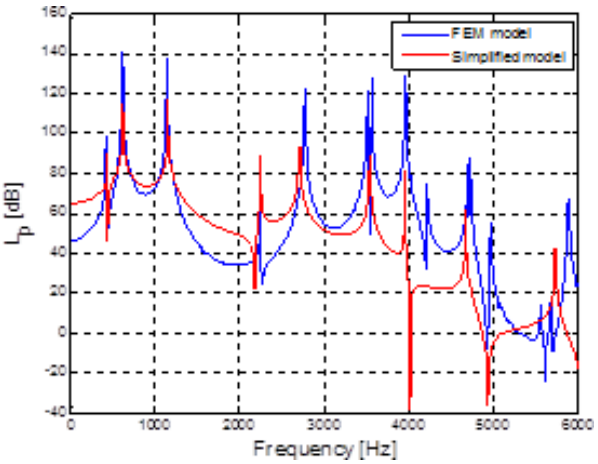


Fig.7 Acoustic pressure response computed at the lips using the full 3D FE model of the vocal tract and the simplified lumped model.

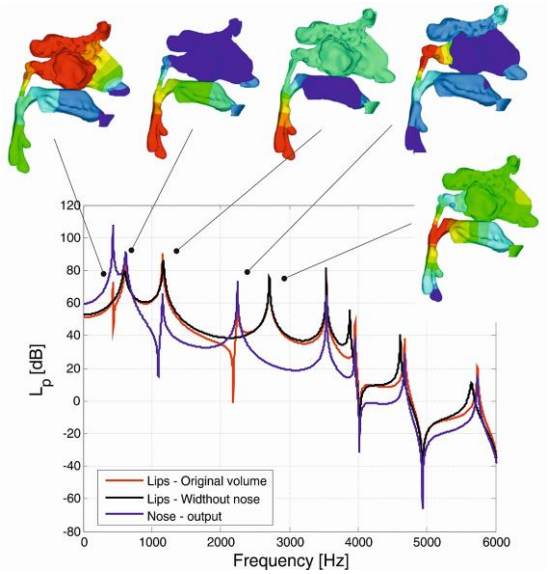


Fig.8 Acoustic pressure response computed at the lips and the nose using the lumped model for the ordinary and nasalized vowel [a:].

Influence of the size of velum region on the output acoustic pressure p_7 is demonstrated by an example shown in Fig. 9. When decreasing of the cavity in the velum region, the first two formats are moving together. The antiresonance-resonance pair between the second and third formant (≈ 2250 Hz) moves to the third formant and increased the energy in the acoustic signal in the frequency range 2.5 – 3 kHz.

The developed lumped model enables to study such changes of the output acoustic spectra very fast and systematically and to find an optimum for the voice

quality. Results for an optimal size of the velo-nasal interconnection are presented in Fig.10.

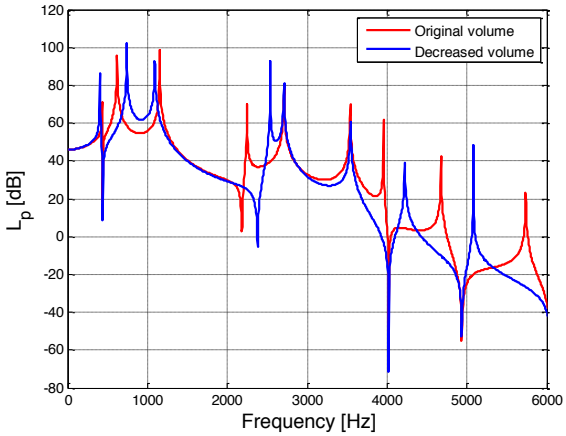


Fig.9 Acoustic pressure response computed at the lips using the lumped model for original and decreased size of the velum region for a nasalized vowel [a:]

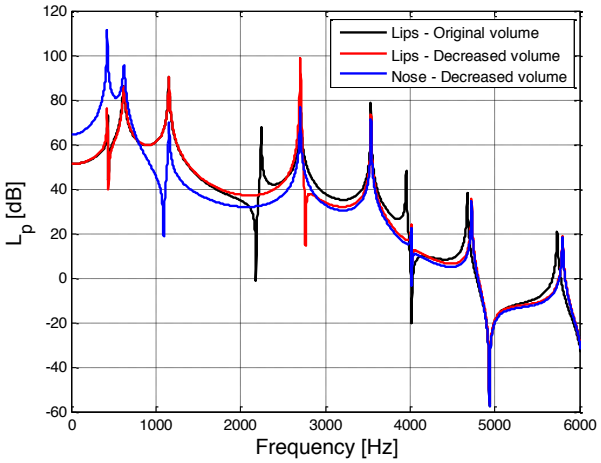


Fig.10 Acoustic pressure response computed at the lips using the reduced model for original and optimal velum region size for the nasalized vowel [a:].

IV. DISCUSSION

For the numerical prediction of the voice quality of the model without the nasal tract the correct modeling of the boundary condition is necessary, see Fig. 6. The differences of the eigenfrequencies for the model with and without lips are up to ca 140 Hz. For a complete model with a nasal tract, the boundary conditions play less important role. Only when acoustic vibrations in the nasal tract are not excited the effect of lips modelling is important.

As a result of tuning of the simplified model to the first 4 resonant peaks of the full 3D FE model, the acoustic pressure response agreement is very good and justifies the use of the simplified model to examine the effects of the parallel acoustic cavities on the voice quality, see Fig.7.

Figure 10 demonstrates that by changing appropriately the interconnection (the cross section) between the vocal and nasal tract the acoustic energy in the frequency range 2.5-3 kHz can be increased.

V. CONCLUSION

The results show that the human vocal tract is a very complex resonator. Side branches are generally known to cause antiresonances, i.e., sharp local minima in the resulting transfer function. In speech research the antiresonance phenomenon is well known from the studies of nasalized vowels where the nose acts as the side branch of the vocal tract. The side cavities act as antiresonators which severely decrease the sound level radiating out of the mouth around the antiresonance frequency. Simultaneously, however, they act also as resonators which amplify the acoustic output at specific frequencies that can be controlled by volume changes of the side cavities. These findings suggest that the side cavities may play a beneficial role in producing the "resonant voice" and formant clustering around 3-4 kHz that plays an important role in the professional speaker's or singer's voice quality [3,7].

Acknowledgement

The research is supported by the Grant Agency of the Czech Republic by project No 16 01246S. *Computational and experimental modelling of self-induced vibrations of vocal folds and influence of their impairments on human voice*.

REFERENCES

[1] B. Delvaux, D. M. Howard, "A new method to explore the spectral impact of the piriform fossae on the singing voice: Benchmarking using MRI-based 3D-printed vocal tracts," *PLoS ONE,* vol.7, no. 9, 2014.

[2] T. Vampola, J. Horáček, J. G. Švec, "Modeling the influence of piriform sinuses and valleculae on the vocal tract resonances and antiresonances," *Acta Acustica United With Acustica,* vol. 101, pp. 594-602, 2015.

[3] T. Leino, "Long-term average spectrum study on speaking voice quality in male actors," in SMAC93, Proceedings of the Stockholm Music Acoustics Conference, eds. A. Friberg, J. Iwarsson, E. Jansson, and J. Sundberg (The Royal Swedish Academy of Music, Stockholm), pp. 206–210, 1993.

[4] J. Sundberg, "Articulatory interpretation of the singing formant," *J. Acoust. Soc. Am.,* vol. 55, pp. 838–844, 1974.

[5] T. Vampola, J. Horáček, J.G. Švec, "FE modeling of human vocal tract acoustics. Part I: Production of Czech vowels," *Acustica United with Acta Acustica*, vol. 94, pp. 433-447, 2008.

DESCRIBING VOICE PERIOD VARIABILITY BY MEANS OF TIME SERIES STRUCTURAL ANALYSIS

G. A. Alzamendi[1, 2], G. Schlotthauer[1, 2]

[1] Laboratorio de Señales y Dinámicas no Lineales, Facultad de Ingeniería, Oro Verde, Argentina
[2] Instituto de Investigación y Desarrollo en Bioingeniería y Bioinformática, CONICET-UNER, Argentina,
{galzamendi, gschlotthauer}@ingenieria.uner.edu.ar

Abstract: **Variability in voice fundamental period is considered as an important indicator describing voice phonation and vocal condition. Acoustical parameters based on perturbation analysis are commonly applied for assessing period variability. Even though perturbation analysis is widely accepted, it is not able to describe in detail the various normally observed phenomena. In this work, period variability is described by means of state space structural analysis, which allows for optimal estimating and computing three components in the period sequence, namely trend, cycle and perturbation. Structural analysis is applied for decomposing period sequences obtained from type 1 sustained vowels, corresponding to both healthy and pathological subjects. Then, the estimated components are independently processed in order to reveal the most relevant information. It is shown that structural analysis suitable describes period variability, where the most important aspects are modeled in the structural components: trend, cycle and perturbation. Results suggest that structural analysis performs well on healthy and pathological cases, and that period variability is explained by cycle and perturbation components.**

Keywords: **Period variability, structural analysis, period fluctuation, jitter, period sequence.**

I. INTRODUCTION

Voice fundamental period, or its reciprocal the voice fundamental frequency, has a strong influence in voice perception, impinging on attributes such as naturalness, intonation, emphasis and vocal quality. Studying and modeling the dynamic of fundamental period, with particular focus on process variability, has proved to be useful in researching normal phonation, voice disorders, speech processing and natural voice synthesis, among others. In clinical practice in particular, it has become a key aspect in measuring the severity of dysphonia and the efficacy of a treatment [1], [2].

Analysis of period variability involves processing a period sequence (PS), i.e., a time series of successive fundamental periods extracted from a voice signal, in order to discover the most representative features of this phenomenon. Different acoustic parameters based on perturbation analysis have been proposed for variability assessment. Even though these acoustical parameters are widely used, they suffer for different technical and structural issues [2]–[4]. In particular, classical methods are not able to describe in detail the slow long-term fluctuations, the cyclic vocal microtremors and the local short-term perturbations (also called Jitter) that are present in the PS [5], [6]. For this reason, several strategies have been developed in the past to explain period variability considering some or all of those components [7]–[11].

Recently, the authors proposed a state space approach for the structural analysis of PS [12], [13]. Briefly, this method allows describing the behavior of a PS by decomposing it into components with simple and straightforward interpretations in terms of the phenomena previously detailed. The present work aims to describe period variability by using the structural analysis, and to compare this method with the classical perturbation analysis. For a thorough review of structural analysis using state space methods, see [14], [15].

This article is organized as follows. In Sec. II the materials used in this work are described, structural analysis is introduced and state space methods are briefly revised. In Sec. III the experimental results are reported and discussed. In Sec. IV the conclusions are presented.

II. MATERIALS AND METHODS

A. Period sequences computation

The PS were obtained by processing voice signals from the Disordered Voice Database, developed by the Massachusetts Eye and Ear Infirmary (MEEI) Voice and Speech Lab [16]. In this work, type 1 (nearly-periodic) sustained vowels /a/ were considered corresponding to 53 subjects (21 males, 32 females) with healthy voices and 74 subjects (29 males, 45 females) diagnosed with different voice disorders.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

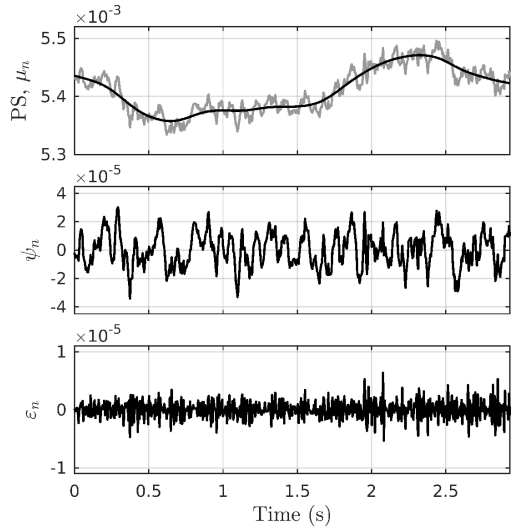


Figure 1: Structural analysis of a PS of a healthy male subject. *Top:* PS in gray, trend μ_n in black. *Center:* cycle component ψ_n. *Bottom:* perturbation ε_n.

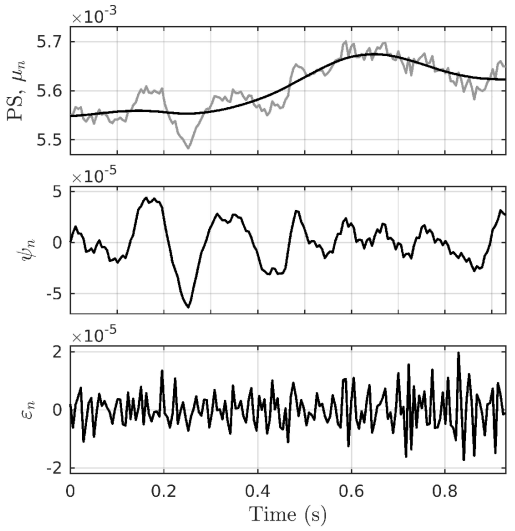


Figure 2: Structural analysis of a PS of a pathological female subject. *Top:* PS in gray, trend μ_n in black. *Center:* cycle component ψ_n. *Bottom:* perturbation ε_n.

Information regarding the recording and digitization conditions is thoroughly described in [17], [18]. The durations were 3 s and 1 s for healthy and disordered voices, respectively.

In order to extract the PS, voice recordings were processed using Praat software (available online at http://www.praat.org/). First, a waveform-matching short-term analysis method was applied to estimate the individual vocal cycles. Next, period length P_n of successive cycles were computed, and the PS was defined as $\{P_1, P_2, \dots, P_N\}$, where N is the number of vocal cycles. Finally, the PS were resampled at a constant sampling frequency equal to the mean fundamental frequency. Examples of PS corresponding to a healthy and a pathological subject are shown at the top of Fig. 1 and Fig. 2, respectively. At first sight, no structural difference can be appreciated between healthy and pathological examples.

B. Structural analysis

Structural analysis considering trend μ_n, cycle ψ_n and perturbation ε_n components was applied. As in [12], [13], it was assumed that μ_n, ψ_n and ε_n represent period fluctuations, vocal microtremors and jitter, respectively. According to this, PS was modeled as follows:

$$P_n = \mu_n + \psi_n + \varepsilon_n. \quad (1)$$

It was also assumed that ε_n behaves as a Gaussian i.i.d. random process, i.e., $\varepsilon_n \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

Trend μ_n component was described considering a *local linear trend* process, defined as follows [14]:

$$\begin{aligned} \mu_{n+1} &= \mu_n + \beta_n + \eta_n, \qquad \eta_n \sim \mathcal{N}(0, \sigma_\eta^2), \\ \beta_{n+1} &= \beta_n + \zeta_n, \qquad \zeta_n \sim \mathcal{N}(0, \sigma_\zeta^2), \end{aligned} \quad (2)$$

where β_n represents the stochastic slope controlling the rate of rise of the trend.

Moreover, cycle component was represented considering an $AR(\rho)$ model, as follows [14]:

$$\psi_{n+1} = -\mathrm{a}_1\psi_n - \mathrm{a}_2\psi_{n-1} - \dots - \mathrm{a}_\rho\psi_{n-\rho+1} + \xi_n, \quad (3)$$

Where $\xi_n \sim \mathcal{N}(0, \sigma_\xi^2)$. Minus signs were for convenience only. To ensure that ψ_n represented a stochastic cycle component, it was a mandatory-requirement that coefficients $\{\mathrm{a}_1, \mathrm{a}_2, \dots, \mathrm{a}_\rho\}$ gave rise to a wide-sense stationary process [14]. Here, all simulations were carried on setting $\rho = 6$.

C. State space methods

State space models are powerful mathematical structures for describing stochastic time series. Moreover, structural analysis can be easily formulated in the form of state space models, where μ_n and ψ_n constitute the unobserved state, and ε_n the signal perturbations [12], [14]. According to this, the state space structural model allowed to implement PS structural analysis by using state space methods.

First, structural components μ_n and ψ_n were coarsely estimated applying Kalman filter, a method

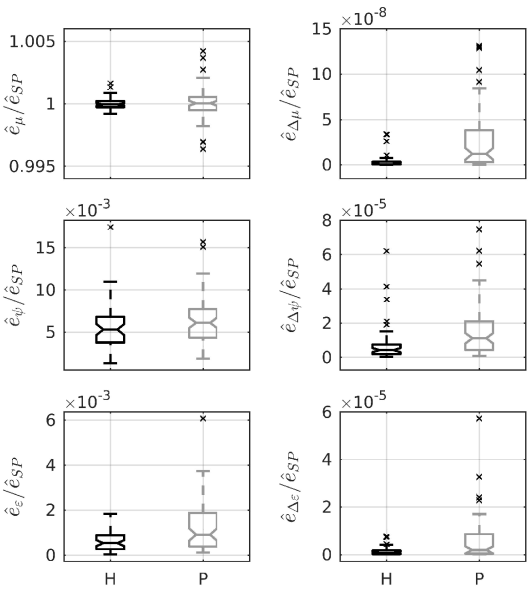


Figure 3: Boxplots comparing normalized RMS structural features computed from Healthy and Pathological PS.

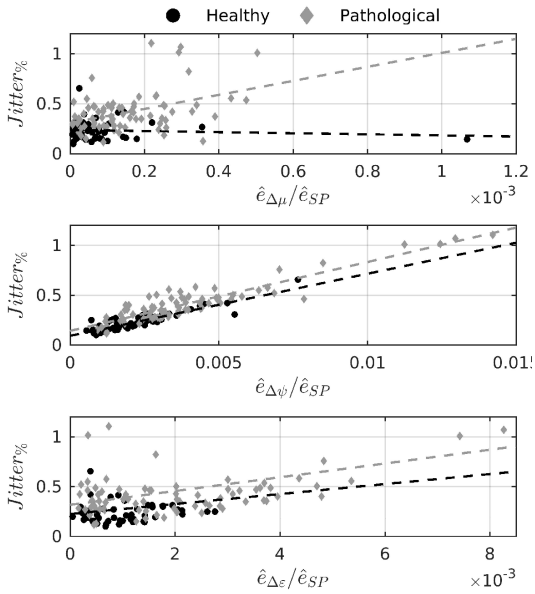


Figure 4: Scatter plots of $Jitter_{\%}$ versus normalized RMS features. *Top:* $\hat{e}_{\Delta\mu}/\hat{e}_{PS}$. *Center:* $\hat{e}_{\Delta\psi}/\hat{e}_{PS}$. *Bottom:* $\hat{e}_{\Delta\varepsilon}/\hat{e}_{PS}$.

for computing the model state considering past and present information. Then, these estimates were improved applying state smoothing, which takes advantage of the entire PS in the state computation. Finally, perturbation smoothing was performed to obtain perturbations ε_n. For further information regarding state space methods, see [15], [19]. This procedure resulted in optimal estimations of μ_n, ψ_n and ε_n components for a given PS.

III. RESULTS

Estimates obtained from the structural analysis of the healthy and the pathological examples previously introduced are displayed in Fig. 1 and Fig. 2, respectively. At the top, it is shown the PS in gray lines, along with the trend μ_n estimates superposed in black lines. It can be observed that trend estimates reproduce the global contour of the PS, describing the slow changes of the fundamental period. In the center, cycle ψ_n estimates are drawn. It can be observed that cycle components capture the oscillatory behavior of the PS because of the auto-regressive formulation, see Eq. (3). Perturbation ε_n estimates are shown at the bottom. It can be observed that these estimates behave as pure stochastic processes. It is interesting to notice that structural components display similar behaviors in both the healthy and the pathological examples.

State space structural analysis was applied for decomposing the healthy and pathological PS, and the structural components μ_n, ψ_n and ε_n were estimated. In order to describe these components, root-mean-square values $\hat{e}$ for the estimated components were obtained. Thus, quantity $\hat{e}_{PS}$ was computed as follows:

$$\hat{e}_{PS} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(P_n\right)^2}\,. \tag{4}$$

Features $\hat{e}_{\mu}$, $\hat{e}_{\psi}$ and $\hat{e}_{\varepsilon}$ were similarly computed. As discrete differences $\Delta P_n = P_n - P_{n-1}$ play an important role in the formulation of acoustic parameters (e.g., absolute jitter and jitter factor), features $\hat{e}_{\Delta\mu}$, $\hat{e}_{\Delta\psi}$ and $\hat{e}_{\Delta\varepsilon}$ were also computed by using Eq. (4). All these RMS features were normalized by dividing by $\hat{e}_{PS}$ in order to reduce the influence of the average fundamental period and the number of vocal cycles.

In Fig. 3, boxplots comparing the normalized RMS structural features computed from healthy (H) and pathological (P) PS are shown. First row suggests that trends μ_n kept most of the global PS information and display very slow dynamics (negligible $\hat{e}_{\Delta\mu}$ values). Second and third rows suggest that cycle components ψ_n and perturbations ε_n are phenomena of comparable magnitudes, where $\hat{e}_{\psi}$ are slightly greater than $\hat{e}_{\varepsilon}$. These results also suggest that $\hat{e}_{\Delta\psi}$ and $\hat{e}_{\Delta\varepsilon}$ are more meaningful (greater values) than $\hat{e}_{\Delta\mu}$, explaining the relevance of ψ_n and ε_n in the classical

perturbation analysis. Finally, the boxplots indicate that discrete difference based features, especially $\hat{e}_{\Delta\psi}$ and $\hat{e}_{\Delta\varepsilon}$, seem to bear significant information for classifying healthy and pathological cases.

Jitter factor, $Jitter_{\%}$, is an acoustical parameter widely accepted in the speech community to assess period variability based on classical perturbation analysis [1], [2]. In order to further study the results obtained with structural analysis, $Jitter_{\%}$ values were computed from the healthy and pathological PS series and then compared with the structural features. For this, only the features based on discrete differences were considered. Simple linear regressions were performed describing whether linear relations were found.

In Fig. 4, scatter plots of $Jitter_{\%}$ versus normalized RMS features for $\Delta\mu_n$ (top), $\Delta\psi_n$ (center) and $\Delta\varepsilon_n$ (bottom) are shown. A linear relationship can be observed only between $Jitter_{\%}$ and $\hat{e}_{\Delta\psi} / \hat{e}_{PS}$. Linear regressions with coefficients of determination $R^2 = 0.791$ and $R^2 = 0.846$ were obtained for healthy and pathological data, respectively, supporting this last statement. In the other cases, linear regressions produced coefficients of determination $R^2 < 0.2$. This analysis agree with the scatter plots, where no clear structures could be observed. These results suggest that cycle components may have a stronger influence on $Jitter_{\%}$, than the perturbations.

V. CONCLUSION

In this work, fundamental period variability was described by means of state space structural analysis. This method allowed for decomposing a period sequence, obtaining simple but physically meaningful components. Structural analysis was performed on type 1 voice signals from healthy and pathological subjects. The results showed that structural components suitable described the different aspects involved in fundamental period variability. The global profile was captured in the trend, while local oscillatory and random behaviors were modeled in the cycle and perturbation components, respectively. This study proved that structural analysis provide more detailed information than classical perturbation analysis, making it a powerful alternative for the assessment of period variability.

REFERENCES

[1] P. H. Dejonckere, P. Bradley, P. Clemente, G. Cornut, L. Crevier-Buchman, G. Friedrich, P. Van De Heyning, M. Remacle, and V. Woisard, "A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques," *Eur Arch Oto-Rhino-Laryngology*, vol. 258, no. 2, pp. 77–82, Feb. 2001.

[2] J. M. Hillenbrand, "Acoustic Analysis of Voice: A Tutorial," *SIG 5 Perspect Speech Sci Orofac Disord*, vol. 21, no. 2, pp. 31–43, 2011.

[3] P. H. Dejonckere, A. Giordano, J. Schoentgen, S. Fraj, B. L., and C. Manfredi, "To what degree of voice perturbation are jitter measurements valid? A novel approach with synthesized vowels and visuo-perceptual pattern recognition," *Biomed Signal Process Control*, vol. 7, no. 1, pp. 37–42, 2012.

[4] C. Manfredi, A. Giordano, J. Schoentgen, S. Fraj, L. Bocchi, and P. H. Dejonckere, "Perturbation measurements in highly irregular voice signals: Performances/validity of analysis software tools," *Biomed Signal Process Control*, vol. 7, no. 4, pp. 409–416, 2012.

[5] J. Schoentgen, "Stochastic models of jitter," *J Acoust Soc Am*, vol. 109, pp. 1631–1650, 2001.

[6] I. R. Titze, *Principles of Voice Production*, 2nd ed. Iowa, USA: National Center for Voice and Speech, 2000.

[7] E. Cataldo and C. Soize, "Jitter generation in voice signals produced by a two-mass stochastic mechanical model," *Biomed Signal Process Control*, vol. 27, pp. 87–95, 2016.

[8] R. Fraile, N. Sáenz-Lechón, V. J. Osma-Ruiz, and J. M. Gutierrez-Arriola, "Characterisation of tremor in normophonic voices," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, 2015, pp. 320–324.

[9] R. F. Leonarduzzi, G. A. Alzamendi, G. Schlotthauer, and M. E. Torres, "Wavelet leader multifractal analysis of period and amplitude sequences from sustained vowels," *Speech Commun*, vol. 72, pp. 1–12, 2015.

[10] C. Mertens, F. Grenez, F. Viallet, A. Ghio, S. Skodda, and J. Schoentgen, "Vocal tremor analysis via AM-FM decomposition of empirical modes of the glottal cycle length time series," in *16th Annual Conference of the International Speech Communication Association (Interspeech 2015)*, 2015.

[11] J. Schoentgen, "Modulation frequency and modulation level owing to vocal microtremor," *J Acoust Soc Am*, vol. 112, no. 2, pp. 690–700, 2002.

[12] G. A. Alzamendi, G. Schlotthauer, and M. E. Torres, "State-Space Approach to Structural Representation of Perturbed Pitch Period Sequences in Voice Signals," *J Voice*, vol. 29, no. 6, pp. 682–692, 2015.

[13] G. A. Alzamendi, G. Schlotthauer, and M. E. Torres, "A new method for structural analysis of perturbed pitch period series," in *VI Latin American Conference on Biomedical Engineering (CLAIB 2014)*, 2014.

[14] A. C. Harvey and N. Shephard, "Structural time series models," in *Econometrics*, vol. 11, G. S. Maddala, C. R. Rao, and H. D. Vinod, Eds. Elsevier, 1993, pp. 261–302.

[15] S. J. Koopman and M. Ooms, "Forecasting Economic Time Series Using Unobserved Components Time Series Models," in *The Oxford Handbook of Economic Forecasting*, M. P. Clements and D. F. Hendry, Eds. Oxford University Press, 2011, pp. 129–162.

[16] Massachusetts Eye and Ear Infirmary (MEEI) Voice and Speech Lab, "Disordered Voice Database." 2009.

[17] M. Markaki and Y. Stylianou, "Voice Pathology Detection and Discrimination Based on Modulation Spectral Features," *IEEE Trans Audio Speech Lang Processing*, vol. 19, no. 7, pp. 1938–1948, 2011.

[18] V. Parsa and D. G. Jamieson, "Identification of Pathological Voices Using Glottal Noise Measures," *J Speech, Lang Hear Res*, vol. 43, no. 2, pp. 469–485, 2000.

[19] J. Durbin and S. J. Koopman, *Time Series Analysis by State Space Methods*, 1st ed. New York, USA: Oxford Univ Pr (Sd), 2001.

Phonation Quality Detection on the Saarbrücken Voice database using Harmonic Spectrum-based Parameters

Wolfgang Wokurek*, Manfred Pützer†
*Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland
†Klinische Phonetik, Institut für Phonetik,
Universität des Saarlandes, Saarbrücken, Deutschland
wolfgang.wokurek@ims.uni-stuttgart.de, puetzer@coli.uni-saarland.de

Abstract: **In this study, voice quality parameters (VQPs) based on amplitude measurements in the harmonic spectrum are surveyed on the "Saarbrücken Voice Database". A trapeziod is used as a simplified version of the volume velocity contour at the vocal folds. It is used to demonstrate the VQPs open quotient and rate of closure in time domain and frequency domain. The VQPs are based on spectral decay gradients to reduce their fundamental frequency dependence. Multivariant analysis of variance (MANOVA) shows significant differences between voice quality parameter means of different speaker groups (group of male and female speakers, group of heathy and pathological speakers, different age groups of speakers).**
Keywords: **speech parametrization, voice quality parameters, harmonic spectrum measurements, statistics**

I. Introduction

The "Saarbrücken Voice Database" is a German database of so called normal and pathological voices [1] [2]. It contains a collection of voice recordings from a longitudinal study of more than 2000 individuals.

Voice quality parameters based on amplitude measurements in the harmonic spectrum were introduced by Stevens and Hanson [3]. These voice quality parameters were found to be robust under realworld disturbances [4] since only little of the noise power is within the narrow band of each harmonic amplitude measurement. The parameters are basically differences of decibel measurements, hence power ratios. Stevens' and Hansons parameter definitions include a reduction of the harmonic amplitudes to compensate the resonant influence of the first formant, which might be viewed as an attempt of inverse filtering the speech sound. The first four formants are compensated including their bandwidths.

A speakers' voice quality is influenced by the fundamental frequency e.g. due to subglottal pressure and strains in the vocal folds and other parts of the larynx. Appart from that an increase of the fundamental frequency increases the spectral amplitude differences due to the spectral decay of the source spectrum (also known as spectral tilt). To reduce this fundamental frequency dependence of the voice quality parameters the spectral amplitude differences are replaced by decay gradients [5].

The Saarbrücken Voice Database

The pathological part of the database was collected in a collaborative project of the department of Phoniatrics and Ear Nose Throat (ENT) at the Caritas clinic St. Theresion in Saarbrücken and the Institute of Phonetics of the Saarland University. The so called normal voices were collected in collaboration with different institutions like schools in a circuit of the city of Saarbrücken, Germany. The collection of the database has combined research methodologies from speech science with phoniatric methods. Methods from speech research which were used are electrogottography (EGG) and recording of the sound pressure waveform (microphone signal). Electroglottogram (EGG) and

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

microphone signals were recorded simultaneously. Both signals were fed directly into a Computerised Speech Lab (CSL) station (model 4300B) at a 50-kHz sampling rate with 16-bits amplitude resolution. The microphone signal was recorded using a headset condenser microphone (NEM 192.15, Beyerdynamic, Heilbronn, Germany). The EGG-signal was acquired with a Portable Laryngograph from Laryngograph Ltd. The phoniatric investigation only applied for the pathological voices consisted of video recordings of the vocal folds, using a laryngoscope and additional stroboscoby.

One recording session contains the following recordings: Recording of the vowels [iː, aː, uː] produced at normal, high and low pitch. Recordings of the vowels [iː, aː, uː] with rising-falling pitch. Recording of the sentence "Guten Morgen, wie geht es Ihnen?" (Good morning, how are you?)

II. Methods

A. *Voice Quality Parameters*

Throughout this paper only signals of sonorant voiced sounds are considered. This implies that a periodic or quasi periodic structure is prevalent and hence a harmonic structure in the spectrum. Minor deviations are acceptable but e.g. if the ratio of the frequencies of peaks of the first two harmoncis deviates more than 10% an error condition is marked. No attempt to deal constructively with such situations is made in this survey. The corresponding frames are excluded from the statistical evaluation.

B. *Trapezoid Model*

The fundamental notion of how harmonic voice quality parameters work may easily be demonstrated by a trapeziod model shown in Figures 1 and 2. There the trapeziod is used as a simplified version of the volume velocity contour at the vocal folds. It is capable to demonstrate two (not so obvious) correspondences between time domain and frequency domain features that are related to phonation behavior.

The dominant phonation parameter is the fundamental frequency F_0 and its reciprocal the fundamental period $T_0 = \frac{1}{F_0}$. These are held constant at 100Hz and 10ms here. The second most obvious phonation parameter is the temporal ratio between the open and the fundamental period technically called duty cycle. As a voice quality parameter this is called the open quotient (OQ). Figure 1 demonstrates that an increase of OQ in time domain increases the height difference between the first two harmonic amplitudes (marked by bold circles). The amplitude is logarithmic (corresponding to a linear decibel scale) and $OQ = H_1 - H_2$. This basic subtration structure is not changed in subsequent modifications. Only formant resonances will be subtracted from the amplitudes. And finally it will be divided by its frequency difference yielding a spectral decay rate.

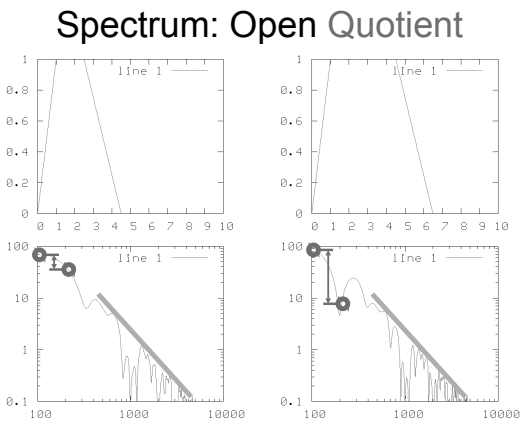


Figure 1: Open quotient: trapezoids duty cycle

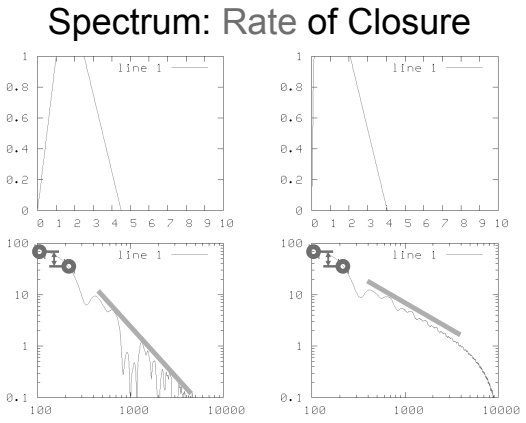


Figure 2: Rate of closure: trapezoids steepest shoulder

The third phonation parameter is related to the

steepest shoulder of the trapezoid. It is tightly bound to both, the phonation cycle and the spectral envelope. What is colloquially called a louder voice is produced with increased subglottal pressure and more strained vocal fold behavior than a softer voice. This phonation behavior basically yields a more rapid interruption of the airflow (volume velocity) by the closing vocal folds. It corresponds to the time domain phonation parameter rate of closure (RC). In the trapezoid model this is modelled by the steeper shoulder. In the signal spectrum this changes the decay rate of the spectral envelope above the first harmonics. In Figure 2 the steeper shoulder in the right trapezoid in time domain yields a flatter (less spectral tilt) envelope in frequency domain (solid line).

Notice that when changing OQ in Figure 1 RC is constant and when changing RC in Figure 2 OQ is constant and so are the corresponding spectral features (bold circles and solid lines). Consider also that the correspondences between the time domain OQ_t and RC_t and the frequency domain OQ_f and RC_f are qualitative and seem to hold for increase and decrease but not for proportionality $OQ_t \not\propto OQ_f$ and $RC_t \not\propto RC_f$.

C. *Spectral Measurements*

The voice quality parameters of this survey are based on amplitude and frequency measurements of several harmonic peaks, an estimate of the fundamental frequency and formant parameter estimates [6]. The harmonic peaks are searched in a short term spectrum with a 25ms (Hamming) window. This window is long enough to show the spectrum of two or more fundamental periods in order to reveal the speech signals harmonic structure. The analysis is repeated every 10ms.

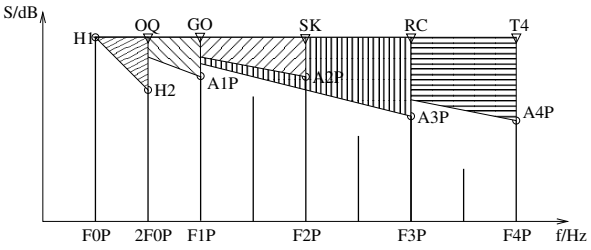


Figure 3: Spectral amplitude measurements and decay gradient triangles

Figure 3 shows a qualitative example of the peak search result. F0P is the frequency where the first harmonic peak is found and H1 is its amplitude in decibels. F1P and H1 are the measurements of the second harmonic peak. Now the LPC based formant frequency estimates F1 - F4 are used. The next four peaks are searched next each formant frequency yielding the harmonics next to each formant F1P - F4P and their amplitudes A1P - A4P.

D. *Error Handling*

Sufficient harmonic structure is necessary for the spectral amplitude measurements of Figure 3. This is checked by the probability of voicing (a by-product of the fundamental frequency estimation procedure) above 50%, and by the first and the second harmonic peak. Frames are accepted if the frequency ratio of these peaks is closer than 10% to two. Only 3.4% of the frames lack the harmonic structure.

Unfortunately there is a single voice quality parameter that requires a signal structure that is not met in half of the frames. The $GO = H1 - A1P$ parameter relates the first harmonic and the harmonic next to the first formant. For high F0 and low F1 this can be the same harmonic, hence the model is not applicable. Currently this situation is simply indicated by an error condition and the frames are dropped (46%).

Due to all error conditions nearly half of the frames (49%) are lost. In particular 69% of the female and 23% of the male frames.

III. Results

First, the results of an analysis of variance (ANOVA) show that parameter means of all VQPs (6) as dependent variables are relevant for the effect of male versus female speaker differentiation in the two groups (normal and pathological voices, see Tab.1) Second, using the same VQPs an ANOVA differentiate so called normal speakers from speakers assigned in the database as pathological. The two groups are also significantly different in six parameters for both genders (normal male versus pathological male, normal female versus pathological female; see Tab.1). Finally, by using all signals in the database an ANOVA with posthoc-test

Table 1: VQPs of normal and pathological speakers: means (standard deviation), $p < 0.001$

	Normal		Pathological	
	male	female	male	female
OQGi	4.61 (2.83)	3.58 (2.01)	5.30 (3.30)	4.08 (2.53)
GOGi	4.52 (2.46)	2.88 (1.21)	4.77 (2.69)	3.37 (1.68)
SKGi	1.98 (0.70)	2.42 (0.86)	2.19 (0.80)	2.52 (0.92)
RCGi	1.05 (0.68)	1.70 (0.75)	1.20 (0.80)	1.64 (0.83)
T4Gi	0.22 (0.75)	0.63 (0.79)	0.15 (0.85)	0.59 (0.88)
IC	0.26 (0.13)	0.32 (0.20)	0.27 (0.16)	0.33 (0.20)

Table 2: VQPs for age group differentiation: mean (standard deviation), $p < 0.05$

	0-30	30-50	50-70	70-
OQGi	4.00 (2.51)	4.75 (2.90)	5.07 (3.19)	5.33 (3.46)
GOGi	3.67 (2.05)	4.24 (3.38)	4.47 (2.59)	4.65 (2.78)
SKGi	2.19 (0.81)	2.25 (0.81)	2.29 (0.87)	2.33 (0.84)

(Scheffé alpha adjustment) significantly differentiate four age groups on the basis of three VQPs (OQGi, GOGi, SKGi; see Tab.2). The latter parameters are highly relevant in separating the four age groups just as they are for gender differentiation and voice classification of normal and pathological voices. The mean values of this parameters provide information about adduction behavior of the focal folds and the degree of glottal opening. They further allow an incomplete glottal closure with restriction of the glottal function to be identified. This phonation behavior can be firstly demonstrated when the parameter means of the two genders are compared in both groups (normal and pathological group). Secondly, a comparison of parameter means between the groups also shows this tendency. Higher means indicate a better adduction behavior than lower once (see Tab.1). Finally, this tendency can also be revealed when different age groups are compared with each other. Less adduction behavior and a more incomplete glottal closure may be generally supposed for ageing voices (see Tab.2).

References

[1] M. Pützer and J. Koreman, "A german database of patterns of pathological vocal fold vibration," *Phonus*, vol. 3, pp. 143–153, 1997.

[2] M. Pützer and W. J. Barry, "Saarbrücken voice database," http://www.stimmdatenbank.coli.uni-saarland.de/, 1997.

[3] K. M. Stevens and H. M. Hanson, "Classification of glottal vibration from acoustic measurements," in *Vocal Fold Physiology*, O. Fujimura and M. Hirano, Eds. Cambridge MA: Hiltop University Press, 1998, pp. 147–170.

[4] M. Lugger, B. Yang, and W. Wokurek, "Robust estimation of voice quality parameters under realworld disturbances," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1, May 2006, pp. 1097–1100.

[5] M. Pützer, W. Wokurek, and J. R. Moringlane, "Evaluation of phonatory behavior and voice quality in patients with multiple sclerosis treated with deep brain stimulation," *Journal of Voice*, vol. 31, no. 4, pp. 483–489, 2016.

[6] W. Wokurek and M. Pützer, "Automated corpus based spectral measurement of voice quality parameters," in *Proc. 15th ICPhS (Barcelona)*, 2003, pp. 2173–2176.

ACOUSTIC TREMOR MEASUREMENT: COMPARING TWO SYSTEMS

M. A. E. Brückl, E. Ibragimova, S. Bögelein

Institute for Language and Communication, Technische Universität Berlin, Berlin, Germany
markus.brueckl@tu-berlin.de
elvira.ibragimova@campus.tu-berlin.de
silke.boegelein@campus.tu-berlin.de

Abstract: **A study is presented comparing two software systems that measure vocal tremor acoustically by analyzing sustained vowels. As measure for the comparison serves the criterion validity, here derived from the determination coefficients of simple linear regressions between the tremor measures and the synthetically given tremor values. For this purpose, the vowels to be analyzed were generated completely by acoustic synthesis. The two systems in comparison are a proprietary and widely, also clinically, used voice quality measurement tool and a self-developed algorithm that is based on autocorrelation of pitch and amplitude contours and implemented as a script of an open-source speech analysis program. The comparison's result is that the open-source software clearly achieves the more valid measurements.**
Keywords : **Vocal tremor, acoustic measurement, system comparison, open-source software**

I. INTRODUCTION

The acoustic measurement of vocal tremor bears a high potential to serve for early diagnosis of several, mostly neuro-degenerative diseases like Parkinson's (PD), Alzheimer's, multiple sclerosis, etc. Tremor often is defined as involuntary cyclic movement, or movement deviation, of the limbs. But, at least if it is caused by deficits of the central nervous system, it is most likely that speech production is affected too, since the production of speech involves the coordinated processing of about 1,400 motor commands per second. So, the more than 80 muscles of the vocal apparatus may all show tremor and thus vocal tremor may have many sources. But once the acoustic output is investigated, all of these organic modulation sources combine to only two types of tremor: subsonic quasi-cyclic modulations of the frequency and of the amplitude. And the acoustic signal may easily be captured.

In spite of the potential of auditive or acoustic vocal tremor assessment, its reliability and therewith its validity still provide great room for improvement. This may be a reason why e.g. simple perturbation measures are used in multi-feature PD detection systems [1, 2], whereas more specific tremor features are either not even evaluated to contribute to the system [1] or they are rather circuitously derived via frequency-domain techniques, but not directly within the time-domain [2], and are thus more error-prone.

Hence, the aim of this study is to compare two acoustic tremor measurement systems according to their criterion validity, that is here defined as goodness in measuring synthetically generated and thus known tremor.

II. METHODS

A. Acoustic synthesis of the test stimuli with known tremor properties in three steps

A completely synthetic sustained vowel is created by formant synthesis. (1) The glottal source signal (3s duration, 200Hz mean fundamental frequency ($\bar{F}_0$) is modelled according to [3] and then (2) filtered by a time-invariant 'female'-/a/-shaped filter function. This /a/-sound, which is perceived as rather natural, serves as the carrier for the frequency and amplitude modulations. (3) These modulations are done by re-synthesis according to the overlap-and-add method [4]. Both modulation types are modelled with a sinusoidal shape that is varied in frequency and amplitude, resulting in 4 synthesis arguments: the frequency tremor frequency (FTrF [Hz]), the amplitude tremor frequency (ATrF [Hz]), the relative frequency tremor intensity (FTrI [%]), and the relative amplitude tremor intensity (ATrI [%]). Each argument is varied in 4 equally spaced steps across each range of naturally occurring values. Additionally, both a frequency (decF) and an intensity decline (decA) are synthesized and varied in order to also simulate these naturally occurring effects. Thus, the synthesis of the modulations may be formulated as functions of time (t):

$$F_0M(t) = F_{0,s} + FTrI \cdot \bar{F}_0 \cdot sin(FTrF \cdot 2\pi \cdot t) - decF \cdot t \quad (1)$$

$$AM(t) = A_s + ATrI \cdot \bar{A} \cdot sin(ATrF \cdot 2\pi \cdot t) - decA \cdot t \quad (2)$$

where $F_{0,s}$ and A_s are the fundamental frequency resp. the amplitude at the sound's start that are depending on the sound's duration, on the means, and on the declines.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

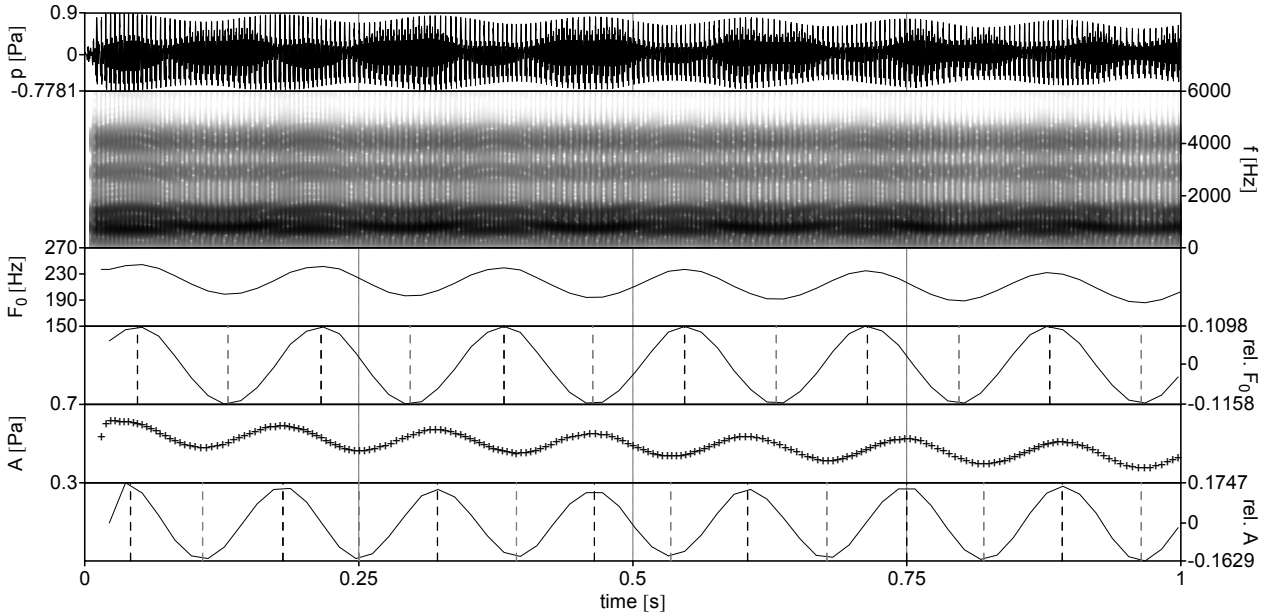


Figure 1: An exemplary synthesized sound and its tremor analysis by TREMOR.PRAAT*: The – from top to bottom – 1st subfigure shows an oscillogram of the first second of the synthesized sound with set tremor values of FTrF=6.0Hz, ATrF=7.0Hz, FTrI=11.5%, ATrI=15.5% as well as declines of decF=15Hz/s and decA=0.15Pa/s. The 2nd subfigure displays a short-time spectrogram of this sound. The contour in Subfigure 3 depicts* TREMOR.PRAAT*'s F_0-analysis. Subfigure 4 contains this F_0-contour, but de-declined and normalized. The short dashed vertical lines denote the times of minima (gray lines) and maxima (black lines) found by* TREMOR.PRAAT*. The 5th subfigure shows the sound's amplitudes per period, extracted by* PRAAT*'s To Amplitude. Subfigure 6 depicts the resampled, de-declined and normalized amplitude contour, again with found minima and maxima.*

A sound example is shown in Fig. 1. The sinusoidal shape and the decline of the amplitude envelope can be seen in particular from SubFig. 1. The frequency modulation may be recognized by the cyclic changes in the density of the glottal pulses in SubFig. 2.

In total $4^6 = 4{,}096$ test sounds result from a complete variation of the 6 synthesis arguments. All 3 synthesis steps as well as the arguments' variation are implemented as a PRAAT [5] script that is added to [6].

B. The tremor measurement systems

The two compared systems are (1) the Multi-Dimensional Voice Program (MDVP) [7] and (2) TREMOR.PRAAT, version 3.01 [6], a revised version of the algorithm presented in [8], including some newly developed tremor measures.

MDVP is a commonly known and widely used voice quality measurement tool. Its standard procedure extracts 4 tremor measures that should correspond to the above mentioned synthesis arguments (MDVP is proprietary software, thus computational details are not known): The frequency of the strongest low-frequency modulation of the fundamental frequency (Fftr [Hz]) or respectively of the amplitude (Fatr [Hz]), and the mean magnitude of the strongest low-frequency modulation of the fundamental frequency (FTRI [%]) or respectively of the amplitude (ATRI [%]).

TREMOR.PRAAT is open-source software and implemented as a PRAAT script. It extracts 14 tremor measures. 4 out of these 14 meet the definitions of the above named MDVP measures, i.e. they also correspond theoretically to the synthesis arguments and are named like them. TREMOR.PRAAT determines the tremor frequencies (FTrF and ATrF) by autocorrelating the F_0-contour, see SubFig. 3 of Fig. 1, and the amplitude contour, see its SubFig. 5. But before the contours get autocorrelated, the linear declines are removed by subtracting the linear regression estimates. Also, the amplitude contour must be resampled at a constant time step, since PRAAT's *To Amplitude* function extracts amplitudes per time-varying periods.

For the computation of the intensity indices (FTrI and ATrI), the contours are normalized, i.e. the deviations about the means ($\bar{F}_0$ or $\bar{A}$) are expressed relative to these means in the analyzed sound – just like in the MDVP:

$$rel.F_0(t) = \frac{F_0(t) - \bar{F}_0}{\bar{F}_0} \;;\; rel.A(t) = \frac{A(t) - \bar{A}}{\bar{A}} \qquad (3)$$

This normalization is needed, since *tremor intensity* shall denote the magnitude of a cyclic deviation, and

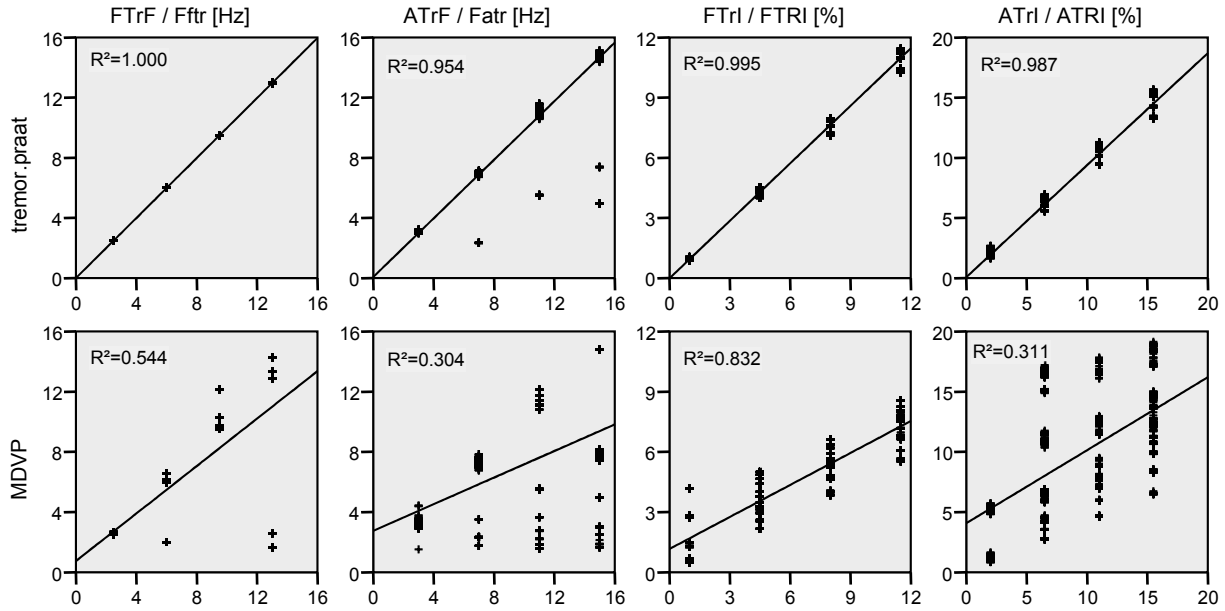


Figure 2: Scatterplots showing the measured values (ordinates) as a function of the values that were set by synthesis (abscissae). The lines are the linear regression models.

thus it should be expressed relative to its mean. The points in time at which this deviation magnitude is largest and that additionally fit to the already determined tremor frequency are found by PRAAT's function *To PointProcess (peaks)*. These steps are visualized in Sub-Fig. 4 and 6 of Fig. 1: The vertical lines mark the times of found extrema. The ordinates of each contour at these times are the searched tremor magnitudes (max, min). Finally, these magnitudes get averaged to the tremor intensity indices:

$$(F,A)TrI = \left(\frac{\sum_{i=1}^{m}|max_i|}{m} + \frac{\sum_{j=1}^{n}|min_j|}{n}\right) \div 2 \qquad (4)$$

where n and m denote the numbers of the found minima resp. maxima.

The default settings of the *search ranges for tremor frequencies* were expanded in both programs to 1.5Hz – 16 Hz. The *amplitude tremor octave cost* was raised to 0.2 in TREMOR.PRAAT in order to compensate for the unnaturally high cyclicality of the synthetically generated tremor contours that induces – together with the rather large analysis window and the sinusoidal shapes – suboctave errors in determining ATrF, see Discussion.

C. Statistical methods

In order to assess the dependence of the 8 measured values on the values that are set by synthesis, 8 simple linear regressions are computed. Their determination coefficients (R^2) denote the proportion of variance in the measured values that can be explained by the set values' variance, thus they may serve as coefficients of validity of the measurement instrument. 99.99% confidence intervals (CIs) around these coefficients are calculated in order to indicate if the populations of corresponding coefficients differ from another.

III. RESULTS

The results of the regression analyses are shown in Fig. 2: MDVP fails to extract amplitude tremor measures in 513 cases and frequency tremor measures in 256 cases. Although TREMOR.PRAAT achieves to extract all measures from all sounds, its errors are highly significantly smaller, i.e. its measures are highly significantly more valid than those of the MDVP. In order to illustrate this significant superiority, Fig. 3 shows that the best estimates of R^2 do not fall within the CIs of corresponding measures of the other system, and that TREMOR.PRAAT's coefficients always denote higher validities.

TREMOR.PRAAT's measurement of FTrF is (nearly) totally valid: The regression line fits all data points and equals the coordinate system's angle bisector. Also, the other TREMOR.PRAAT measures can be considered excellent. In contrast the MDVP's extractions exhibit considerably more and greater measurement errors.

The MDVP is not built to be able to cope with naturally occurring declines, neither of the amplitudes nor of the frequency. In order to adjust for this, a further statistical analysis was executed that was reduced to the 4^4 = 256 sounds without any decline. But the highly significant differences between the two measurement systems remain – again with a confidence greater than 99.99%, just like in the analysis that comprises all 4.096 sounds.

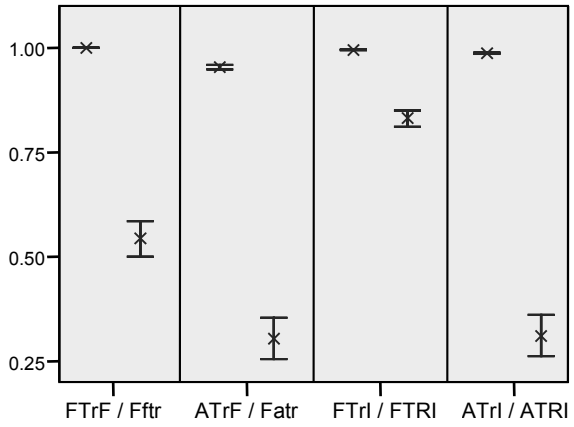


Figure 3: The best estimates (x) and the 99.99% CIs (double-T-bars) of the regressions' determination coefficients (R^2): TREMOR.PRAAT*'s measures are highly significantly more valid than those of the MDVP.*

IV. Discussion

All errors in TREMOR.PRAAT's measurements may be reduced by shortening the *analysis time step* (default value: 0.015s), at the cost of an exponentially increasing computational load.

TREMOR.PRAAT's tremor intensity measures (FTrI and ATrI) exhibit greater underestimations at greater synthetically set values. These errors are due to the combination of the sinusoidal shapes of the modulations with the averaging of these shapes within analysis windows: Sinusoids reach extreme values only punctually, whereas analysis windows mandatorily span a duration.

If ATrF gets extracted deficiently, then exactly one or two octaves too low, cp. Fig. 2. These octave errors result from correctly detecting sub-harmonics of the modulation frequencies that – again – are artificially induced by sampling the synthetically exactly sinusoidal contours at a rather low rate. Additionally to reducing these errors by shortening the *analysis time step*, they can be avoided by further raising the *tremor octave cost* argument. Apart from that, these errors will hardly occur when analyzing natural sounds, since natural tremor modulations are far less cyclic, wherefore a "rough" sampling seldom will construe sub-harmonics.

Errors in the MDVP's extractions seem to be far less systematic. Their sources must remain unrevealed, since the MDVP's algorithm is proprietary and thus unknown.

Besides, TREMOR.PRAAT still is developed to comprising more indices that in their totality are perceptually and biologically more valid for the concept of *tremor* than those alone that are already known and implemented: The newly developed indices FTrP and ATrP, for example, combine tremor frequency and intensity. As reported in [9] they seem to better picture the medical concept of *tremor severity* than the known intensity indices and thereby indicate PD – provided that the speakers' age and sex is considered. Furthermore, the concept of *cyclicality* is highly likely to contribute to a holistic concept of *tremor strength* or *severity*, just as well as to consider the fact that often there is not just one, the strongest, tremor frequency in a voice. Consequently, the most recent inventions in TREMOR.PRAAT [6] are indices that integrate tremors at multiple frequencies, whereat considering each cyclicality and intensity.

V. Conclusion

Although TREMOR.PRAAT is still under development, it has been shown that it is already far more valid in measuring vocal tremor than the standard program MDVP. Thus, it can only be advised to use TREMOR.PRAAT for acoustic tremor measurement. Furthermore, formerly gained results that were based on the MDVP's tremor measures are very likely to improve in precision and variety if they were re-measured with TREMOR.PRAAT. Also, the PD detection rates of the approaches described in [1] and [2] are likely to improve, if the measures of TREMOR.PRAAT were added to the feature sets.

References

[1] A. Tsanas et al., "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease", *IEEE Transactions on Biomedical Engineering*, vol. 59 no. 5, pp. 1264–1271, 2012.

[2] D. Hemmerling et al., "Automatic detection of Parkinson's disease based on modulated vowels", *INTERSPEECH–2016*, San Francisco, pp. 1190–1194, 2016.

[3] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels", *Journal of the Acoustical Society of America*, vol. 49, pp. 583–590, 1971.

[4] E. Moulines, F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Communication*, vol. 9, pp. 453–467, 1990.

[5] P. Boersma, D. Weenink, *Praat: doing phonetics by computer (Version 6.0.29)* [Computer program], University of Amsterdam. http://www.praat.org/, accessed on 25 June 2017.

[6] M. A. E. Brückl, *tremor.praat (Version 3.01)* [Computer program], Technische Universität Berlin. http://brYkl.de/tremor3.01.zip, 2017.

[7] Kay Elemetrics Corp. / PENTAX Medical, *Multi-Dimensional Voice Program (MDVP), Model 5105 (Version 2.6.2)* [Computer program], 1993/2003.

[8] M. A. E. Brückl, "Vocal tremor measurement based on autocorrelation of contours", *INTERSPEECH–2012*, Portland (OR), 715–718, 2012.

[9] M. A. E. Brückl, A. Ghio, F. Viallet, "Measurement of tremor in the voices of speakers with Parkinson's disease", *1st International Conference on Natural Language and Speech Processing*, Algiers, 44–48, 2015.

THE DATABASE OF NORMAL AND PATHOLOGICAL SINGERS' VOICES: AN APPROACH TO COLLECTING DATA

V.V. Evdokimova[1], K. V. Evgrafova[1], P.A.Skrelin[1], T.V. Chukaeva[1]
[1] Saint Petersburg State University, Department of Phonetics, Saint-Petersburg, Russia
postmaster@phonetics.pu.ru, karinaevgr@mail.ru, skrelin@phonetics.pu.ru, chukaeva68@mail.ru

Abstract: **The given paper presents an approach to recording a database of normal and pathological singers' voices. In the paper the procedure of recording is described. The subjects are classified according to the healthy state of their voices. The database can be used for different biomedical and phonetic studies. The data obtained can be applied to many applications such as speech/speaker recognition, speech synthesis, emotion identification, age identification, speech coding and various medical applications.**

Keywords: **speech signal, singing voice, voice pathologies**

I. INTRODUCTION

Traditionally, the acoustic analysis of voice for clinical purposes has been made on sustained vowels [1] and a set of parameters measuring voice instability are of common use in clinical software for voice analysis [2]. However, it is not an easy task to extrapolate these results and methods to running speech [2], since the stationarity assumption only holds for sustained phonations. Besides, the sustained vowels are not easy to use for phonetic analysis of speech. In this paper, a small database of speech material for speech in healthy state and speech of ill people is presented.

II. METHODS

The recording experiments were conducted with the use of Ling Wave microphone recording system. The goal of the experiments was to obtain the samples of singing, reading and producing isolated vowels from the same subject in different physical states: normal and with certain pathologies resulting in phonation problems.

Eleven trained singers of the Mikhailovsky Theatre in Saint-Petersburg were involved in the experiments (6 female singers and 5 male singers). They were instructed to read a phonetically representative text in a comfortable rate and also produce isolated Russian vowels at comfortable pitch level and at higher and lower than comfortable pitch level. The subjects were also asked to sing one of the Russian classical romances for not more than 2 minutes. An average length of the recording for each speaker was from ten to fifteen minutes. Besides the video signal of the vocal cords phonation process was recorded simultaneously with the audio data.

III. RESULTS

The professional phoniatrician was employed in the process of recording the informants and establishing diagnosis. The recorded database contains the speech of the subjects with different diagnoses: 3 subjects had vocal cord nodules, 2 subjects were tested after the vocal cord hemorrhage state, and 1 subject had age-related larynx hypotonia.

IV. CONCLUSION

The principles of the database construction allow obtaining the samples of a singer's normal and pathological (e.g. with acute respiratory disease) voice. Besides, the database makes it possible to investigate patients with the same diagnosis.

We plan to extend the database to obtain representative sampling.

At the moment, the recordings contain different speech types: a 2 minute of Russian classical romance, read phonetically representative text, isolated vowels uttered in different frequency registers.

The recording procedure design allows processing the voice samples with the use of medical software products such as LingWave. Besides, the complex phonetic analysis (auditory, articulatory, and perceptive) is also possible. On the whole the data can be applied in the research of singing voice [3, 4, 5].

REFERENCES

[1] V. Parsa, and D.G. Jamieson, "Acoustic Discrimination of Pathological Voice: Sustained Vowels Versus Continuous Speech" *J Speech Lang Hear Res* vol.44, pp.327-339, 2001.

[2] Y. Zhang, J.J. Jiang, "Acoustic Analyses of

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

Sustained and Running Voices from Patients with Laryngeal Pathologies", *J Voice*, vol.22, pp.1-9, 2008.

[3] K. Evgrafova, V. Evdokimova, "Perception of Russian Vowels in Singing," *Frontiers in Artificial Intelligence and Applications*, *EEE*. *Trans*. *Biom*. *Eng*, vol. 247: Human Language Technologies – The Baltic Perspective, pp. 42-49, 2012.

[4] K. Evgrafova, V. Evdokimova, "Acoustic analysis of vocal fatigue in professional voice users", *International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications - MAVEBA 2011*, Florence, pp. 153-156, 2011.

[5] K. Evgrafova, V. Evdokimova, P.Skrelin, T.Chukaeva "The study of Acoustic-Articulatory Relations in Producing Singing Vowels with the Use of EMA", *International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications - MAVEBA 2015*, Florence, pp. 95-98, 2015.

SESSION II:

VOICE QUALITY MONITORING

MONITORING VOICE CONDITION USING SMARTPHONES

F. Schaeffler, J. Beck

CASL Research Centre, Queen Margaret University Edinburgh, Scotland, UK
fschaeffler@qmu.ac.uk, jbeck@qmu.ac.uk

Abstract: **Smartphone mediated voice monitoring has the potential to support voice care by facilitating data collection, analysis and biofeedback.**

To field-test this approach we have developed a smartphone app that allows recording of voice samples alongside voice self-report data. Our long-term aim is convenient and accessible voice monitoring to prevent voice problems and disorders. Our current study focussed on the automatic detection of voice changes in healthy voices that result from common transient illnesses like colds.

We have recorded a database of approximately 700 voice samples from 62 speakers and selected a subset of 225 voice samples from 8 speakers who had submitted at least 10 recordings and reported at least one instance of a moderate cold. We extracted 12 acoustic parameters and applied multivariate statistical process control procedures (Hotelling's T^2) to detect whether instances of cold caused violations of distributional control limits.

Results showed significant association between control limit violations and reporting of a cold. While there is scope for further improvement of sensitivity and specificity of the procedure, it could already support early detection of voice problems, especially if mediated by voice experts.

Keywords: **voice problems, monitoring, acoustic analysis, smartphones**

I. INTRODUCTION

Modern smartphones offer entirely new approaches to personal health by facilitating data collection, analysis and biofeedback. This offers new methods for tackling occupational voice problems, which are endemic in some professions [1], [2].

Most occupational voice problems are behavioural (i.e. arising from ineffective voice use) [3], so can potentially be prevented through early recognition and behavioural changes. We aim to develop an early warning system for voice problems via a smartphone app, whereby people in vocally demanding professions can routinely monitor their voice and receive tailored advice if necessary. Smartphones are widely used nowadays and a number of studies suggest smartphone audio recordings can reliably be used to extract acoustic voice parameters (see e.g. [4], [5]).

Health monitoring systems often consider patterns of deviation from baseline performance as well as static thresholds. Many human physiological factors (e.g. blood pressure, body temperature) show fluctuation patterns that can be indicative of health state [6]. For voice, too, fluctuation patterns in acoustic parameters could be indicative of vocal health. To study acoustic voice fluctuation patterns we are currently recording a longitudinal database of typical and 'at risk' voices, sampled frequently over several weeks through a smartphone app. This app records voice samples and a number of voice-related self-reports alongside each recording.

To monitor voice condition in individuals we are employing statistical process and quality control procedures [7]. These procedures are designed to detect variations in patterns that indicate non-random or 'special' causes and can be applied to univariate and multivariate situations.

We assume that stability over time is an indicator of system integrity for healthy voices. Our current aim is to analyse whether acoustic parameters derived from mobile phone recordings are a) robust enough to remain stable under normal conditions, i.e. do not exceed limits expected due to normal cause variation and b) sensitive enough to pick up minor variations in the acoustic voice profile of voice users, i.e. successfully detect special cause variation that is due to changes in the user's voice.

As a test case for detecting deviations from regular voice patterns we chose instances of self-reported common colds and similar illnesses by participants, as we have so far mainly recorded speakers who do not report regular problems with their voices. Upper respiratory tract infections (URTIs), especially when accompanied by acute laryngitis, have effects on the voice that may be similar to those encountered in occupational voice problems (e.g. hoarseness, weak voice or voice loss). Successful detection of cold-related voice symptoms

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

would therefore indicate a level of sensitivity that could support a broader range of applications.

Detection of such changes could also have more direct benefits. URTIs are a recognized risk factor in development of voice disorder [8], so if detection of cold-related changes could trigger provision of appropriate advice when most needed (e.g. reduction of voice use and techniques for reducing vocal fold impact), this could help to prevent occupational voice problems. In addition, the ability to track whether voices return to baseline after a cold may help to differentiate transient voice changes from longer lasting or chronic ones.

II. METHODS

To collect frequent voice samples from a range of speakers we developed a mobile phone app, the 'voicecheck' app, which is available for Apple iOS and Google Android devices in UK app stores. The app records audio data in uncompressed wav (pcm) format with a sampling frequency of 44 kHz, and prompts a survey alongside each recording.

For the current study the app prompted the recording of two sustained [a] vowels at a comfortable pitch and loudness, with each vowel sustained for at least 3 seconds. Afterwards participants read 9 sentences and a short passage of text (a modified and shortened version of the 'dog and duck story' [9]).

Participants were instructed to control microphone distance by holding the phone approximately a handspan (20 cm/8 inches) from their mouth.

The survey consisted of 12 questions that addressed voice use prior to recording, psychological stress, room size/configuration, current state of the participant's voice, recent throat sensations and whether the participant currently had a cold on a scale with four levels: no cold, mild, moderate, severe. For further analysis in the present study, the 'cold' variable was transformed into a binary variable by counting "no cold" or "mild cold" as 0 and all other instances of "cold" as 1.

After audio recording and survey completion, all data was securely transferred to a central server.

Participants signed up and provided consent for the project through a website (voicecheck.org.uk). After sign-up, participants received an electronic schedule of 50 recordings as a calendar (ics) file for integration into their smartphone calendar app of choice.

Success for triggering automatic reminders by this method was variable as some calendar apps did not recognise the trigger. Recording events were distributed over twelve weeks, with more intensive and less intensive weeks and 2-3 recordings per recording day. Triggers prompted recordings at 7am, 1pm and 7pm on weekdays and 9 am and 7 pm on weekends. Over the course of the project it turned out that many participants found it difficult to stick to the schedule and were therefore instructed to provide recordings whenever suitable, but leaving at least 4h between recordings.

The database currently contains around 700 recordings from 62 speakers. For the present study we selected data from 8 speakers who had completed at least 10 recordings and had recorded instances of a cold or similar illness once or more at moderate level over the course of their recordings. Table 1 provides general information about the individual speakers.

We extracted 12 acoustic parameters from the connected speech samples, using Praat [10]. Audio processing was performed in two steps, using two different Praat scripts. The first script separated sustained vowels from both sentences and connected speech, and removed pauses and unvoiced stretches from the signal, applying the method described and using parts of the script published in [11]. Only these pre-processed connected speech samples (i.e. sentences and passage of text combined) were used for further analysis in the current study.

The second script extracted the 12 acoustic parameters from the pre-processed audio files. These comprised all AVQI parameters as described in [12], using the implementation in [11]. These were smoothed cepstral peak prominence (CPPS), harmonics-to-noise ratio (HNR) as implemented in Praat, shimmer local (Shim) and shimmer local dB (ShdB), the general slope of the spectrum (Slope) and the tilt of the regression line through the spectrum (Tilt). To this we added mean F0 (Praat's cross-correlation algorithm), jitter (RAP), jitter (PPQ5), Glottal Noise Excitation Ratio [13], [14] and uncorrected (H1-H2) and corrected (H1*-H2*) first and second harmonic difference in our own implementation, following the procedure described in [15].

Prior to analysis we calculated correlations for all extracted parameters and inspected correlations of Pearson's r above 0.7. This led to the exclusion of both jitter measures as they showed high correlation with CPPS. Shim correlated highly with ShdB and the latter was kept as it showed less correlation with CPPS. H1-H2 showed high correlation with H1*-H2*. We kept the corrected version as it should provide a better estimate of harmonic energy at the glottis.

For the remaining 8 parameters we constructed multivariate Hotelling T^2 control charts using the 'hm' method and alpha-levels of .05 and .01 [7] and recorded speaker-specific upper control limit (UCL) violations. T^2 UCL violations were then compared to the presence or absence of a cold in order to see

whether instances of colds and similar illnesses would affect the acoustic profile of individuals.

Performance of the procedure was evaluated by analyzing sensitivity and specificity at group and individual level.

Table 1: Speaker age range (Age), gender (Gen), smartphone type (Phone), number of recordings (Rec) and instances of cold (Cold).

Nr	Age	Gen	Phone	Rec	Cold
1	25-29	M	Samsung Galaxy S6 Edge+	33	2
2	60-64	F	iPhone 5s	66	11
3	45-49	M	iPhone 6s	34	4
4	45-49	M	HTC One (M8) & Samsung Galaxy S7 Edge	22	3
5	25-29	F	Galaxy S6	21	3
6	35-39	F	iPhone 5c & iPhone 6s	24	2
7	35-39	F	HTC One	15	2
8	25-29	M	iPhone 6	10	2
	Sum			225	29

III. RESULTS

Table 2 shows the contingency tables for presence of a cold and T^2 UCL violations across all speakers for p-values of .05 and .01. Fisher's exact test showed a significant association between cold state and UCL violations for p=.05 (p=.007) and p=.01 (p=.001). Hit rate/sensitivity for p=.05 was 62%, specificity 66%, for p=.01 sensitivity was 55%, specificity 76%.

Results for individuals show large differences in performance of the procedure. Table 3 shows individual values for sensitivity and specificity. We investigated whether individual sensitivity and specificity values were connected to the number of recordings per participant. Figure 1 shows both sensitivity and specificity as a function of the number of submitted recordings. The graph shows that specificity increases with sample size, suggesting that false alarms become rarer when speakers provide more data. Acceptable specificity values are reached for both approaches (p=.01 and p=.05) with a sample size around 30.

Table 2: Contingency table for 'hm' method and p-levels of .05 and .01

	No cold		*Cold*		*Sum*	
	.05	.01	.05	.01	.05	.01
Below UCL	129	149	11	13	140	162
Above UCL	67	47	18	16	85	63
Sum	196		29		225	

Table 3: Sensitivity and specificity per speaker for each p-level

Speaker	*Sensitivity*		*Specificity*	
	.05	*.01*	*.05*	*.01*
2	0.5	0.5	0.8	0.8
6	0.6	0.5	0.9	0.9
9	0.3	0.0	0.9	0.9
18	1.0	1.0	0.5	0.5
40	1.0	1.0	0.5	0.5
43	1.0	1.0	0.9	0.9
61	0.0	0.0	0.5	0.5
67	0.5	0.5	0.1	0.1

The pattern for sensitivity does not show a clear relationship with sample size but there is a tendency for the p=.05 method outperforming the p=.01 method with higher sample sizes.

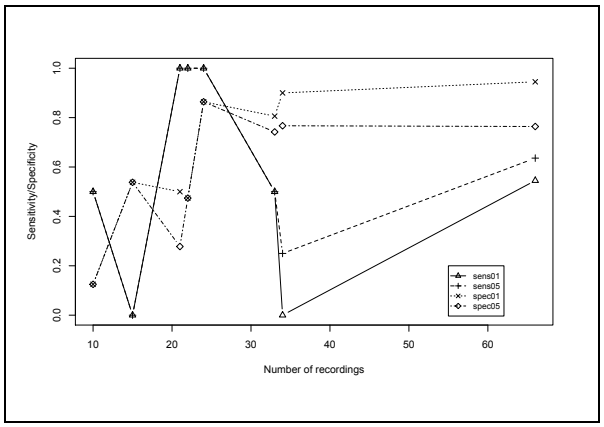


Figure 1: Changes in sensitivity and specificity of cold detection with number of recordings (sens01 – sensitivity with alpha level .01 etc).

IV. DISCUSSION

Our first analyses indicate that longitudinal monitoring of voice recordings via smartphones has potential for providing important information about the state of a voice. The current setup still generates too many misses and false alarms for unsupervised monitoring, but could be useful for supervised monitoring with voice expert support.

We have so far not excluded any recordings based on background noise levels, and we have not yet considered field effects like background noise and room size. Incorporation of these variables is likely to decrease false alarm rates in the future.

Another important future aim will be increasing the sensitivity of the method. The current acoustic parameters have not yet been analysed for their individual contributions to outlier patterns, and exclusion or addition of parameters, alongside alternative analytical approaches (e.g. machine learning) could improve hit rate.

Besides further development of the database and incorporating speakers with frequent voice problems, future research will focus on increased calibration of the method, e.g. by developing normative thresholds for acoustic parameters collected with various types of smartphones and quantifying the effects of various potential confounds that can occur in the field, e.g. background noise and room size.

V. CONCLUSION

This study presented evidence that semi-regular monitoring of voices with smartphones has potential to provide important cues about the health state of a voice. This information could be used to trigger tailored advice provided by voice experts via remote channels and thus make an important contribution to the prevention of voice problems and disorders.

VI. ACKNOWLEDGEMENTS

This study was supported by a Research Incentive Grant from the Carnegie Trust for the Universities of Scotland (grant reference 70230). We would like to thank Matthias Eichner for app programming, Tess Whittaker for help with app testing and study design and all our participants for their efforts.

VII. REFERENCES

[1] R. Martins, E. Pereira, C. Hidalgo, and E. Tavares, “Voice Disorders in Teachers. A Review,” *Journal of Voice*, vol. 28, no. 6, pp. 716–724, 2014.

[2] L. Lehto, P. Alku, T. Bäckström, and E. Vilkman, “Voice symptoms of call-centre customer service advisers experienced during a work-day and effects of a short vocal training course,” *Logopedics Phoniatrics Vocology*, vol. 30, no. 1, pp. 14–27, 2009.

[3] L. Mathieson, *The Voice and Its Disorders*, vol. 6. London: Whurr Publishers, 2001.

[4] E. Lin, J. Hornibrook, and T. Ormond, “Evaluating iPhone recordings for acoustic voice assessment,” *Folia phoniatrica et logopaedica*, vol. 64, no. 3, pp. 122–130, 2012.

[5] C. Manfredi, J. Lebacq, G. Cantarella, and J. Schoentgen, “Smartphones offer new opportunities in clinical voice research,” *Journal of Voice*, vol. 31, no. 1, pp. 111–e1, 2017.

[6] E. O'Brien, A. Coats, P. Owens, and J. Petrie, “Use and interpretation of ambulatory blood pressure monitoring: recommendations of the British Hypertension Society,” *BMJ: British Medical Journal*, vol. 320, no. 7242, p. 1128, 2000.

[7] E. Santos-Fernández, *Multivariate statistical quality control using R*. Springer Science & Business Media, 2012.

[8] N. Roy, R. M. Merrill, S. D. Gray, and E. M. Smith, “Voice disorders in the general population: Prevalence, risk factors, and occupational impact,” *Laryngoscope*, vol. 115, no. 11, pp. 1988–1995, 2005.

[9] A. Brown and G. J. Docherty, “Phonetic variation in dysarthric speech as a function of sampling task,” *European Journal of Disorders of Communication*, vol. 30, no. 1, pp. 17–35, 1995.

[10] P. Boersma and D. Weenink, “Praat: doing phonetics by computer (version 6.0.21) [computer software].” 2016/09/25-2016.

[11] Y. Maryn and D. Weenink, “Objective Dysphonia Measures in the Program Praat: Smoothed Cepstral Peak Prominence and Acoustic Voice Quality Index,” *Journal of Voice*, vol. 29, no. 1, pp. 35–43, 2015.

[12] Y. Maryn, M. Bodt, and N. Roy, “The Acoustic Voice Quality Index: Toward improved treatment outcomes assessment in voice disorders,” *Journal of Communication Disorders*, vol. 43, no. 3, pp. 161–174, 2010.

[13] D. Michaelis, T. Gramss, and H. W. Strube, “Glottal-to-noise excitation ratio - a new measure for describing pathological voices,” *Acustica*, vol. 83, no. 4, pp. 700–706, Jul. 1997.

[14] J. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, P. Gómez-Vilda, M. Blanco-Velasco, and F. Cruz-Roldán, “The Effectiveness of the Glottal to Noise Excitation Ratio for the Screening of Voice Disorders,” *Journal of Voice*, vol. 24, no. 1, pp. 47–56, 2010.

[15] M. Iseli and A. Alwan, “An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation,” presented at the ICASSP 04, 2004, vol. 1, pp. 666–669.

MYORTHO – A VOCAL COACH APPLICATION WITH VISUAL FEEDBACK FOR MONITORING AND STORING OF PATIENT PROGRESS IN A HOME ENVIRONMENT

I. Verduyckt [1], P. Cardinal [2], A. Loubnani [2], A. Alpan [3]

[1]Université de Montréal, École d'orthophonie et d'audiologie, Montréal, Canada
[2]École de technologie supérieure de Montréal, Département de génie et logiciels et des TI, Montréal, Canada
[3]Université Libre de Bruxelles, BEAMS department; Brussels, Belgium
Ingrid.verduyckt@umontreal.ca; Patrick.Cardinal@etsmtl.ca; abdelali.loubnani.1@ens.etsmtl.ca; aalpan@ulb.ac.be

Abstract: **This study is the first step in a project aiming at improved treatment compliance to SLP therapy in the dysphonic population. A vocal coach application able to support the patient in his performance of the vocal exercises at home will be developed. Here, we describe the first developmental stage of the graphical interface of the vocal coach. This stage included the selection of clinically relevant exercises and the definition of pertinent criteria for evaluation of success as well as visually clear and supportive feedback modalities. The process and the results will be presented.**

Keywords : **Dysphonia, Treatment compliance, Visual feedback, Voice therapy, GIU**

I. INTRODUCTION

Behavioral management is one of the primary treatment options for voice disorders, either isolated or in combination to a surgical intervention. It involves physiological exercises that the patient performs regularly to achieve and maintain a healthy vocal behavior [1]. The exercises are taught to the patient by a speech language pathologist (SLP) during a clinic based session. They then have to be practiced daily at home between the clinical sessions in order for the therapy to be efficient. Although good outcomes are generally reported for behavioral voice therapy, patients face several obstacles and up to 65% of them are reported to drop out from therapy before completion [2]. One recurring factor patients identify as a barrier to comply with the treatment is the perceived difficulty in carrying out the vocal exercises. Replicating voice exercises at home without the support and feedback of the SLP is difficult and leads patients to shorten, cancel or simply forget about the exercises [3]. Increasing patients' motivation and patients' confidence in their ability to correctly execute the exercises are key factors for treatment adherence [3]. Research has shown that adherence to voice therapy can be improved by providing exercise examples on mobile devices [4], by giving visual feedback to the patient on how well the exercise is performed [5], and by monitoring compliance through audio-recordings of the patient's exercises at home [6]. Today there are, to our best knowledge, no clinical tools available that are combining these three adherence improving features.

Our objective is thus to develop a mobile application for home based voice training offering these features in the form of audio examples of the exercises, visual feedback of patient performance, and monitoring of exercise compliance by recording and storing of the patient's exercises.

The present study has three specific aims 1) the evidence-based selection of therapeutic exercises, 2) the determination of relevant success criteria and feedback modalities for each exercise, and 3) the development and testing of the graphical interface for visual feedback.

II. METHODS

1) *Evidence-based selection of vocal exercise program*:

Criteria for determining the exercise program were based on the one hand on discussions with the SLP of the team, who has over 10 years of clinical experience with voice patients, and on the other hand on the scientific literature where evidence concerning specific exercises' efficiency was sought for.

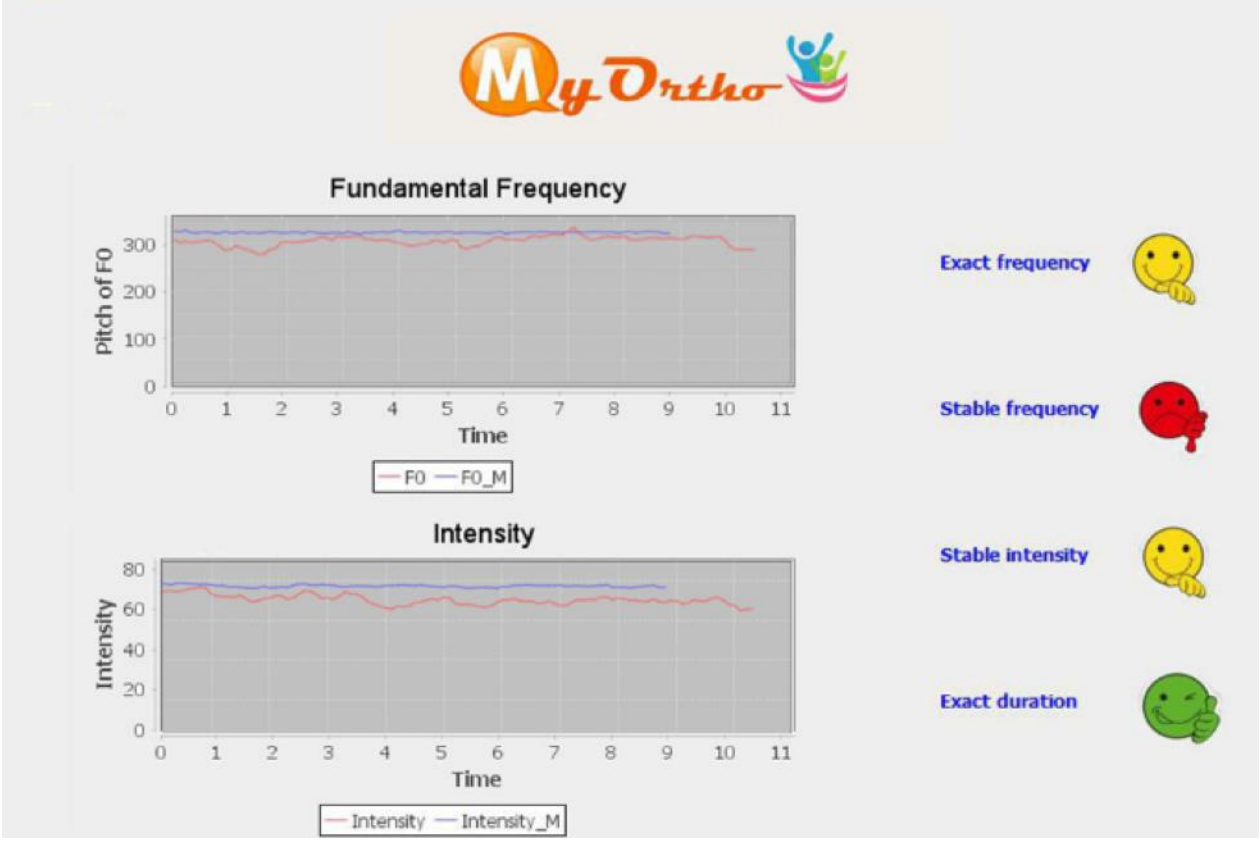
2) *Determining relevant success criteria and feedback modalities for each exercise*:

Relevant success criteria for the exercises and feedback modalities were defined in discussion with the SLP of the team and with regard to the specific aims that the chosen exercises were targeting.

3) *Design and testing of a graphical interface for the visual feedback of exercise efficiency*:

As the method for the design of the graphical interface is dependent on the results of step 1 and 2, this step will be reported on in the results section only.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)



III. RESULTS

1) *Evidence-based selection of vocal exercise program*:

The discussions with the SLP and the literature search lead to the choice of the therapeutic program Vocal Function Exercises (VFE) developed by J. Stemple. This program has a well referenced efficiency and is relatively easy to perform, it is also used by SLPs worldwide [7]. It comprises four exercises that have to be produced two times each at two occasions per day for several weeks. The exercises target specific respiratory, vibratory and resonance goals.

Exercise 1 targets maximum phonation time on a specific note in a soft, stable and resonant voice. Exercise 2 and 3 are stretching or gliding exercises targeting maximum frequency range by vocal glides from the lowest to the highest note and vice versa. Exercise 4 is an adductory strengthening exercise where five sequential notes have to be sustained for as long as possible in a resonant and stable voice quality.

2) *Determining relevant success criteria and feedback modalities for each exercise*:

Exercise 1 and 4 have common aims, namely: expanding phonation time, maintaining a stable pitch and a stable intensity, and vocalizing at a specific pitch. Common success criteria were thus developed for these exercises namely: a) Exercise length, b) Stability of pitch c) Stability of intensity and c) Pitch accuracy.

Exercise 2 and 3 also have common aims that are: expanding the frequency range and achieving smooth vocal glides without pitch breaks, relevant success criteria for these exercises were determined as: a) Continuity of pitch increase or decrease b) Pitch accuracy, and c) Magnitude of pitch range.

Based on the literature and on discussions with the SLP, two visual feedback modalities were envisioned: one that would display the patient's exercise *in real time* by plotting the patient's production in comparison to the sample exercise and one *delayed* visual feedback that would inform the patient of his level of performance on each of the success criteria after completion of the exercise. It was decided to maintain the two visual feedback modalities as simple and visually clear as possible in order to render the feedback easily understandable for a majority of patients, keeping in mind that patients' age and cognitive level can vary greatly in the dysphonic population.

3) *Design and testing of a graphical interface for the visual feedback of exercise efficiency*:

Real time visual feedback: Since VFE are predefined exercises that the patient has to match, we decided to plot both the model exercise and the patient exercise on the visual interface to encourage patients to match the model. The Praat software [8] was used to extract both the fundamental frequency and the intensity, and the curve analysis and the graphical interface were designed in Java. Figure 1 shows the user interface developed for the proposed application with the real time visual feedback displayed on the right.

The SLP of the team recorded the model exercises and she as well as three SLP students recorded "failed" exercises mimicking different types of failures to test the graphical visual feedback. The failures for the stable tone exercises 1 and 4 were: a) too short phonation time, b) on pitch but unstable voice, c) out of pitch. The failures for the glide exercises 2 and 3 were: a) shorter phonation time than the model, b) pitch break during the glide, c) narrower frequency range than the model.

Although phonation time is not a criteria for exercise 2 and3, this kind of "failure" was important to include precisely because it should not affect success levels for these exercises.

In the example used to produce the screenshot, the aim for the patient was to be able to maintain a specific and stable pitch (around 300 Hz) with a stable intensity for 11 seconds. The interface shows the sample exercise in blue lines (what the patient should do) and a failed exercise in red lines (what the patient did).

Delayed visual feedback:

On the left of the screenshot, the results of the evaluation using criteria related to this exercise are shown using simple and meaningful pictures and colors. A green smiley indicates that the exercise has been done perfectly, a yellow one indicates that the patient has done well but should try continue to improve that particular criteria and a red one indicates a failed exercise.

In order to determine if the exercise has been correctly executed, we calculated the coefficient correlation between the expected pitch curve and intensity curve respectively with the ones obtained from the patient recording. In this work, two different types of correlation coefficients have been experimented.

The first one is the Pearson's correlation coefficient (PCC), which is defined as:

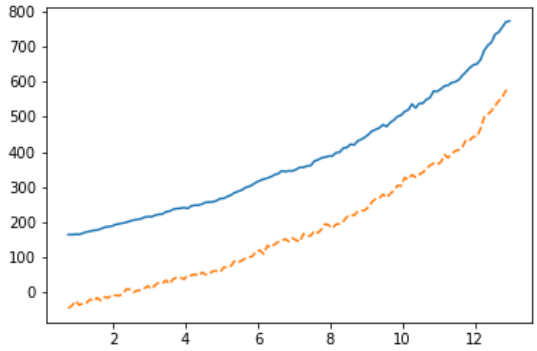
$$\rho_{M,Y} = \frac{cov(M,Y)}{\sigma_M \sigma_Y}$$

where $cov(M,Y)$ is the covariance between the model M and the patient result Y, σ_M is the variance of M and σ_Y is the variance of Y.

In figure 2, the use of the PCC is demonstrated with a glide exercise in which the patient has to increase F0 from 200Hz to 750Hz while maintaining a constant intensity. We can see that the patient has been able to produce the same curve but with an F0 shift of about 200Hz. This example is a case where the exercises has been executed correctly only in part, but not enough to be considered as a complete success.

However, the PCC ($\rho_{M,Y} = 0.99$) for this example is almost perfect. This is caused by the fact that the PCC doesn't take into account the position of the values, but only the variance. In this example, both curve shapes are almost identical, except that the one of the patient is 200Hz lower than the sample.

Figure 2: An example of an exercise in which the patient has to gradually increase his vocal pitch. The solid line represents the model produced by the therapist and the dashed line is produced by a patient.



In order to take the position of the curve into account, we used the concordance correlation coefficient (CCC). The CCC is defined as:

$$\rho_c = \frac{2\rho_{M,Y}\sigma_M\sigma_Y}{\sigma_M^2 + \sigma_{Y\,M}^2 + (\mu_M - \mu_Y)^2}$$

where $\rho_{M,Y}$ is the Pearson's correlation between M and Y, μ_M is the mean of M and μ_Y is the mean of Y.

In order to obtain a CCC of 1.0, both curves have to be identical also in terms of position. This is shown in the example of Figure 2 where the CCC is penalized by the shifting ($\rho_c = 0.59$). In this case, the CCC gives precious information on the level of success of the patient. However, both results are interesting. The CCC informs us that the curve is not the expected one but the PCC says that the shape is very similar to the model, which is also an important aspect of the exercise. The PCC and CCC scores are used to generate the scores defining the success levels illustrated by the smileys. The PCC scores are used to compute stability of pitch and intensity for exercise 1 and 4, and continuity of pitch and magnitude of pitch range in exercise 2 and 3. The CCC scores are used to compute pitch accuracy in all 4 exercises. Success of duration is based on a comparison between expected length and actual length of the

patient's performance in exercises 1 and 4. In the glide exercises where duration is not a success criterion, the curve will be normalized before the computation of both correlation coefficients. That will allow the system to make a fair comparison between the model and the patient's result.

IV. Discussion

The final objective of our project is to create a vocal coach application that will help patients sustain motivation for their voice treatment and support them in performing their vocal exercises. The overall aim being to improve therapeutic efficiency by improved compliance to the therapeutic program. The present study is the first step in the development of the virtual vocal coach and focused on three objectives, namely selecting the vocal exercises to be included in the vocal coach, determining relevant success criteria and feedback modalities, designing the graphical interface and test its capacity to accurately analyze the patient exercises and display corresponding visual feedback.

The VFE were chosen because they are well researched and numerous studies have attested of their efficiency for multiple vocal disorders. Further, they have been very precisely described and are available in audio-samples [9] which makes them easy for clinicians worldwide to include in their practice. However, VFE are not the panacea and SLPs are typically using other types of vocal exercises in their toolbox as well. A future development of our vocal coach could therefore be the possibility to feed it with other types of exercises.

The vocal objectives of VFE are multiple. The success criteria that we developed in this study do not encompass all vocal objectives of the VFE. For instance, we have not yet developed a criterion regarding the vocal quality in which the exercises are performed. This objective is of high importance in the therapeutic process of improving the voice of dysphonic patients and should be one of the future objectives of our project. Other criteria, such as quality of vocal onset and offset should be looked upon as well.

The graphical interface was developed in discussion with the team's SLP and inspired by the literature on visual feedback in voice therapy. It was designed to be easy to understand but still provide enough information to make sure that the user will know what and how to improve himself. However, achieving user friendliness will require the input from real users that will be testing the application in real life setting. Their input will be valuable to improve the graphical interface as to make it as user friendly as possible. Further studies with users will then be required to evaluate the value of the suggested feedback in supporting the patients in his daily exercise program.

The graphical interface in its current development state is a desktop application. It is likely that a mobile application will be required to maximize the usefulness of our vocal coach and integrate it smoothly to the modern lifestyles of our dysphonic patients. Development of the application on ios and android will be targeted in the future steps of our project. The visual feedback in the form of graphs are intended to be displayed in real-time, however, the current state of our software only allows delayed display. Real time display will have to be integrated into the graphical interface before testing with user groups will start.

V. Conclusion

Our study is the first step in the development of a vocal coach application aiming at increasing adherence of dysphonic patients to their voice training program. We concentrated on the selection of exercises, the determination of relevant success goals and feedback modalities for each of these goals, and the development of a graphical interface for visual feedback. The first prototype is a desktop application. However, mobile devices (ios and android) are targeted for the final version of the application. Future steps will be to develop our graphical interface to display the visual feedback of patient performance in real-time and to test its utility with a group of patients.

REFERENCES

[1] Vinney, L.A. and L.S. Turkstra, The role of self-regulation in voice therapy. J Voice, 2013. 27(3): p. 390 e1-390 e11.

[2] Portone-Maira, C., et al., Differences in temporal variables between voice therapy completers and dropouts. J Voice, 2011. 25(1): p. 62-6.

[3] van Leer, E. and N.P. Connor, Patient perceptions of voice therapy adherence. J Voice, 2010. 24(4): p. 458-69.

[4] van Leer, E. and N.P. Connor, Use of portable digital media players increases patient motivation and practice in voice therapy. J Voice, 2012. 26(4): p. 447-53.

[5] Zaki-Azat, J.N., The Influence of Real-Time Visual Feedback Training on Vocal Control. 2016.

[6] Ellis, L.W. and S.A. Beltyukova, Effects of compliance monitoring of vocal function exercises on voice outcome measures for normal voice. Percept Mot Skills, 2011. 112(3): p. 729-36.

[7] Barnett, S., The Advancement of Voice Therapy and the Contribution of Vocal Function exercises. Lewis Honors College Capstone Collection, 2017. 30.

[8] Boersma, P. and D. Weenink, Praat, software for speech analysis and synthesis. 2005.

[9] Stemple, J. Vocal Function Exercises, Plural Publishing, 2006.

PARTICIPATORY ENQUIRY FOR A BIONIC VOICE

M. Hagmüller[1], A. K. Fuchs[1], C. Bath [2]
[1]Signal Processing and Speech Communication Laboratory, Graz University of Technology
[2]Technische Universität Braunschweig, Germany
{hagmueller, anna.fuchs}@tugraz.at, c.bath@tu-braunschweig.de

Abstract: **People who have lost their larynx and thus speech functionality need a substitution voice to regain speech. Three main approaches exist, all of which have severe disadvantages. Previously, we have been working on improving the state-of-the-art for an electronic speaking aid. The current stage of our project has a special focus on a gender appropriate voice for laryngectomised speakers. To better understand the needs of the potential users of a bionic voice we adopted a participatory inquiry that involved interaction with 17 people without a larynx, of which 9 were female. All common substitution voices were used in the test sample. We spent between 1.5 and 6 hours with the individuals per session and had one to four visits. We learned that for all of them a natural voice is important. Most of the laryngectomees reject the use of a speaking aid, because of its bad sound. Women were specifically against the speaking aid. Desired properties of a bionic voice were an assertive voice, a voice matching ones personality. Women want to be recognized as female and have an attractive voice. They suffer from the low fundamental frequency of all substitution voices.**

Keywords: **alaryngeal speech, bionic voice, participatory enquiry**

I. INTRODUCTION

For people who suffer from laryngeal cancer or similar diseases, the last resort is a total laryngectomy, which results in a loss of speech. Currently there are around 25.000 people who have undergone a laryngectomy in Germany, around 10% of which are female.

After the larynx is removed surgically, the anatomy has changed dramatically, as depicted in Figure 1 (a) and (b). The trachea ends at the so called tracheostoma at the neck and the vocal tract is shortened. The vocal folds are missing, and thus the possibility to produce voiced speech.

There are three alternatives for people to regain their speech. (1) For esophageal voice air is gulped and then released in a controlled manner and the tissue of the pharyngo-esophageal segment in the pharynx vibrates. (2) A Tracheo-esophageal shunt valve is placed between the trachea and esophagus and therefore speech can be generated as above but with the air coming from the lungs (Fig. 1c). Although in Western Europe the tracheo-esophageal voice is the primary method of speech rehabilitation the situation is different in other countries and often it causes problems due to a leaking valve [1]. (3) The transcutaneous electronic speaking aid device (EL) is a small, hand-held and battery-driven device. The vibrating coupler disk of the device is held against the neck. The signal of the coupler disk is carried into the vocal tract. The EL is the focus of our research.

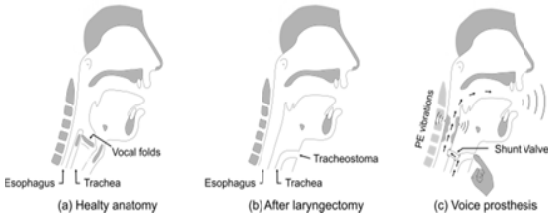


Figure 1: Anatomical details of (a) a healthy neck, (b) after laryngectomy, and (c) speaking with a voice prosthesis (fromv [2])

Major drawbacks of the resulting speech using the EL are the directly radiated noise of the device itself, the unnatural, monotonous quality of speech and the need of one hand to operate the device [3]. For the past years we have been working to improve the EL in order to increase the communication quality of the users. Regarding device operation, it is inconvenient to use one hand to operate the device. The main disadvantage is the inadequate quality of the resulting speech. The current technology of electronic speaking aids has been available for more than half a century [4] and there has been no major improvement of intelligibility and naturalness since then. An overview of the state-of-the-art and our research results has been published in [5]. In the scientific literature, we encounter two streams of approaches to improve the situation. (1) Technical approaches: the properties of EL speech and its differences to healthy speech are analysed (e.g. [6]); filtering techniques or similar

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

approaches to reduce the differences are applied (e.g. [7], [8], [9]). The resulting speech is evaluated objectively or subjectively with more or less appropriate listeners. (2) Researchers try to learn about the situation of affected people by sending out questionnaires, analyse and draw conclusions from the answers (e.g. [1]).

Female and male laryngectomees have different needs and requirements. Women are much less likely to be laryngectomised, therefore there is only limited research focused on women and much of the research done on men cannot be generalized to also include women [10], [11]. Much more data seems to be available on research about transgender women (e.g. [12]). The challenge to acquire a new voice seems to be similar to our topic of research. For transwoman there are clinical guidelines to support them to develop an appropriate female voice.

In order to overcome the shortcomings of existing approaches we planned to give the users a voice in the research process to reduce the bias that is unavoidable when only researchers make up their minds without incorporating potential users of their research. We wanted to address the specific problems women have to face, when they are forced to use a substitution voice.

There are several questions we wanted to explore together with people using a substitution voice. (1) What are the requirements of people who on a substitution voice concerning their verbal communication? (2) Do different user groups have different requirements? (3) What are specific situations that make it especially difficult to communicate with a substitution voice? (4) What is the reason so few people use an EL?

The rest of the paper is organized as follows. We first describe our methodology and the available subjects. In the results section we summarize the findings and the discussion section we reflect on the interactions with the users and we finally draw some conclusions.

II. METHODOS

We were guided by the methods of contextual design that are used for getting to know the work process of potential users of new software that should improve those processes [13].

A. Interview partners

We performed informal interviews with the potential users and most of the time, spent a longer period of time (1.5-6 hours) with them. Most of the interviews were with a single laryngectomised person, sometimes together with their partners. In addition we had two meetings with a group of people. We are aware of one important bias in our study, as we only had contact with socially active people, who were interested in the research. When possible they went back to their workplace, are involved in social life and have learned to cope with their new situation. Others withdraw themselves from social interaction and people from that group were not interested in an interaction with us. They might have different requirement than the active group, but we don't have a possibility to assess their needs with this methodology.

We originally planned to work with regular users of an EL. We took a lot of effort to find women who use an EL, but we were not able to find any woman that uses this as her primary substitution voice. Therefore we included users of any substitution voice. For an overview on gender and means of communication see Table 1. Our small statistic reflects qualitatively what is reported in literature on the distribution of substitution voices. We only have a high proportion of EL users because we were specifically looking for them. One woman was communicating with pen and paper only. The person who whispers had only the vocal cords removed.

We complemented this first-hand information with a discussion with the team of phoniatricians and speech and language therapists (SLPs) at the phoniatric department at the ENT university clinic Graz.

We organized the interaction in several meetings that were structured as follows:

Table 1: Distribution of gender and means of communication: EL ... electronic speaking aid, ES ... esophageal voice, TE ... tracheo-esophageal voice, PN ... Pen, WH ... whisper

	EL	**ES**	**TE**	**PN**	**WH**	**Total**
Male	4	1	3	0	0	**8**
individual visit	4	1	0	0	0	5
group talk	0	0	3	0	0	3
Female	0	3	4	1	1	**9**
individual visit	0	2	2	1	1	6
group interview	0	1	2	0	0	3
Total	**4**	**4**	**7**	**1**	**1**	**17**

B. Structure of interaction

1) The first visit was aimed at getting to know the person and introducing ourselves. We emphasized that we visited them because they are the experts concerning their voice and we wanted to better understand their specific needs and problems. We then suggested spending up to half a day with them to get to know them better. We also ask for a specific scenario that is a challenge for their communication abilities and whether we could be take part in it and observe

them. 2) For the second visit we observed a challenging communication scenario, e.g. pub or shopping. We observed the interaction with other people and the challenges that arose because of the specific situation. 3) For the third visit we continued from the second session in a different situation and then presented our bionic voice test system.

C. Bionic Voice System

Our bionic voice test system is an improved version compared to what we presented in [5]. We use a small transducer that is attached to the neck with a neck-collar above the tracheo-stoma. The transducer is driven by a headphone amplifier that gets the signal from a notebook. We also use a head-set microphone to pick up the speech sound and use this information to calculate an F0 contour. The Matlab based system allows modifying the voice quality by means of changing the parameters of the LF-model, which is used to generate the excitation signal. The users get a wireless button to turn the signal on and off.

At the current stage, we have gone through the whole cycle with the four subjects using the EL. For the non EL users we did only complete the interviews and with some we tried to do the hands-on experiment with our bionic voice test system. We realized without a sufficient proficiency regarding speaking with an EL, the experiment didn't make much sense.

III. RESULTS

A. Interviews

a) The learnings can be summarized in three categories.

1) **Specific problems female speakers have when using a substitution voice**: Even though losing the voice is a traumatic experience for everyone, female speaker especially suffer from the quality of the substitution voices. The low pitch frequently leads to being identified as a male, which is especially critical when using telephony based services that require some form of identification. This has an impact on the feeling of self-worth and the question of attractiveness as a woman.

2) **Insights why we weren't able to find a female of an EL**. The robotic and monotonous sound of the electronic speaking device seems specifically repelling for women. A frequent comment of the female subjects on why they didn't want to use an EL was that they would rather communicate by writing than having such a strange voice.

3) **Requirements for an electronic speaking aid.** The most important shortcoming of all substitution voices seems to be the reduce loudness of their voice, that results in not being able to take part in conversations in acoustically difficult settings. Examples we witnessed were settings such as in a restaurant, a shopping centre, an intercom at a barrier. b) Hands-free operation of the EL is another important requirement. Currently, conversation is very limited when doing something where both hands are needed, such as driving a car, cooking, or eating. People using the EL with the right hand have to change the device e.g. when shaking hands. c) Battery life. When talking a lot than the batteries drain a lot. We witnessed the use of up to four packs of battery for a period of half a day. d) The conversation over the telephone is a problem for all. We often hear that they only actively call but don't pick up the phone if they can avoid it. They report people hang up the phone when they hear the substitution voice. For EL users this seems particularly relevant. A more natural voice would reduce such situations.

B. Testing the Bionic Voice System

When testing our Bionic Voice System we got valuable feedback. All mentioned it was not loud enough and therefore could not solve one of their most important requirements. While the neck-collar was well received by some for others it seems not to be a good solution. Almost every neck was different, often due to additional problems, such as a neck dissection. A custom fit coupler disc would be necessary in some cases. One woman had issues with the pharyngeal reflex, so the collar was not an option for her.

The hands-free option, though it was not implemented in a way that would work in everyday life was confirmed as a very important feature.

The varying fundamental frequency was disturbing at first for all subjects. While some started to prefer it over the static pitch, some were not getting used to it.

In addition to voice related learning, we also learned methodological lessons. The first issue that we had to reflect was what impression our laboratory setup would leave on the users. A very complex setup with lots of cables and unfamiliar electronics might be intimidating and could create an unnecessary barrier between the scientists and the users.

IV. DISCUSSION

Once the volume of the voice was satisfied also the male speakers were concerned how they sounded. Women explicitly expressed that they were much more concerned how they sounded. We learned that woman have difficulties to accept the new voice because it doesn't sound feminine at all [11]. Some women decide to rather not speak at all than sounding like a male. In a study with 218 larygectomees (on average 6

years after surgery), 17% remain voiceless and 40% withdrew socially [14].

One older user explicitly mentioned that he didn't like technology so much. We therefore tried to reduce the visible technical complexity, while being open for those interested in the technology to explain what is going on behind the scenes. On the other hand, for younger subjects state-of-the-art technology was important, such as a connection to the smart phone.

We found it helpful to record the meetings with an audio recorder and not to rely to collect interview notes from memory in order to make the description as less subjective as possible. Since most of the times very personal issues came up, we also felt it not being appropriate when one of us was taking written notes during the conversation. Of course audio recordings were authorized by the users.

V. CONCLUSIONS

The interviews and the test of our bionic voice system showed, that there is a great need for an improved way of speaking for people without a larynx. Especially women are in need of a voice that is in line with their gender. The main problem of the current bionic

VI. ACKNOWLEDGEMENTS

This work was funded by the FEMTech project #849824 administered by the Austrian Research Promotion Agency (FFG) and the HEIMOMED Heinze GmbH & Co. KG

REFERENCES

[1] S. R. Cox and P. C. Doyle, "The Influence of Electrolarynx Use on Postlaryngectomy Voice-Related Quality of Life," *Otolaryngology -- Head and Neck Surgery (OTO-HNS),* vol. 150, pp. 1005-1009, 2 2014.

[2] J. Lohscheller, "Dynamics of the Laryngectomee Substitute Voice Production," 2003.

[3] G. S. Meltzner and R. E. Hillman, "Impact of Aberrant Acoustic Properties on the Perception of Sound Quality in Electrolarynx Speech," *Journal of Speech, Language, and Hearing Research,* vol. 48, pp. 766-779, 8 2005.

[4] J. L. Flanagan, "Artificial larynx". US Patent 3,291,912, 1966.

[5] A. Fuchs, M. Hagmüller and G. Kubin, "The New Bionic Electro-Larynx Speech System," *IEEE Journal of Selected Topics in Signal Processing,* vol. 10, pp. 952-961, 2016.

[6] M. S. Weiss, G. H. Yeni-Komshian and J. M. Heinz, "Acoustical and perceptual characteristics of speech produced with an electronic artificial larynx," *The Journal of the Acoustical Society of America,* vol. 65, pp. 1298-1308, 1979.

[7] E. A. Goldstein, J. T. Heaton, J. B. Kobler, G. B. Stanley and R. E. Hillman, "Design and implementation of a hands-free electrolarynx device controlled by neck strap muscle electromyographic activity," *IEEE Transactions on Biomedical Engineering,* vol. 51, pp. 325-332, 2004.

[8] K. Nakamura, "Speaking-Aid Systems using Statistical Voice Conversion for Electrolaryngeal Speech," Ph.D. dissertation, Nara Institute of Science and Technology, 2010.

[9] H. R. Sharifzadeh, "Reconstruction of Natural Sounding Speech from Whispers," Ph.D. dissertation, Nanyang Technological University, Singapore, School of Computer Engineering, 2011.

[10] W. H. Gardner, "Adjustment problems of laryngectomized women," *Archives of Otolaryngology,* vol. 83, pp. 31-42, 1966.

[11] S. R. Cox, J. A. Theurer, S. J. Spaulding and P. C. Doyle, "The multidimensional impact of total laryngectomy on women," *Journal of Communication Disorders,* vol. 56, pp. 59-75, 2015.

[12] S. Davies, V. G. Papp and C. Antoni, "Voice and Communication Change for Gender Nonconforming Individuals: Giving Voice to the Person Inside," *International Journal of Transgenderism,* vol. 16, pp. 117-159, 7 2015.

[13] H. Beyer and K. Holtzblatt, Contextual Design: Defining Customer-centered Systems, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998.

[14] H. Danker, D. Wollbrück, S. Singer, M. Fuchs, E. Brähler and A. Meyer, "Social withdrawal after laryngectomy," *European Archives of Oto-Rhino-Laryngology,* vol. 267, pp. 593-600, 9 2009.

SESSION III:

VOICE AND NEUROCOGNITION

BABIES' VOICES: A COLLABORATIVE RESEARCH PROGRAM ON THE AUTOMATED ACOUSTICAL ANALYSIS OF THE PRETERM NEWBORN CRY

R. Viellevoye[1], D. Melino[2], S. Orlandi[2], G. Pieraccini[2], G. Donzelli[3], A.Torres-García[4], C.A. Reyes García[4],C. Manfredi [2]

[1]Department of Pediatrics, Neonatal Intensive Care Unit, University of Liege, Liege, Belgium
[2]Department of Information Engineering, Università degli Studi di Firenze, Firenze, Italy
[3]Department of fetal and neonatal medicine, Università degli Studi di Firenze, Firenze, Italy
[4]Computer Science Department, Instituto Nacional de Astrofisica, Optica y Electronica (INAOE), Puebla, Mexico
renaud.viellevoye@chrcitadelle.be;donatellamelino@gmail.com;orlandisilvia85@gmail.com;
gianandrea.pieraccini@stud.unifi.it;alejandro.torres@ccc.inaoep.mx;kargaxxi@inaoep.mx;
claudia.manfredi@unifi.it

Abstract: **The aim of this collaborative work is to provide the automated assessment of the melodic shape of the newborn cry with the BioVoice software tool. The method was tested on synthetic signals with 100% matching. Acoustical parameters of cries obtained from preterm and term newborns in Liege (Belgium) and Firenze (Italy) were estimated with BioVoice. The automated classification was first compared to the perceptual (visual) analysis considered as the gold standard on a set of healthy at term newborns with a matching up to 85%. Then, significant differences were found between at term and preterm babies up to 85%. Our study suggests that some melodic characteristics of the newborn cry could be detected to predict the belonging to term/preterm group of patients with an acceptable accuracy.**

Keywords: **Newborn, cry, preterm newborn, cry melody, automated analysis.**

I. INTRODUCTION

The acoustical analysis of infant crying is a promising non-intrusive and cheap approach as an aid to early diagnosis of neurological disorders [1]. The most relevant clinical parameter is the fundamental frequency f0, which reflects the regularity of the vibration of the vocal folds of the newborn. To date, the analysis of the infant crying melody, that is the temporal trend of f0 over time, is carried out by the paediatrician/neurologist with a perceptive examination based on listening to the cry and visually inspecting the fundamental frequency f0 shape. This approach is not widespread as the procedure is operator-dependent and requires a considerable amount of time often prohibitive in daily clinical practice [2]. The aim of our collaborative work is to provide a fast and fully automated method for assessing the melody shape of the newborn cry that could be used routinely to assess at risk newborns such as preterm infants.

Indeed, preterm newborns are at high risk for developing cerebral palsy, cognitive impairment, behavioural difficulties and/or neurosensory disabilities [3]. Early diagnosis of neurological impairment is crucial to initiate neuromodulatory interventions supporting cerebral plasticity, while delayed recognition increases the risk for co-morbidities and poor outcome. Systematic automated analysis of the newborn cry performed at term-equivalent age could thus help clinicians to identify particular pattern of cry that could be predictive of poor neurological outcome.

Furthermore, cry is a developmental process influenced by acoustical environment and stimulations. Premature birth causes a sudden transition from the physiological intrauterine environment towards the noisy world of the neonatal intensive care unit (NICU), depriving the baby of the biological maternal voice. The cry, as the first way of communication experienced by the newborn to elicit caretaking, could be negatively influenced by this modified environment. Developmental care is a broad category of interventions designed to minimize the stress of the NICU environment [3, 4]. These interventions may include elements such as control of external auditory stimuli and facilitation of parental involvement. Again, a systematic analysis of the newborn cry could identify adequate strategies that could minimize the impact of the postnatal environment on neurodevelopment, in particular speech and language acquisition.

II. METHODS

BioVoice is a multi-purpose voice analysis tool developed under Matlab® at the Biomedical Engineering Lab., Department of Information

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

Engineering, Università degli Stud di Firenze [5-8]. Newborn infant cry recordings, that may last even several minutes, are made up by a number of "cry units" (CUs), commonly of different length and separated by "silence" frames. A CU is defined here as a high-energy frame lasting >260ms. The purpose is to automatically perform the classification of the cry melody of each CU detected within a recording. Detection of CUs is performed using a robust Voiced/Unvoiced (V/UV) detection procedure with variable energy thresholds that avoids incorrect splitting of a single event into several intervals [8,9].

BioVoice performs the estimation of several acoustic parameters that gained great scientific interest in the last years, including the fundamental frequency f0 and the first three resonance frequencies of the vocal tract, along with their std. To successfully assess the CU melodic shape, the first step is the accurate removal of the f0 outliers, that are irregularly distributed within the CU and that could distort the shape identification. This is performed through several steps each one based on specific conditions. Afterwards, BioVoice allows the classification among 5 basic melodic shapes: Plateau (P), Rising (R), Falling (F), Symmetric (S) and Complex (C). To perform the classification each CU is subdivided into 12 equally spaced time segments each one described by its f0 mean value plus the first f0 value (13 Perc). To find the best number of segments, the same procedure is applied with 20 equally spaced points plus the first f0 value (21 Perc).

First, the method was tested on synthetic signals. To this aim a new synthesizer was developed made up of a pulse train generator and a vocal tract filter. To get variable frequency control vectors capable of synthesizing the 5 basic melody shapes, a spline interpolation was implemented with a time varying f0[n]. Settings allowed obtaining melodic shapes close to newborn cry, within the range 400 Hz - 650 Hz, and f0 mean between 500 Hz and 550 Hz. To test the proposed method synthetic white noise of increasing amplitude (0%, 1%, 5% and 10% of the signal maximum amplitude) was added to the synthesized signals [10].

The automated classification has then been applied to cry recordings coming from at term healthy newborns. Recordings were performed with a Shure SM58 microphone and a Tascam US144MK2 sound card. The microphone was kept at fixed distance (15 cm) from the newborn's mouth and cry signals were recorded at the awakening of the baby before feeding, thus they were supposed to be feeding cries. Results have been compared to the perceptual analysis (considered as the gold standard) performed by trained raters. A melodic shape was classified as belonging to one of the five categories (P, F, R, S and C) if at least two out of the three raters agreed with the same shape. If the three raters disagreed, the CU was defined as "borderline" and temporarily eliminated.

Once the method was validated, it was applied to cry episodes recorded from both preterm and at term newborns in Liege, Belgium (Neonatal Intensive Care Unit, CHR Liege) and Firenze, Italy (Neonatal Intensive Care Unit, Meyer Children Hospital and Neonatal Unit, Azienda Sanitaria Ospedale San Giovanni di Dio) at term-equivalent age. Percentage of the different melodic shapes and several acoustic parameters such as mean duration of CUs, mean fundamental frequency f0 and standard deviation were calculated by automated analysis using BioVoice. Perceptual analysis was performed by trained raters. Results were compared between Belgian and Italian newborns as well as between at term and preterm babies among the entire and local populations. Statistical analysis was performed using a t-test and statistical significance was considered for p-value <0.05.

Finally, a classification technique based on the 10-fold cross-validation method was applied to the set of estimated melodic shapes in order to obtain the automated classification of a cry episode as belonging to term vs preterm infant [11].

III. RESULTS

A. Synthetic data

Both 13 and 21 Perc were tested. For all the synthetic melodic shapes and all the levels of added noise the fitting procedure gave the best results with the 4th order polynomial for which the R-square parameter R^2 (ranging between 0 and 1 where 1 is the best fitting):

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \qquad (1)$$

was found as: $R^2 \cong 0.993$ for 13 Perc and $R^2 \cong 0.995$ for 21 Perc. Both perceptual and automated melodic classifications were successful at 100%: all melody shapes were correctly classified [10, 11].

B. Real data

CUs from 6 healthy at term newborns (3 male and 3 female) were recorded in the maternity unit in CHR Liège (Belgium). Each cry episode lasts 1-2 minutes and consists of several CUs. A total of 466 CUs was collected. 48 CUs were excluded because of the absence of consensus between the three raters. Moreover, other 116 CUs were not perceptually recognized by any rater as belonging to one of the five basic shape considered here and were also excluded from analysis.

Considering all the 5 shapes, the automated analysis with 21 percentiles matched the perceptual one in 89.5% of cases while only in 80.3% with 13 percentiles. Excluding the C shape, the match increases to 96.7% with 21 percentiles and to 89.3% with 13 percentiles. Therefore, the best number of Perc proved to be equal to 21. We point out that these results concern the case of full agreement among the three raters. Table I summarizes the results.

TABLE I – Automated vs perceptual classification of the melodic shapes. Results for the full agreement among the three raters are presented. Best result: 96.7% match with 21 Perc and the 4 basic shapes P, R, F and S (C excluded).

3/3 raters	P, F, R, S, C	P, R, F, S
Automated – 13 Perc	80.3%	89.3%
Automated – 21 Perc	89.5%	**96.7%**

C. Newborn cry melody in term and preterm newborn

A larger data set was analyzed consisting of a total of 9 preterm infants (21 cry episodes and 382 CUs) and 24 term infants (41 cry episodes and 2532 CUs) delivered from Belgian mothers in the maternity and neonatal intensive care units at CHR Liege. Moreover, 9 preterm infants (24 cry episodes and 1787 CUs) and 25 term infants (70 cry episodes and 5187 CUs) delivered from Italian mothers were recorded in the neonatal intensive care unit at Meyer Children Hospital and San Giovanni di Dio Hospital, Firenze, respectively. Melody of each CU was then classified as belonging to one of the five main categories (P, F, R, S and C) or to additional categories (LU – Low Up, UL – Up Low, FS – Frequency step, D – Double, U – Unstructured, NC – Not a cry or O – Other) using both automated and perceptual analysis. Details about shapes can be found in [10, 11].

Matching between automated and perceptual analysis was 58%, 65%, 59% and 61% for Belgian preterm, at term, Italian preterm and at term newborns respectively, which was lower than the results obtained in our preliminary study, even when limited to the five main basic shapes. This was mainly due to the low level of experience of the operator(s) that however increased during the testing phase: a deeper training brought to a better matching rate. Moreover, percentage of occurrence of each of these five categories was not statistically different between groups regarding to the method of analysis we used.

The mean value of f0 was not statistically different between term and preterm infant in Belgium (P =0.22), Italy (P =0.30), or from both countries (P =0.45). Mean CU duration was significantly shorter for term newborns than for preterm infants (2,57s vs 5,04s respectively, $p<0.05$). We did not find any significant differences between groups regarding to the percentage of each of the five main categories considering both automated and perceptual analysis: Belgian vs Italian preterm infants (Pauto=0.25; Pperc=0.40) or term infants (Pauto=0.11; Pperc=0.50), term vs preterm infant in Belgium (Pauto=0.28; Pperc=0.46) and Italy (Pauto=0.44; Pperc=0.23), term vs preterm infants from both countries (Pauto=0.45; Pperc=0.24) or Belgian vs Italian newborns (Pauto=0.10; Pperc=0.45). However, preterm newborns seemed to have a trend in favour of more C and NC shapes as compared to term infants who more frequently present the P pattern, while other categories show similar frequencies in each population. Table II summarizes these results.

TABLE II – Percentage of C, NC and P shapes in Belgian, Italian and overall at term and preterm newborns.

Shape	Population	Preterm	Term
Complex	Liege	32.5%	23.7%
	Firenze	29.8%	25.4%
	Total	**33%**	24.9%
Not a cry	Liege	7.9%	6.5%
	Firenze	9%	4.6%
	Total	**8.8%**	5.3%
Plateau	Liege	11.3%	28.7%
	Firenze	17.8%	28.3%
	Total	16.6%	**28.4%**

Finally, results obtained with automated classification of cry episodes using the 10-folds cross-validation method were encouraging. The method, applied to the whole set of melodic shapes, was able to discriminate between preterm or at term infants with an accuracy ranging from 74.47 to 85.48%. (Table III).

Table III – Accuracy of the automated classification to discriminate between at term and preterm newborns

PRETERM vs AT TERM	Correct	Not correct
Liege	**85,48%**	14.52%
Firenze	74,47%	25.53%
Liege + Firenze	80.77%	19.23%

IV. Discussion

This work presents a methodological approach to the classification of the newborn cry melody whose features are considered clinically relevant for the assessment of the neurological status of the newborn at birth. The method has the advantage of being totally contactless and thus applicable also to very delicate subjects such as newborn babies are. The required equipment is low cost and easy to use and therefore easily implementable in any paediatric clinic both public and private.

Results obtained with the BioVoice software for the automated classification of newborn cry compared to the perceptual analysis are encouraging. We assumed as gold standard the blinded perceptual (visual) analysis made by a panel of trained raters. The automated analysis with 21 percentiles better matched the perceptual one than the automated analysis with 13 percentiles. However, imperfect matching still remains, especially if more than the five basic categories of melodic shapes are considered. We point out that the inter-observer reliability was not perfect in our study, maybe due to different levels of experience between raters. For a more reliable gold standard reference, future developments should consider a larger and more experienced group of evaluators.

Comparison of the main acoustical characteristics and the pattern of melodic shapes between at term and preterm newborns did not reach statistical significance, except for the mean duration of CU that was found shorter in term newborns. However, our study shows some trends between at term and preterm babies in the percentage of some categories of melodic shapes.

Finally, the results obtained with the automatic classification suggest that some characteristics of the newborn cry could be selected to predict the belonging to a defined group of patients with an acceptable accuracy. Systematic recording of cry from preterm newborn at term-equivalent ages is currently under progress. This could make it possible to retrospectively characterize the acoustical parameters and melodic patterns of infants with typical development compared to infants with abnormal neurological development such as cerebral palsy, cognitive or speech delay in order to develop a predictive model based on the most relevant features.

V. Conclusion

This methodological work is a first step towards the establishment of procedures for the analysis of infant cry. The overall matching percentage between automated and perceptual analysis was found around 60%. This was mainly due to the low level of experience of the operator(s) that however increased during the testing phase: a deeper training brought to a better matching rate.

In future work the automated analysis will be improved by a refined control on the f0 shape variations and on the estimation of frequencies steps.

These results could be compared with the patients' follow-up, especially for preterm babies, in order to track their development and to study relationships between the automated/perceptual results and possible diseases in the central nervous system for such delicate patients. Therefore, the automated analysis of newborn cry melody could become a reliable support to the "time-consuming and subjective perceptual analysis and, if properly assessed, could even replace it and become part of clinical standards in the neonatal screening.

References

[1] Corwin MJ, Lester BM, Golub HL. The infant cry: what can it tell us, Curr. Probl. Pediatr. 1996 Oct;26(9):325-34.

[2] Larroque B, Ancel PY, Marret S, Marchand L, Andre M, Arnaud C, Pierrat V, Roze JC, Messer J, Thiriez G, Burquet A, Picaud JC, Breart G, Kaminski M; EPIPAGE Study group. Neurodevelopmental disabilities and special care of 5-year-old children born before 33 weeks of gestation (the EPIPAGE study): a longitudinal cohort study. Lancet 2008; 371(9615): 813-820.

[3] Marx V, Nagy E. Fetal behavioural response to maternal voice and touch. PlosOne 2015;10(6):e0129118

[4] Als H. Developmental care in the newborn intensive care unit. Curr Opin Pediatr. 1998 Apr;10(2):138-42

[5] Wermke, K., Mende, W., Manfredi, C., Bruscaglioni, P., "Developmental aspects of infant's cry melody and formants", (2002) Medical Engineering and Physics, 24 (7-8), pp. 501-514. doi: 10.1016/S1350-4533(02)00061-9

[6] Manfredi, C., Tocchioni, V., Bocchi, L., "A robust tool for newborn infant cry analysis", Conf Proc IEEE Eng Med Biol Soc 2006 ;1:509-12

[7] Manfredi, C., Bocchi, L., Orlandi, S., Spaccaterra, L., Donzelli, G.P., "High-resolution cry analysis in preterm newborn infants", (2009) Medical Engineering and Physics, 31 (5), pp. 528-532. doi: 10.1016/j.medengphy.2008.10.003

[8] Orlandi S, Manfredi C, Bocchi L, Scattoni ML, "Automatic newborn cry analysis: a non-invasive tool to help autism early diagnosis.", Conf Proc IEEE Eng Med Biol Soc. 2012;2012:2953-6. doi: 10.1109/EMBC.2012.6346583.

[9] Orlandi S, Dejonckere PH, Schoentgen J, Lebacq J, Rruqja N, Manfredi C. Effective pre-processing of long term noisy audio recordings: An aid to clinical

monitoring. Biomedical Signal Processing and Control 2013; 8: 799– 810
[10] Orlandi S, Bandini A, Fiaschi FF, Manfredi C. Testing software tools for newborn cry analysis using synthetic signals, Biomedical Signal Processing and Control 2017; 37: 16-22.
[11] Orlandi, S., Reyes Garcia, C.A., Bandini, A., Donzelli, G., Manfredi, C., "Application of Pattern Recognition Techniques to the Classification of Full-Term and Preterm Infant Cry", (2015) Journal of Voice, doi: 10.1016/j.jvoice.2015.08.007

RELATIONSHIPS BETWEEN NEWBORNS' CRY MELODY SHAPES AND NATIVE LANGUAGE

C.Manfredi[1], G.Pieraccini[1], R.Viellevoye[2], A.Torres-García[3] C.A. Reyes-García[3]
[1] Department of Information Engineering, Università degli Studi di Firenze, Firenze, Italy
[2]Department of Pediatrics, Neonatal Intensive Care Unit, University of Liege, Liege, Belgium
[3]Computer Science Department, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Puebla, Mexico
claudia.manfredi@unifi.it; gianandrea.pieraccini@stud.unifi.it; renaud.viellevoye@chrcitadelle.be; alejandro.torres@ccc.inaoep.mx; kargaxxi@inaoep.mx

Abstract: **Recent research studies have shown that since the last trimester of pregnancy the human fetus is able to listen to and possibly memorize auditory stimuli from the external world, both as music and language are concerned. In particular, they exhibit a specific sensitivity to prosodic features such as melody, intensity, and rhythm, that are essential for an infant to learn and develop the native language. This paper presents first results concerning the mother language assessment of a set of about 7.500 cry units coming from French, Arabic and Italian mother-tongue healthy at term newborns. A number of acoustical parameters and 12 different melodic shapes are detected with the BioVoice SW tool and their classification is performed with Random Forest and 4 neuro-fuzzy classifiers. Results show up to 94% differences among the three languages.**
Keywords: **newborn cry melody, mother language, automated acoustical analysis, classification algorithms.**

I. INTRODUCTION

During the last three months of pregnancy the human fetus is able to perceive sounds and distinguish the maternal voice. Adult-like processing of pitch intervals allows newborns to appreciate musical melodies as well as emotional and linguistic prosody and language [1]. Cry is the first means of communication for humans, and is the result of a developmental process influenced by the acoustical environment and stimulations, therefore some studies suggest that the newborn cry melody (the trend of the fundamental frequency with time) could be shaped by the maternal native language [2]. Prosodic features such as melody, intensity, and rhythm are in fact essential for an infant acquiring language dominion [3].
This paper presents first results concerning the mother language assessment of a large set of about 7.500 cry units coming from French, Arabic and Italian mother-tongue newborns. The acoustical parameters and the melodic shapes are detected with BioVoice [3-6] and their classification is performed with Random Forest and 4 neuro-fuzzy classifiers. Results show up to 94% difference among these languages, thus suggesting that newborns pick up acoustic elements of their parents' language before they are even born, and certainly before they start to babble themselves.

II. METHODS

The automated newborn cry analysis is performed with BioVoice, a multi-purpose voice analysis tool developed under Matlab® at the Biomedical Engineering Lab., Department of Information Engineering, Università degli Studi di Firenze [3-6]. Typically newborn infant cry recordings, that may last even several minutes, are made up of a number of "cry units" (CUs) of different length and separated by "silence" frames. A CU is defined as a high energy voiced frame lasting >260ms. With BioVoice the detection of CUs is performed using a robust Voiced/Unvoiced (V/UV) procedure that avoids incorrect splitting of a single event into several intervals [7]. On each CU BioVoice estimates several acoustic parameters, among which the fundamental frequency F0 and the first three resonance frequencies F1-F3, along with their maximum, minimum and standard deviation values, as well as other statistical parameters. It applies autoregressive (AR) parametric techniques, well suited to deal with quasi-stationary high-pitched signals as newborn cries are. After an accurate automated removal of outliers, the melodic assessment of the CU shapes is made through several steps, each one based on specific conditions. BioVoice allows both automated and perceptual classification of a CU among 12 basic melodic shapes: Plateau (P), Rising (R), Falling (F), Symmetric (S), Complex (C), Low-Up (LU), Up-Low (UL), Frequency Step (FS), double (D), Unstructured (U), Not a cry (NC), and Other (O). More details can be found in [8]. Some examples of the above mentioned melodic shapes are reported in Fig.1.

The very simple user interface implemented in BioVoice is shown in Fig.2. Several recordings can be

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

uploaded and analysed sequentially. The plot in the lower part of Fig.2 shows the result of the automated CUs selection (dotted line). Fig.3 shows the interface for the melodic assessment. As an example, a P shape is shown in the picture.

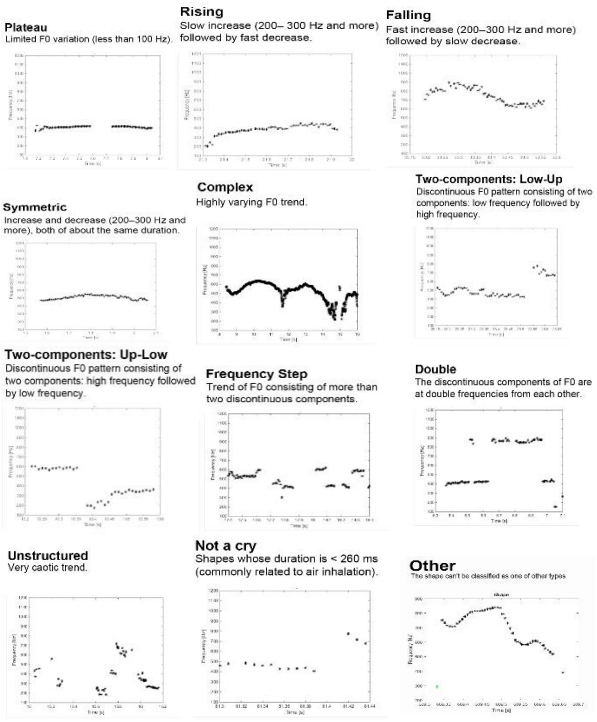


Figure 1 – Examples showing the 12 melodic shapes assessed with BioVoice. For each shape a short description is reported in the legend.

Once the acoustical parameters and the melodic shapes are detected and estimated, the automatic classification consists in detecting which classifier must infer a function (from training data) that allows recognizing new cases (test data) not used during the training process. Specifically, we assessed the Random Forest (RF) classifier, which is an ensemble of classification trees, and 4 neuro-fuzzy classifiers: Adaptive Neuro-Fuzzy with linguistic Hedges (ANFCLH) [9], Adaptive Neuro-Fuzzy with feature selection based on linguistic hedges (ANFCLH-FS) [10], Neuro Fuzzy Classifier (NFC) [11] and Speed-up Scaled Conjugated Gradient Neuro-Fuzzy Classifier (NFC-SCG) [12]. Neuro-Fuzzy models are hybrid systems that combine the capabilities of both representing the knowledge using linguistic expressions of fuzzy systems and learning of neural networks.

Specifically, NFC, NFC-SCG, ANFCLH and ANFCLH-FS create fuzzy rules using the K-means algorithm, whose input membership functions and output functions are later trained as a neural network. The main difference between NFC and NFC-SCG is that the second one implements an improvement for speeding up the scaled conjugate gradient algorithm used for training the NFC classifier. Whereas ANFCLH adds a layer of linguistic hedges for applying to each membership function. Finally, ANFCLH-FS takes advantages of the linguistic hedges for selecting a subset of features, which are used later for the classification stage using ANFCLH.

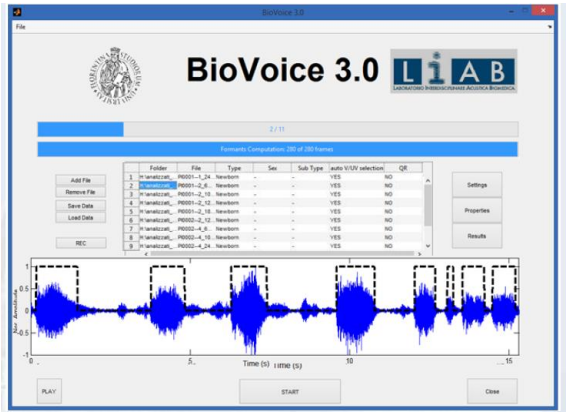


Figure 2 – BioVoice user interface

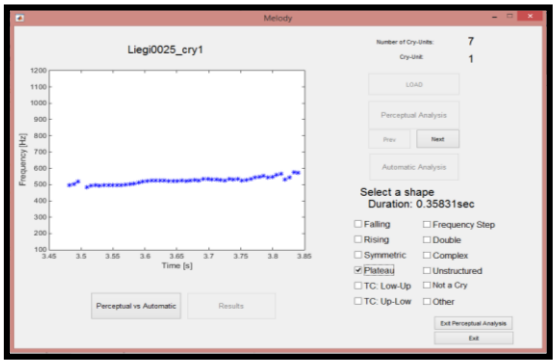


Figure 3 – BioVoice interface for the melodic assessment of each detected CU. A P-shape is shown.

III. RESULTS

For the experiments, we recorded infant cries from healthy at term babies whose mother's native languages are: Arabic, French, and Italian. Specifically, recordings come from two data sets: a set of 24 at term newborns (2532 CUs) from the Children Hospital «La Citadelle», Liege, Belgium (French and Arabic) and 28 at term newborns (5187 CUs) from the San Giovanni di Dio Hospital, Firenze, Italy (Italian). All recordings were made according to the same protocol: a Shure58 microphone was kept at a fixed distance of 20cm from the baby's mouth and connected to a laptop through a Tascam audio board. Recordings, lasting from few seconds to even 1 minute, were made 1-2 days after birth, in a non-noisy environment, before feeding:

therefore, all recorded signals should concern hunger cries.

BioVoice was applied first to raw data, i.e. the overall set of 52 recordings of variable length. Within each recording all the CUs were automatically detected and for each CU about 25 parameters were estimated. Also, the melodic shape of each CU was automatically assessed among the 12 listed in the previous section.

Looking for any relationship between CUs and mother's native language, we assessed the performance obtained from the automatic classifiers. All the infant cry's instances were characterized with the 12 qualitative features described above: falling, rising, symmetrical, plateau, complex, low-up, up-low, frequency step, double, unstructured, not a cry, and other. Also, we tested if adding two quantitative features (mean and standard deviation of the fundamental frequency F0) could help to improve the recognition of native language from infant CUs.

In a first instance, we assessed the performance of the automatic classifiers for the recognition of pairs of languages: Italian vs French, Arabic vs French and Arabic vs Italian. Table 1 shows the performances which are above the chance level for two classes (50%) and show that the use of mixed features is somewhat better than using qualitative features only (for the best performance's cases). Best results show up to 94% of correct classification (Arabic vs Italian).

Table 1. Accuracy percentages obtained for the classifiers using pairs of languages. Q - qualitative features. M - both qualitative and quantitative features.

Classifier	Italian/French		Arabic/French		Arabic/Italian	
	Q	M	Q	M	Q	M
RF	**90.52**	89.47	67.64	70.58	91.35	91.35
NFC	82.22	87.71	72.5	**77.5**	85.13	**93.75**
NFC-SCG	82.22	87.71	72.5	80	90.13	**93.75**
ANFCLH	82.22	85.51	72.5	**77.5**	88.88	87.5
ANFCLH-FS	85.33	88.82	69.16	72.5	88.75	88.88

Moreover, we evaluated which of these performances could be kept, or improved, when all languages are simultaneously classified. Table 2 shows the classification performances for all methods and using three types of features: qualitative (melodic shape), quantitative (acoustical parameters) and mixed. In this case, we observed that the best performances (nearly 84%) were obtained using mixed and qualitative features either with RF or NFC-SCG classifiers. Moreover, almost all performances were above the chance level (33.33% for the 3 classes).

Table 2. Accuracy percentages obtained for the classifiers using simultaneously all languages. Q - qualitative features; q - quantitative features; M - both qualitative and quantitative features. The low ANFCLH-FS q accuracy is obtained applying feature selection in which one feature from two variables is selected This means that one qualitative feature is not good enough for classifying among languages.

Classifier	All languages		
	Q	q	M
RF	82.85	60	**83.80**
NFC	74.18	72.45	76.40
NFC-SCG	73.27	67.72	81.31
ANFCLH	74	63	79.31
ANFCLH-FS	75.18	16	66.01

IV. Discussion

In this paper, a preliminary assessment of the relationship between the native language of the babies' mothers and qualitative features computed from infant cry recordings is presented. Results point out strong differences of newborn cry melody between Italian, Arabic and French mother tongues, even when all languages are simultaneously classified. Furthermore, the methods' performances are above the chance level for 3 classes, highlighting the performances obtained by RF (using qualitative and mixed features) and NFC-SCG (using mixed features). Thanks to the robust estimation and classification techniques the percentage of classification accuracy was found quite high: 90.52% between Italian and French and even higher (93.75%) between Italian and Arabic. The percentages vary according to the language and the classification method applied: the most robust methods are RF and NFC when the Italian language is compared to the other two, nevertheless the percentages are always above 67%.

We notice that the highest percentage was found between Italian and Arabic: maybe this could be related to the quite low number of guttural sounds that are found in Italian with respect to Arabic and also to French language.

V. CONCLUSION

In this paper, first results are presented concerning the automated classification of the newborn's cry melody. The melodic shapes as well as several acoustical parameters of the newborn cry are estimated with the BioVoice software tool, whose performance was tested with synthetic signals [13]. The classification is performed with several methods on a large data set, made up of more than 7.500 cry units, coming from French, Arabic and Italian mother language newborns, according to a specific protocol. Results show differences up to 94% thus suggesting that newborns pick up elements of their parents' language before they are even born, and certainly before they start to babble themselves.

Although the outcomes are promising, an extensive study should be further carried on for a better understanding. Also, improvements in the SW could help with a better assessment of the 12 melodic shapes as well as for the detection and classification of other shapes. Finally, recording a larger set of infant cries, especially the Arabic ones, would help to evaluate if the methods' performances could be kept.

REFERENCES

[1] Marx V, Nagy E. Fetal behavioural response to maternal voice and touch. PlosOne 2015;10(6): e0129118

[2] Mampe B, Friederici AD, Christophe A, Wermke K. Newborns' cry melody is shaped by their native language. Curr Biol. 2009 Dec 15;19(23):1994-7. doi: 10.1016/j.cub.2009.09.064. Epub 2009 Nov 5.

[3] Wermke K., Mende W., Manfredi C., Bruscaglioni P. Developmental aspects of infant's cry melody and formants, Medical Engineering and Physics,24,7-8, 2002

[4] Manfredi C., Bocchi L., Orlandi S., Spaccaterra L., Donzelli G.P., High-resolution cry analysis in preterm newborn infants, Medical Engineering and Physics,31, 5, 2009

[5] Manfredi C., Tocchioni V., Bocchi L., A robust tool for newborn infant cry analysis, Annual International Conference of the IEEE Engineering in Medicine and Biology – Proceedings, 2006

[6] Orlandi S., Manfredi C., Bocchi L., Scattoni M.L., Automatic newborn cry analysis: A Non-invasive tool to help autism early diagnosis, Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2012.

[7] Orlandi S., Dejonckere P.H., Schoentgen J., Lebacq J., Rruqja N., Manfredi C. Effective pre-processing of long term noisy audio recordings: An aid to clinical monitoring Biomedical Signal Processing and Control,8, 6, 2013.

[8] Orlandi S., Reyes Garcia C.A., Bandini A., Donzelli G.P., Manfredi C., Application of Pattern Recognition Techniques to the Classification of Full-Term and Preterm Infant Cry, Journal of Voice, 30, 6, 656-663, 2016, https://doi.org/10.1016/j.jvoice.2015.08.007.

[9] Sun CT, Jang JSR. A neuro-fuzzy classifier and its applications. Proc. of IEEE Int. Conf. on Fuzzy Systems, San Francisco 1:94–98. Int. Conf. on Fuzzy Systems, San Francisco 1993.

[10] B. Cetişli, A. Barkana. Speeding up the scaled conjugate gradient algorithm and its application in neuro-fuzzy classifier training. Soft Computing 14(4):365–378, 2010.

[11] B. Cetişli. Development of an adaptive neuro-fuzzy classifier using linguistic hedges: Part 1. Expert Systems with Applications, 37(8), pp. 6093-6101, 2010.

[12] B. Cetişli. The effect of linguistic hedges on feature selection: Part 2. Expert Systems with Applications, 37(8), pp 6102-6108, 2010.

[13] Orlandi S., Bandini A., Fiaschi F.F., Manfredi C., Testing software tools for newborn cry analysis using synthetic signals, Biomedical Signal Processing and Control, 37, 16-22, 2017, https://doi.org/10.1016/j.bspc.2016.12.012

FACIAL EXPRESSION RECOGNITION WITH FUZZY EXPLAINABLE MODELS

E. Morales-Vargas, C.A. Reyes-García, H. Peregrina-Barreto, F. Orihuela-Espina

División de ciencias computacionales, Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, México
{emoralesv, kargaxxi, hperegrina, f.orihuela-espina} @ccc.inaoep.mx

Abstract: **The performance of current algorithms of facial expressions recognition are still insufficient for certain applications such as facial rehabilitation. We aim at alleviating some current limitations of these algorithms by exploiting explainable fuzzy models over sequences of frontal face images. In this work, facial expressions are characterized in terms of action units. Fuzzy models maintain a semantic relation between the facial muscle appearance and the fuzzily associated facial expression. First, heuristic guided affine transformations align facial landmarks of the neutral and target expression. Second, features are extracted describing face movements in terms of changes in orientation (angle and magnitude) of distinctive facial areas. Third, the full featured representation is embedded into a compact one by means of pooling. Finally, a Sugeno-type adaptive neuro fuzzy inference system is used for each action unit to generate a description of the movements in the face that identifies the facial expression present in an image sequence. The proposed model discriminates facial expressions with mean accuracy of 89.04±0.91% with a maximum accuracy of 91.41±28%. Further, distinctly to current solutions the model can also describe why is reaching such decision. The current solution brings application in facial rehabilitation a step closer.**

Keywords: Facial expression recognition, fuzzy explainable models, facial action coding system.

II. INTRODUCTION

Facial expressions communicate emotions. Facial expression recognition (FER) is concerned with the automatic identification of the overt manifestation of affective states of a user by a computer. FER plays an important role in social communication [1], and has applications in security, human computer interaction, driver safety, psychology, neuroscience, education or sociology [2], [3] and health care such as detection of facial neuromuscular disorders [4], among others [5].

FER systems often proceed by extracting features from the input image set to feed a subsequent classifier that outputs the inferred facial expression [6], [7].State of the art algorithms in FER report maximum accuracies in the range of 90.51±0.64%. Current solutions have favoured discriminative over explicative power e.g. [5], [8]. Consequently, a general limitation of current developments are its explicative capacities. Explicative models go beyond predictive and discriminative models affording not only an output label but also accompany it with procedural mechanics. In FER, there are initial steps in this direction [1], [9], but admittedly, there is room for improvement.

Here, we question whether and how fuzzy explainable models can be tapped to afford both discriminative and explicative outcomes of facial expressions from frontal facial image sequences. We hypothesized that under controlled conditions (negligible camera rotations or illumination changes, absence of zooming operations in the image sequence and occlusions) fuzzy rules defined over action units (AU) can exhibit a high discriminate power between facial expressions whilst concomitantly explaining its actions. Each facial movement is called an AU and describes the smallest visually discriminable facial deformation. We propose a dynamic approach for FER based on an explainable fuzzy model of basic expressions (anger, contempt, disgust, fear, happiness, sadness and surprise). The movements of facial distinctive areas are used to encode an image sequence into AU. Facial expressions are modeled using an automated generation of fuzzy rules through subtractive clustering. Results suggest that our model can recognize facial expressions performing above state of the art performances, in addition to explaining why a facial expression has been labeled accordingly.

II. METHODS

The proposed method consists in four steps: (i) facial landmarks alignment, (ii) feature extraction, (iii) pooling, and (iv) fuzzy modeling. The input to the system are a sequence of frontal facial images of the same subject going from a certain neutral expression to a target expression. The method initiates by detecting and aligning facial landmarks. Affine transformations based on a novel heuristic guides the superimposition

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

alignment and ensuring invariance to scale, orientation and translation. Then, features describing the face movements are extracted from AU such as eyes or mouth. Features are concatenated into a vector of magnitudes and angle orientation of the detected movements of the facial landmarks with changes in size of facial areas. The raw representation is embedded to a more compact representation using average pooling. Finally, a set of Adaptive Neuro-Fuzzy Inference System (ANFIS) models describe facial movements of each sequence in terms of the AU.

A. Facial landmark alignment

Active Appearance Models (AAM) [14] retrieve 68 coordinates pairs from the face images describing the face shape, each one corresponding to a vertex of a face descriptors set. Procrustes Analysis (PA) superimposes shapes by optimally translating, rotating and uniformly scaling objects [5], [10]. PA mitigate geometric distortions in images or landmarks in terms of affine transformations. Classical PA is highly sensitive to noise and outlier values. To reduce such sensitivity, a heuristic based on eye canthus alignment is proposed. First, a rotation aligns the face along the x axis using the canthus of the eyes. Then, the facial landmarks are normalized to range [0, 1] for both neutral and facial expression for scale invariance. Later, to maintain the deformations caused by facial movement a correction in terms of the neutral state is performed. Finally, a translation to the neutral state minimizes the distance between the canthus of the eyes for both states superimposing both shapes. The full process is depicted in Fig. 1

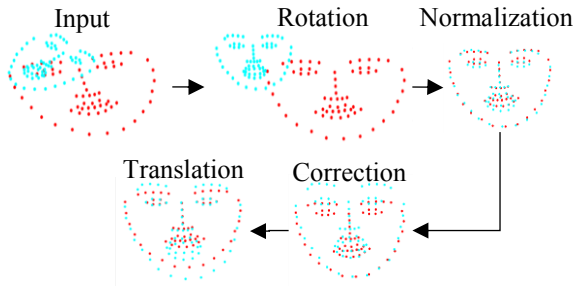


Fig. 1. Landmarks alignment using the novel heuristic.

B. Feature extraction

The performance of a predictive model is often dependent on the chosen representation for data [11].Choosing an appropriate representation is relevant to boost classification rates [9]. Feature extraction here consists of three substeps. First, the magnitude and orientation angle of each facial landmark between the final frame t_f and the initial frame t_0 of a i-th sample are computed and concatenated in the following tuple, $mo_i = [\mathrm{m}_1, \mathrm{o}_1, \mathrm{m}_2, \mathrm{o}_2, \ldots, \mathrm{m}_n, \mathrm{o}_n]$ where m_i and o_i are the magnitude and orientation of the movement of the i-th sample respectively, with n being the number of landmarks. Then, a triangulated shape of the facial landmarks is calculated forming a new vector of triangle areas $ac = [a_1, a_2, \ldots, a_m]$ with a_i being the change in area of a triangle, and m being the number of triangles. Finally, mo and ac are further concatenated to obtain a raw 243-dimensional feature vector $br = [mo, ac]$. The feature vector br is subsequently pooled to obtain a compact representation.

C. Pooling

Fuzzy models with a higher number of rules usually exhibit higher accuracies than one with a less number (a trivial consequence of increasing the model parameters), but in losing simplicity, they lose the ability to explain why the model is making a decision and are more prone to overfitting. We thus strive to reduce the dimensionality of the representation to generate a simpler model affording fewer rules. Given a chosen endpoint, many automatic strategies can search for optimal or suboptimal representations. However, given our interest in explicative models, we opted for a manual exploration of the data. Following this exploration, distinctive areas of the face were manually chosen; *Magnitude and orientation:* [Inner eyebrow, Outer eyebrow, eyelids, nose, upper lip, lower lip, right corner lip, left corner lip, jaw, Lips corners] *Areas:* [eyes, mouth]. Average pooling [12] then aggregates the selected local descriptors into a subset of the feature representation describing one facial distinctive area. The original br 243-dimensional feature representation is thus reduced to a 22-dimensional representation in which 20(=10x2) values are related to mo and 2 to ac.

D. Classification and explanation

The final stage of the model is the explaining classifier considering both the classification and explanation of the labelling. Knowledge is generated using granular fuzzy models in which the information is represented by hyperboxes. A hyperbox is a region of the decision space. For this task, a Takagi-Sugeno fuzzy inference system is used due to its extended flexibility in system design over the Mamdani fuzzy inference systems [13]. For each AU associated to a distinctive area, a Takagi-Sugeno model is generated using a rule generation algorithm [14] which consists in two steps: hyperboxes generation and rule generation. For the hyperbox generation subtractive clustering was used. Subtractive clustering depends on a γ parameter limiting the radius of influence and hence controlling the number of clusters. A Gaussian membership function was used for the hyperboxes.

In fuzzy system, fuzzy rules imply that vectors being evaluated are labelled not with a single value or class, but instead they are assigned a degree of membership to each class or label. One fuzzy rule is obtained for each hyperbox. The algorithm for fuzzy rule generation was modified here to assign a semantics to each rule facilitating interpretation. Such semantics assignment was made by partitioning the membership range [0,1]. For each subset of the partition, a semantics is assigned (very weak, weak, medium, strong and very strong presence). Then, the AU models are used to generate the descriptive fuzzy rules; i.e. a facial expression model. Here we have parameters γ_{au} and γ_{ex} for the AU and facial expression models respectively.

III. EXPERIMENS AND RESULTS

The Cohn Kanade Plus dataset (CK+) was obtained from [3]. It consists of 327 labelled image sequences from 123 healthy subjects in which one of seven facial expressions is represented. CK+ is labelled by expert judges according to the Facial Action Coding System (FACS) and AU [15]. Each sequence begins with a subject in a neutrally affective state (t_0) and ends with a facial expression (t_f) related to an affective state. Table 1 indicates the number of samples for each facial expression present in the dataset.

Table 1: Frequency of facial expressions in CK+ [3]

Emotion	N	Emotion	N
Angry (An)	45	Fear (Fe)	25
Disgust (Dis)	59	Sadness (Sad)	28
Contempt (Con)	18	Surprise (Sur)	83
Happy (Hap)	69		

The proposed method was evaluated using leave one out replication for validation purposes. We conduct experiments allowing variation in the radius influence for the AU and facial expressions models in the range $\gamma_{au} = \gamma_{ex} = [0.2,1]$ with steps of 0.1. To facilitate comparison of our work against other approaches, rules outputs were defuzzified using a max operation, but such defuzzification is strictly not part of the model.

The results in Table 2 were obtained varying the parameters of γ for the models yielding a mean accuracy of 89.04±0.91% and a maximum accuracy of 91.41±28% with either γ_{au} =0.4 and γ_{ex} = 0.2 or γ_{au} = 0.2 and γ_{ex} =0.8. The clustering results for the parameters γ_{au} and γ_{ex} are similar (ANOVA: p>0.05) for values greater than 0.2. Notwithstanding, the parameter values affect recognition rates for specific facial expressions. Tables 3 and 4 suggest a prevalence in false positive errors. In the cases of anger and contempt this may be due to the likeness between the AU related to these emotions. In the case of the rules, lower values of γ yielded higher number of rules. The number of rules for AU models was of 74±34 and for facial expressions of 23±3. Fig. 2 shows the association map of the AU with the emotion, inferred from the rules generated by the model. The rules generated were validated by psychologists. Examples of rules are:

- IF dimpler IS weak AND Lips part IS strong THEN contempt IS very weak.
- IF dimpler IS strong AND lips part IS very weak THEN contempt IS strong.

Table 2: Mean accuracies following leave-one-out replication for combinations of values γ_{au} and γ_{ex}.

γ	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	$\bar{x}$	σ
0.2	0.90	0.89	0.89	0.89	0.90	0.90	**0.91**	0.90	0.90	**0.90**	0.008
0.3	0.90	0.89	0.90	0.88	0.88	0.89	0.90	0.89	0.88	**0.89**	0.009
0.4	**0.91**	0.89	0.89	0.90	0.88	0.89	0.88	0.88	0.89	**0.89**	0.012
0.5	0.89	0.88	0.89	0.89	0.89	0.90	0.90	0.90	0.90	**0.89**	0.007
0.6	0.88	0.90	0.90	0.90	0.89	0.89	0.88	0.89	0.90	**0.89**	0.008
0.7	0.90	0.88	0.89	0.89	0.89	0.89	0.90	0.88	0.88	**0.89**	0.007
0.8	0.90	0.88	0.88	0.90	0.90	0.89	0.89	0.89	0.88	**0.89**	0.009
0.9	0.89	0.89	0.89	0.89	0.89	0.90	0.90	0.88	0.88	**0.89**	0.007
1.0	0.89	0.88	0.87	0.87	0.89	0.89	0.90	0.89	0.89	**0.88**	0.011
$\bar{x}$	**0.90**	**0.89**	**0.89**	**0.89**	**0.89**	**0.89**	**0.89**	**0.89**	**0.89**		
σ	0.011	0.005	0.010	0.008	0.007	0.007	0.011	0.007	0.011		

Table 3. Confusion matrix of facial expression recognition for $\gamma_{au} = 0.4, \gamma_{ex} = 0.2$.

	An	Con	Dis	Fe	Hap	Sad	Sur
An	**0.89**	0.04	0.04	0.00	0.00	0.02	0.00
Con	0.06	**0.72**	0.00	0.00	0.06	0.17	0.00
Dis	0.00	0.03	**0.93**	0.00	0.00	0.03	0.00
Fe	0.00	0.00	0.00	**0.80**	0.12	0.04	0.04
Hap	0.00	0.00	0.01	0.00	**0.99**	0.00	0.00
Sad	0.07	0.00	0.00	0.04	0.00	**0.89**	0.00
Sur	0.00	0.01	0.00	0.04	0.00	0.01	**0.94**

Table 4. Confusion matrix of facial expression recognition for $\gamma_{au} = 0.2, \gamma_{ex} = 0.8$.

	An	Con	Dis	Fe	Hap	Sad	Sur
An	**0.82**	0.09	0.07	0.00	0.00	0.02	0.00
Con	0.06	**0.78**	0.00	0.00	0.06	0.11	0.00
Dis	0.03	0.02	**0.91**	0.03	0.00	0.00	0.00
Fe	0.00	0.00	0.04	**0.84**	0.04	0.00	0.08
Hap	0.00	0.00	0.01	0.00	**0.99**	0.00	0.00
Sad	0.07	0.00	0.00	0.04	0.00	**0.89**	0.00
Sur	0.01	0.01	0.00	0.02	0.00	0.00	**0.95**

IV. DISCUSSION

In [8], the authors reported an accuracy of 86% for the CK+ dataset using a descriptor which capture the change of 560 angles obtained from the combination of the 68 landmarks points. The proposed methodology use the location change of 68 landmarks plus the area of tree polygons (mouth, left and right eye) which necessitates of less processing time and still improves accuracy on average. Further, with the proposed methodology a finite number of rules explain why the model is making a decision.

The overlap of the description of the emotions through AU presented in [5] with that generated by this work has a Jaccard index of 0.43 which means that both share some AU for each emotion.

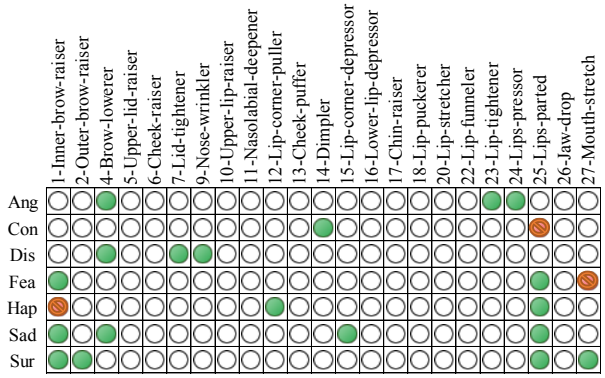


Fig. 2. Emotion description in terms of AU generated by the model. Green dots indicate when the AU must be present. Red dots indicate when an AU must be absent. White dots indicate the AU is not related to the emotion.

V. Conclusion

A system for decoding and explaining facial expressions in terms of AU using fuzzy logic has been presented. We obtained an overall accuracy of 89.04±0.91% with the maximum being 91.41± 28% for $\gamma_{au} = 0.4$, $\gamma_{ex} = 0.2$. A remarkable characteristic of the model is its ability to keep the semantic meaning between the feature representation, action units, and facial expressions models. The controlled conditions limit the generalizability of the model. Validation on people suffering face paralysis is pending.

VI. Acknowledgments

EMV was supported by scholarship #702647 from the Mexican Consejo Nacional de Ciencia y Tecnología (CONACyT). This research has been funded by project "Analysis and classification techniques of voice and facial expressions: application to neurological diseases in newborns and adults" from AMAXID supported by Mexico and Italy governments.

References

[1] A. Khanam, M. Z. Shafiq, and M. U. Akram, "Fuzzy Based Facial Expression Recognition," in Congress on Image and Signal Processing, 2008. CISP '08, 2008, vol. 1, pp. 598–602.

[2] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: the state of the art," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 12, pp. 1424–1445, Dec. 2000.

[3] P. M. Cole, P. A. Jenkins, and C. T. Shott, "Spontaneous Expressive Control in Blind and Sighted Children," Child Dev., vol. 60, no. 3, pp. 683–688, 1989.

[4] V. Cohn, J. F. J. M., "Depression, smiling and facial paralysis," Facial Palsies Amst. Neth. Lemma Holl, 2005.

[5] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2010, pp. 94–101.

[6] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on Local Binary Patterns: A comprehensive study," Image Vis. Comput., vol. 27, no. 6, pp. 803–816, May 2009.

[7] M. R. Mohammadi and E. Fatemizadeh, "Fuzzy local binary patterns: A comparison between Min-Max and Dot-Sum operators in the application of facial expression recognition," in 2013 8th Iranian Conference on Machine Vision and Image Processing (MVIP), 2013, pp. 315–319.

[8] F. Iglesias, P. Negri, M. E. Buemi, D. Acevedo, and M. Mejail, "Facial expression recognition: a comparison between static and dynamic approaches," in International Conference on Pattern Recognition Systems (ICPRS-16), 2016, pp. 1–6.

[9] R. Ghasemi and M. Ahmady, "Facial expression recognition using facial effective areas and Fuzzy logic," in 2014 Iranian Conference on Intelligent Systems (ICIS), 2014, pp. 1–4.

[10] I. Matthews and S. Baker, "Active Appearance Models Revisited," Int. J. Comput. Vis., vol. 60, no. 2, pp. 135–164, Nov. 2004.

[11] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[12] Y.-L. Boureau, J. Ponce, and Y. Lecun, "A Theoretical Analysis of Feature Pooling in Visual Recognition," in 27th International Conference on Machine Learning, Haifa, Israel, 2010.

[13] A. Hamam and N. D. Georganas, "A comparison of Mamdani and Sugeno fuzzy inference systems for evaluating the quality of experience of Hapto-Audio-Visual applications," in 2008 IEEE International Workshop on Haptic Audio visual Environments and Games, 2008, pp. 87–92.

[14] A. Priyono, M. Ridwan, A. J. Alias, R. A. O. K. Rahmat, A. Hassan, and M. A. M. Ali, "Generation of Fuzzy Rules with Subtractive Clustering," J. Teknol., vol. 43, no. 1, pp. 143–153, Feb. 2012.

[15] P. Ekman and W. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, 1978.

A CORRELATION STUDY BETWEEN SPEECH-RELATED FEATURES AND PERSONALITY TRAITS

A.Guidi[1,2], C. Gentili [3], E.P. Scilingo[1,2] and N.Vanello [1,2]

[1] Research Center "E. Piaggio", University of Pisa, Pisa,
[2] Dipartimento di Ingegneria dell'Informazione, University of Pisa, Pisa, Italy
[3]Department of General Psychology, University of Padua, Padua, Italy
guidi.andrea84@gmail.com, c.gentili@unipd.it, e.scilingo@ing.unipi.it, nicola.vanello@unipi.it

Abstract: **Voice signal has been widely investigated to characterize mood and emotional states. A further interesting dimension could regard the personality traits. In fact, speech production can be related to personality traits evaluated by others. The relationship between personality traits and specific speech features is not yet fully understood and requires further investigation. In this study, a correlational analysis between some speech-related features and the personality traits, as described by the Zuckerman-Kuhlman model, is performed. An experimental protocol was administered to eighteen healthy subjects to investigate both fundamental frequency and voice quality related features. Results showed that a skewness-like measure of the fundamental frequency is negatively correlated with the Sociability dimension. The impact of personality traits and speech production studies on the characterization of mental disorders and the estimation of emotional/mood state of the speaker are discussed.**

Keywords : **personality traits, Zuckerman-Kuhlman model, Fundamental frequency, spectral slope, jitter**

I. INTRODUCTION

The analysis of speech signal allows to explore several psychological dimensions: emotion [1], mood [2], and stress [3] were widely studied in relation to the speakers' speech production. A further interesting dimension could be related to the personality traits, whose effects might overlap to the ones related to emotion and/or mood. Speech intonation parameters might be related to a set of individual and sociocultural means that can allow reaching different communication goals [4]. Probably, such a relation might be stronger in people showing some particular personality trait. According to the trait theory [5], traits can be defined as "stable internal characteristics that people display consistently over time and across situations". Different studies attempted to investigate the relationship between voice and personality. Sapir [6] proposed the hypothesis of "speech as a personality trait". Addington [7] reported that a higher pitch variation in males was perceived as more dynamic, feminine and aesthetically inclined, while in female was rated as more dynamic and extravert. Again, some prosodic features, such as mean pitch, pitch variation and speaking rate, were found to be related to the perception of competence, benevolence, extraversion, dominance and political charisma in [8,9]. These studies showed that a voice characterized by a high (low) pitch variation and a high (low) speaking rate was perceived as index of high (low) competence, while a voice showing a low pitch variation and a high speaking rate was judged with low benevolence ratings, and vice versa. Similarly, a negative (positive) correlation between mean pitch and both extraversion and dominance was detected in American female (male) speakers [8]. Political charisma and leadership were reported to be positively correlated with higher pitch in [9]. Spectra and voice quality features were also investigated in relation to personality perception in [10,11]. A correlation between speech fluency and both extroversion and neuroticism was observed in [12]. The INTERSPEECH 2012 Speaker Trait Challenge showed that the classification of personality traits, as defined by the OCEAN five personality dimensions [13], is feasible. Within this challenge, hundreds of short clips, on average lasting 10 s, were evaluated by a pool of judges to assess the personality traits by using the Big Five Inventory questionnaire [14]. Acoustic features outperformed linguistic and psycholinguistic features to achieve an automatic recognition of speaker personality trait [15].

Although some useful indications about a significant relationship between personality traits and voice production can be drawn from the literature review, the work on this topic is far from being concluded. For instance, the currently available studies mostly rely on the estimation of the perceived personality traits, without exploring the possibility of using dedicated personality tests. Moreover, the relationship of personality traits and specific speech features have still to be clarified.

In this study, a correlational analysis between some speech-related features and the personality traits, as

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

described by the Zuckerman-Kuhlman model [16], is performed. Specifically, a correlation analysis between features related to speech fundamental frequency (F_0) and voice quality, and the six factors of personality traits defined in the above cited models, will be conducted. This study will be performed on healthy subjects using a structured speech task.

II. METHODS

Experimental Protocol: Eighteen healthy subjects (12 females, 23.66 ± 2.28 year) without any history of psychiatric disorder were enrolled. Subjects were asked to fill out the Zuckerman-Kuhlman Personality Questionnaire (ZKPQ) at home, about 4 days before performing the experimental protocol. The ZKPQ is a self-report questionnaire that provides information about personality in terms of five dimensions: Impulsive Sensation Seeking (ImpSS), Aggression-Hostility (Agg-Host), Sociability (Sy), Neuroticism-Anxiety (N-Anx) and Activity (Act). Subjects were asked to read a neutral text ("The universal declaration of Human rights", lasting 3 minutes) twice, at the beginning and at the end of the experimental protocol, after about 30 minutes. In the following, the first and the second reading task will be indicated as Rd1 and Rd2, respectively. In addition, they were asked to comment a set of Thematic Apperception Test (TAT) images [17], between the two neutral text reading tasks. Subjects were driven to comment all of them or were stopped after 3 minutes of speaking. One task was chosen to provide a neutral baseline of the vocal production (reading of neutral text), while the other was customized to emphasize some particular phenomena related to personality traits. In fact, TAT is a traditional projective test used to assess personality disorders. Furthermore, since anxiety can play a role as a confounding factor in speech-related features dynamics [19,20] we asked subjects to fill out the short form of the State-Trait Anxiety Inventory (STAI) for state anxiety, i.e. the STAI-X2 test [20]. This form has shown comparable psychometric properties to the original one and therefore is preferred in case of multiple administrations [21]. Subjects were asked to compile the scale at the beginning and at the end of each single task. Audio signals were acquired by means of a high quality system (AKG P220 Condenser Microphone, M-Audio Fast-Track), with a sampling frequency equal to 48 KHz and a resolution of 32 bits.

Speech Feature Extraction. The estimated speech features took into account the overall F_0 dynamics and the voice quality of the speakers. More in detail, skewness-like measurement of F_0 (*Median/Mean*), a frame-to-frame Jitter Factor (*LPJit*) estimate, and the Glottal Flow Spectral Slope (*Slope*) were investigated. The *Median/Mean* provides a global information about the tone of the speaker, while the *LPJit* and the *Slope* carry information about the quality of the speakers' voice. Specifically, *LPJit* describes the short-term variability of the voice, while *Slope* can be used to describe different phonation types (e.g creaky, tense or breathy). As a first step, voiced sounds are extracted from speech signals by means of a Voice Activity Detection algorithm that exploits signal energy and Zero Crossing Rate as described in [22]. Then, the proposed features are estimated within each segment from the F_0 contour, obtained according to the double iteration method as described in [18], based on Camacho's SWIPE' [23] algorithm. This latter algorithm estimates speech fundamental frequency using a spectral matching approach. The *Median/Mean* is computed as the ratio of median over mean of F_0. This ratio also acts as a normalizing procedure across subjects to face individual differences in tone. The *LPJit* is estimated in each segment using 4 glottal cycles-long time windows according to the following formula

$$LPJit = \frac{1}{N-1}\sum_{i=1}^{N-1}|F_{i+1} - F_i| \Big/ \frac{1}{N}\sum_{i=1}^{N} F_i \qquad (1)$$

where F_i is the fundamental frequency at the i-th window. *LPJit* represents a low-pass version of the classical jitter measure. *Slope* is obtained according to the procedure described in [24]. According to this approach, the glottal flow spectrum is estimated after the removal of the vocal tract effects. This result is obtained by averaging all the energy-normalized frames, obtained from voiced speech spectra using sliding windows. At the end, the glottal flow spectral slope is estimated by fitting a straight line over 300-3000Hz frequency band of the glottal flow spectrum. Both *LPJit* and *Slope* are already normalized measures and can be directly used in a correlation study at group level.

Statistical Analysis. A non-parametric Sign Test [25] is used to compare short-form STAI scores acquired before and after each task to evaluate possible effects on subject anxiety due to task execution. Moreover, a correlation analysis between monitored anxiety levels and speech features is also performed at group level by means of the non-parametric Spearman method. Similarly, the Spearman method is used to estimate the correlation coefficient between the personality trait dimensions and the corresponding speech-related features ($\alpha<0.05$). The Benjamini-Hochberg procedure is used to correct p-values for the false discovery rate.

III. RESULTS

No statistically significant differences were observed by investigating short-form STAI scores acquired before and after the reading tasks, thus confirming that these tasks did not induce anxiety level

changes. Interestingly, no statistically significant differences between short-form STAI scores related to the beginning and the end of the whole experiment were found, while no significant correlation coefficients, between STAI scores and speech-related features, were reported. In Table 1, the Spearman's correlation coefficients between ZKPQ scores and speech features are reported.

Table 1. Spearman's correlation coefficients between ZKPQ scores and speech features.

#task	Feat	Imp-SS	Agg-Host	Sy	N-Anx	Act
Rd 1	Median/Mean	0.01	0.29	**-0.66**	0.16	0.05
Rd 1	LPJit	-0.51	-0.08	-0.07	0.09	-0.27
Rd 1	Slope	0.09	0.55	0.10	-0.39	0.11
TAT	Median/Mean	-0.22	0.21	-0.41	-0.07	0.43
TAT	LPJit	**-0.60**	-0.01	-0.10	0.19	0.02
TAT	Slope	0.00	**0.66**	-0.02	-0.29	0.08
Rd 2	Median/Mean	-0.02	0.02	**-0.82**	0.35	-0.06
Rd 2	LPJit	-0.24	-0.29	0.23	0.00	0.00
Rd 2	Slope	0.00	0.55	0.32	-0.48	0.10

The values that report a significant p-value according to the Benjamini-Hochberg procedure are highlighted in bold.

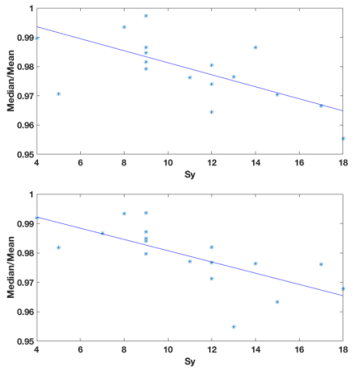


Figure 1. Reading task: scatter plot of *Median/Mean* vs *Sy*. Upper: Rd1 (ρ=-0.66). Lower: Rd2 (ρ=-0.82).

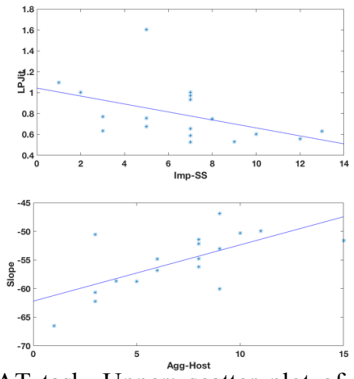


Figure 2. TAT task. Upper: scatter plot of *LPJit* and *Imp-SS*. (ρ=-0.60). Lower: Slope and Agg-Host (ρ=0.66).

Interestingly, *Median/Mean* reports a negative Spearman's correlation coefficient with the Sociability trait dimension in both reading tasks. Scatter plots of *Median/Mean* vs *Sy* values obtained in both reading tasks are shown in Fig. 1. In addition, the analysis of the commenting of TAT image task shows that *LPJit* correlates negatively with Impulsive Sensation Seeking trait dimension and *Slope* positively with Aggression-Hostility trait. In Fig. 2 while the scatter plots related to the TAT task are shown.

IV. Discussion and Conclusion

The results obtained on this study revealed some significant correlations. Interestingly, different results were obtained with the two tasks. Neutral text reading and TAT image commenting might in fact emphasize specific phenomena related to personality traits. As regards neutral text reading, a negative Spearman's correlation between *Median/Mean* feature and the Sociability trait dimension score is reported. This is verified in both repetitions of this task. Such a result could indicate that the more the speaker shows a sociable personality, the more the F_0 distribution shows a negative-skewed behaviour. A negative-skewed F_0 distribution is usually reported in relaxed and calm voices. The results on TAT images showed a negative correlation between *LPJit* and *Imp-SS*. This result might indicate a possible less hoarse voice in persons with a marked Impulsive Sensation Seeking trait. In this task, a positive correlation between *Slope* and *Agg-Host* was found. According to the fact that Slope is always negative and a steeper value is usually associated with a breathier voice, while a flat spectrum to a tenser or creakier voice [26], this result might indicate that a more aggressive trait is associated to a tenser or creakier voice. The coherent results, obtained between the two repetitions of the neutral reading task, seem to indicate a robust behavior of *Median/Mean*. Since personality traits have long-temporal dynamics, an analysis performed on longer time intervals might further elucidate the relevance and the robustness of this feature. Interestingly, no significant correlations were found between speech-related features and anxiety levels. We have to stress that anxiety levels were not significantly different before and after the recordings. Future studies could explore the use of stressful tasks to further investigate possible interactions of anxiety, personality traits and speech features.

The results of this study could have an impact on the comprehension of mental disorders. In fact, according to Zuckerman [16] severe personality disorders such as psychopathy, antisocial behaviour and forms of paranoid hostility would be a

combination of research of impulsive sensation seeking and low sociability.

REFERENCES

[1] P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, “Analysis of Emotional Speech---A Review,” in Toward Robotic Socially Believable Behaving Systems-Volume I, Springer, 2016, pp. 205–238.

[2] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, “A review of depression and suicide risk assessment using speech analysis,” Speech Commun., vol. 71, pp. 10–49, 2015.

[3] C. L. Giddens, K. W. Barron, J. Byrd-Craven, K. F. Clark, and A. S. Winter, “Vocal indices of stress: a review,” J. Voice, vol. 27, no. 3, pp. 390--e21, 2013.

[4] A. S. Silnitskaya and A. N. Gusev, “character and temperamental determinants of prosodic parameters in natural speech,” Psychol. Russ. State art, vol. 6, no. 3, 2013.

[5] D. Bernstein, Essentials of psychology. Cengage Learning, 2013.

[6] E. Sapir, “Speech as a personality trait,” Am. J. Sociol., pp. 892–905, 1927.

[7] G. B. Ray, “Vocally cued personality prototypes: An implicit personality theory approach,” Commun. Monogr., vol. 53, no. 3, pp. 266–276, 1986.

[8] K. R. Scherer and U. Scherer, “Speech behavior and personality,” Speech Eval. psychiatry, pp. 115–135, 1981.

[9] F. Weninger, J. Krajewski, A. Batliner, and B. Schuller, “The voice of leadership: Models and performances of automatic analysis in online speeches,” Affect. Comput. IEEE Trans., vol. 3, no. 4, pp. 496–508, 2012.

[10] G. Mohammadi, A. Origlia, M. Filippone, and A. Vinciarelli, “From speech to personality: Mapping voice quality and intonation into personality differences,” in Proceedings of the 20th ACM international conference on Multimedia, 2012, pp. 789–792.

[11] C. Hu, Q. Wang, L. A. Short, and G. Fu, “Speech spectrum’s correlation with speakers’ Eysenck Personality Traits,” PLoS One, vol. 7, no. 3, p. e33906, 2012.

[12] B. Gawda, “Neuroticism, extraversion, and paralinguistic expression,” Psychol. Rep., vol. 100, no. 3, pp. 721–726, 2007.

[13] P. H. Lodhi, S. Deo, and V. M. Belhekar, “The Five-Factor Model of Personality,” in The five-factor model of personality across cultures, Springer, 2002, pp. 227–248.

[14] B. Rammstedt and O. P. John, “Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German,” J. Res. Pers., vol. 41, no. 1, pp. 203–212, 2007.

[15] F. Alam and G. Riccardi, “Fusion of acoustic, linguistic and psycholinguistic features for speaker personality traits recognition,” in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 955–959.

[16] M. Zuckerman, “Zuckerman-Kuhlman Personality Questionnaire (ZKPQ): an alternative five-factorial model,” Big five Assess., pp. 377–396, 2002.

[17] H. A. Murray, “Uses of the thematic apperception test,” Am. J. Psychiatry, vol. 107, no. 8, pp. 577–581, 1951.

[18] N. Vanello, A. Guidi, C. Gentili, S. Werner, G. Bertschy, G. Valenza, A. Lanata, and E. P. Scilingo, “Speech analysis for mood state characterization in bipolar patients,” in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2012.

[19] E. Moore, M. A. Clements, J. W. Peifer, L. Weisser, and others, “Critical analysis of the impact of glottal features in the classification of clinical depression in speech,” Biomed. Eng. IEEE Trans., vol. 55, no. 1, pp. 96–107, 2008.

[20] C. D. Spielberger, “Manual for the State-Trait Anxiety Inventory STAI (form Y)(‘ self-evaluation questionnaire’),” 1983.

[21] T. M. Marteau and H. Bekker, “The development of a six-item short-form of the state scale of the Spielberger State---Trait Anxiety Inventory (STAI),” Br. J. Clin. Psychol., vol. 31, no. 3, pp. 301–306, 1992.

[22] B. Atal and L. Rabiner, “A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition,” Acoust. Speech Signal Process. IEEE Trans., vol. 24, no. 3, pp. 201–212, 1976.

[23] A. Camacho and J. G. Harris, “A sawtooth waveform inspired pitch estimator for speech and music,” J. Acoust. Soc. Am., vol. 124, no. 3, pp. 1638–1652, 2008.

[24] A. Ozdas, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes, “Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk,” Biomed. Eng. IEEE Trans., vol. 51, no. 9, pp. 1530–1540, 2004.

[25] J. D. Gibbons and S. Chakraborti, Nonparametric statistical inference. Springer, 2011.

[26] J. Kuang and M. Libermann, “The effect of spectral slope on pitch perception,” in Sixteenth Annual Conference of the International Speech Communication Association, 2015.

TRANSFER LEARNING ON IMAGINED SPEECH ELECTROENCEPHALOGRAM USING BAG OF FEATURES

Jesús S. García-Salinas, Luis Villaseñor-Pineda, Carlos A. Reyes-García, A. Torres-García
Biosignals and Medical Computing Laboratory, Instituto Nacional de Astrofísica Óptica y Electrónica, Puebla, México
{jss.garcia, villasen, kargaxxi, alejandro.torres}@inaoep.mx

Abstract: Brain Computer Interfaces require a model generation which solves a specific task. However, the models have a drawback when they must be expanded to include new tasks. Although imagined speech is a recent neuroparadigm for such interfaces, it also has the same inconvenient, for example, if new imagined words need to be added, a new training process is necessary. In this work, a Bag of Features representation is explored to expand the vocabulary. This method extracts characteristic units from electroencephalogram signals and then represents the imagined words from them. Initially, transfer learning without a calibration step was used to add a new word to the vocabulary. Later, a calibration step with different training set sizes of the new word was tested. The obtained results showed that Bag of Features method allows the extension of the vocabulary with a small accuracy decrement. The average accuracy for all subjects for "up" word transferring was 65.27%, meanwhile the average accuracy for the baseline, when no transfer learning is applied, was 68.93%. Moreover, applying a calibration step increases the accuracy for a word transferring.

Keywords: EEG, Imagined Speech, Transfer Learning, Bag of Features.

I. INTRODUCTION

A Brain Computer Interface (BCI) can be used to transform the brain signals in commands to control a device. For this task, the user must produce a brain activity pattern, which can be evoked internally or produced by an external stimulus. This brain activity pattern will be identified by the BCI system and transformed into commands for a particular device. In this work electroencephalograms (EEG) were used to record brain electrophysiological activity. Also, this work is focused on brain signals evoked by imagined speech, i.e. to imagine a word diction without emitting any sound nor articulating any facial movement.

When a BCI is trained for a specific task, to extend it normally requires to train again the BCI adding the new task information [1]. In the case of imagined speech, it could be needed to extend the BCI vocabulary for recognizing new imagined words. The objective of this work is to analyze an imagined speech vocabulary extension, using transfer learning in a Bag of Features (BoF) model. The proposed method would help to improve the BCIs scalability in practical applications.

The Bag of Features method consists in calculate a *codebook* that contains a set of *codewords* for representing a document, a signal or an image; as histograms of the generated codewords. Although the order information of codewords is ignored, the BoF model is very effective to capture characteristic units [6].

II. METHOD

A. Classification based on Bag of Features

In Fig. 1, the general method's flowchart is shown [3]. In the feature extraction step, each imagined word EEG epoch (i.e. repetition) w_i for the s_j subject can be seen as a 14 by n matrix $X_{wi,sj}$, where 14 is the channels number and n is the samples number recorded for this epoch, see (1).

$$X_{w_i,s_j} = \begin{bmatrix} x_{1,1} & x_{2,1} & \cdots & x_{14,1} \\ x_{1,2} & x_{2,2} & & x_{14,2} \\ & \vdots & \ddots & \vdots \\ x_{1,n} & x_{2,n} & \cdots & x_{14,n} \end{bmatrix} \quad (1)$$

From the $X_{wi,sj}$ matrix a feature extraction is applied to keep the spatial information of the signal in a new representation, this will allow a pattern detection from different areas of the brain and their activation during imagined speech. This representation, shown in (2), takes the microvolt values of the signal as features.

Then the y instances are made by taking samples from all channels at the same time instant, and concatenating them in a single vector. Resulting in n instances per epoch.

$$y = \left\{ \begin{matrix} [x_{1,1}, \dots, x_{14,1}] \\ [x_{1,2}, \dots, x_{14,2}] \\ \vdots \\ [x_{1,n}, \dots, x_{14,n}] \end{matrix} \right\} \quad (2)$$

Later, a clustering method is applied to obtain the

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

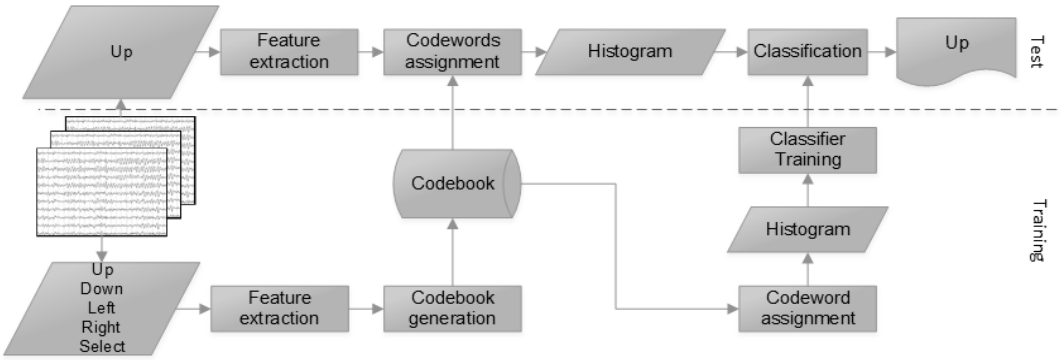


Figure 1. General method flowchart.

codebook from these features, this clustering is applied to the training set from each class independently [2], the resulting clusters are later joined in a complete codebook [4]. Each cluster prototype will receive the name of **codeword**. In this step, kMeans algorithm is applied to find 200 clusters. Thus, from each class are obtained k_i clusters, where i is the class number, and they are computed as follows

$$k_i = \frac{K}{C} \qquad (3)$$

where C is equal to classes number which participate in cluster generation and K is the clusters number. Then each k_i is equal to 40 clusters using five classes.

The next step, is to replace every y instance with one codebook's codeword, this results as a sequence of the codewords over the $X_{wi,sj}$ matrix. Then, the original signals become sequences of codewords.

Later a histogram is calculated for each imagined word epoch. Due to the differences in signal lengths, all the histograms must be normalized.

Once the data are converted in a set of histograms, they become the training set for a classifier, where each histogram represents a word epoch. The classification step is performed with a Naive Bayes classifier.

III. TRANSFER LEARNING RESULTS

A. Experimental setup

An imagined speech data set was recorded in [5], which is composed of the EEG signals of 27 native Spanish speaking subjects, registered through the Emotiv EPOC headset, which has 14 channels and a sampling frequency of 128 Hz. The data consist of 5 Spanish words (i.e. "arriba", "abajo", "izquierda", "derecha", "seleccionar"; translated to English as "up", "down", "left", "right", "select" respectively) with 33 epochs each one, with a rest period between them.

Data were processed with Common Average Reference (CAR). Also, a low-pass filter to reduce the noise was applied, such filter is an infinite impulse response Butterworth filter with a stop-band frequency of 50 Hz and a pass-band frequency of 40 Hz. No additional signal processing was applied, and the microvolt signal values are used as features.

All the experiments were realized taking 75% of the imagined words epochs randomly for the codebook generation, and the remaining data was used for testing purposes. Also, all the experiments were repeated ten times due to the random properties of the method. Thus, the results are given as averages from all subjects and their repetitions.

B. Transfer learning

A vocabulary extension was simulated by excluding one of the five words in the data set. The codebook was generated using four words, and the new word was represented using the previously calculated codewords.

Also, the method was individually applied to each subject's dataset. The obtained results from the 27 subjects are shown in Fig 2. The average accuracy when transfer learning is applied to the "up" word is 65.27% ± 12.64. Whereas, the method accuracy for the five words (i.e. when no class is excluded in the codebook generation) is 68.93% ± 12.43. Also, when transfer learning is applied to the "down" word, the average accuracy is 65.49% ± 12.37. It must be highlighted that transfer learning results were obtained without calibration data.

To contrast the transfer learning behavior, a baseline confusion matrix is shown (i.e. with five words) in Table 1 and the confusion matrix with transfer learning for the "up" class in Table 2.

Table 1. Baseline confusion matrix. (Class prediction is shown in columns).

	Up	Down	Left	Right	Select
Up	74.95	9.76	4.72	6.06	4.49
Down	8.33	67.26	6.8	13.37	4.21
Left	3	7.36	66.34	10.74	12.25
Right	4.62	13.19	12.5	62.17	7.5
Select	3.19	6.57	12.54	4.76	72.91

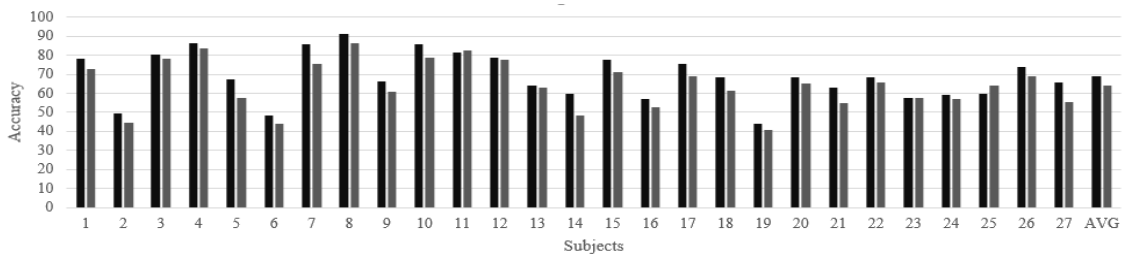


Figure 2. Transfer learning results for the word "up" (baseline results in dark color)

The confusion matrices were generated averaging all the subject's confusion matrices and they are presented as global accuracy percentages for an easy interpretation. The confusion matrix in Table 1 corresponds to the results obtained in Fig. 1.

Table 2. "Up" class exclusion confusion matrix.

	Up	Down	Left	Right	Select
Up	62.04	18.56	9.91	5.83	3.66
Down	13.06	73.19	5.79	4.54	3.43
Left	10.6	11.76	61.57	9.35	6.71
Right	7.04	10.56	10.6	59.49	12.31
Select	5.6	9.58	4.31	10.46	70.05

In addition, Table 3 corresponds to the "down" word transferring classification matrix, this table shown a different behavior from transferred class. The accuracy obtained from "up" word decreases 14.35 in comparison to the baseline. Otherwise the "down" word accuracy increased in 3.2.

Table 3. "Down" class transferring confusion matrix.

	Up	Down	Left	Right	Select
Up	70.23	7.45	11.67	6.39	4.26
Down	20.14	67.87	4.12	4.4	3.47
Left	22.18	3.34	57.87	9.68	6.85
Right	16.25	1.94	8.61	60.93	12.27
Select	13.1	3.19	3.29	9.86	70.56

In a different analysis, the average histograms from all subjects were obtained to calculate which codewords are used by the excluded classes. This analysis should complement the confusion matrix analysis, comparing the confusion among classes and the percentage of codewords used.

The Table 4 summarizes the percentage of codewords used for the transferred classes averaging the results from all subjects.

Table 4. Codewords' distribution to represent the transferred classes.

Percentage used	Up	Down	Left	Right	Select
Up	-	39.94	20.72	22.92	16.42
Down	27.03	-	28.82	33.13	16.02

From Table 4, it is expected that the confusion matrix of the "up" word transferring shown a bigger confusion with "down" word than others words. Otherwise, when "down" word is transferred, it is expected that it would be more confused with "right" word than others words. Nevertheless, Table 3 shows a different behavior which is discussed in next section.

C. Calibration

To improve the classification results, a small amount of epochs of the transferred word were used in the codebook generation. As mentioned before, 75% of the imagined word's epochs are used to generate the BoF, resulting in the use of 25 epochs for each class.

The Table 5 shows the method's accuracies in a transfer learning approach, including different amounts of epochs for the "up" word.

Table 5. "Up" class accuracies using epochs of this class in the codebook generation.

Epochs number	"Up" accuracy	Total accuracy
0	62.04 ± 20.5	64.35 ± 12.64
1	62.91 ± 19.31	65.84 ±13.16
3	65.92 ± 19.74	67.2 ± 13.47
5	65.5 ± 20.28	68.19 ± 13.57
8	66.34 ± 20.3	68.77 ± 12.98
10	67.54 ± 17.86	68.25 ± 13.46

Table 6 shows the accuracies from the "down" word in a transfer learning approach using different amounts of its epochs in the codebook generation step.

Table 6. "Down" class accuracies using epochs of this class in the codebook generation.

Epochs number	"Down" accuracy	Total accuracy
0	67.87 ± 20.5	66.67 ± 12.77
1	71.38 ± 23.96	66.17 ± 13.47
3	70.6 ± 23.71	67.18 ± 13.32
5	70.78 ± 23.41	67.5 ± 12.88
8	70.64 ± 25.06	67.62 ± 13.23
10	71.66 ± 23.65	68.05 ± 12.77

Both last two tables show the accuracy of the transferred classes and the total accuracy of the five classes, averaging the result of all the subjects.

IV. DISCUSSION

The baseline accuracy, when no transfer learning is applied is 68.93% ± 12.43. When applying transfer learning, a total accuracy of 65.27% ± 12.6 was obtained for "up" word, this is an accuracy decreasing of 3.66. When "down" word is transferred a total accuracy of 65.49% ± 12.77 was achieved, this is an accuracy decreasing of 3.44. The blind transferring shows a slight decreasing compared to the baseline accuracy. Moreover, a Kruskal-Wallis test showed that there is no statistical significance between the transfer learning approach and the baseline results.

Also, the obtained classification results are similar to related work, in [5] an accuracy of 68% ± 16 for the same database was reported, without using a transfer learning approach. In Table 2, the "up" word transferring confusion matrix shown a confusion between this word and "down" word. This result was expected due to the results in Table 4, which show that when "up" word is transferred, it uses codewords from "down" word more than others to be represented.

The "down" word shows a different behavior. The results in Table 3 do not correspond to the codewords distribution in Table 4. The "down" word codification has more codewords of "right" word, and the confusion matrix shows a higher confusion with "up" word.

In Table 5, the accuracy when a calibration step is added to the "up" word transferring is shown. As it is expected, the accuracy increases as the imagined word's data are increased. Otherwise, Table 6 shows a variable total accuracy while the data of the of the transferred "down" word is added. Although as it was expected the recognition accuracy for the new word was improved.

V. CONCLUSION

The proposed method allows the extension of the imagined speech vocabulary with a small accuracy decrement. The exclusion of the class data from generation step has obtained similar accuracy results as if the word's data were used. Also, the use of a calibration step increases the classification accuracies. Nevertheless, for practical applications this step must use the less information possible from the transferred word. Moreover, using small amounts of epochs may have an important issue, if these data are not representative of the word, the codebook will not be able to represent correctly the words.

The analysis of the used codewords when "down" word is transferred, shown that the quantity of codewords are not correlated to the confusion among classes. Thus, a deeper analysis must be done to explain these results.

It is also interesting to highlight that the feature extraction step takes only into account the signal microvolt values. Hence, the impact of noise in the same bandwidth of the filtered signals must be explored. Additionally, in future experiments, the frequency information of the signal could be taken into account to search for patterns related to the brain activity's frequency bands. Nevertheless, the lack of frequency features extraction step of the signal makes the method more suitable for a real time BCI.

ACKNOWLEDGMENTS

The present work was partially supported by CONACyT (scholarship 401887 and project grant CB-2015-01-257383). Also, the authors thank the support of the Italian Foreign Affairs and Cooperation Ministry, and the International Cooperation for Development Mexican Agency for the project MX14MO06.

REFERENCES

[1] F. Lotte and C. Guan, "Learning from other subjects helps reducing Brain-Computer Interface calibration time," 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, 2010, pp. 614-617.

[2] S. Lazebnik and M. Raginsky, "Supervised Learning of Quantizer Codebooks by Information Loss Minimization," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 7, pp. 1294-1309, July 2009.

[3] P. Ordonez, T. Armstrong, T. Oates and J. Fackler, "Using Modified Multivariate Bag-of-Words Models to Classify Physiological Data," 2011 IEEE 11th International Conference on Data Mining Workshops, Vancouver, BC, 2011, pp. 534-539.

[4] A. Plinge, R. Grzeszick and G. A. Fink, "A Bag-of-Features approach to acoustic event detection," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, 2014, pp. 3704-3708.

[5] A. A. Torres-García, C. A. Reyes-García, L. Villaseñor-Pineda, G. García-Aguilar, "Implementing a fuzzy inference system in a multi-objective EEG channel selection model for imagined speech classification", Expert Systems with Applications, Volume 59, 15 October 2016, Pages 1-12.

[6] J. Wang, P. Liu, Mary F.H. She, S. Nahavandi, A. Kouzani, "Bag-of-words representation for biomedical time series classification", Biomedical Signal Processing and Control, Volume 8, Issue 6, November 2013, Pages 634 - 644.

ROUND TABLE:

VOICE AND SPEECH PROCESSORS (VSP): READY FOR GLOBAL DEPLOYMNT IN MOBILE DEVICES?

ROUND TABLE

VOICE AND SPEECH PROCESSORS (VSP): READY FOR GLOBAL DEPLOYMENT IN MOBILE DEVICES?

K. Izdebski[1],[2], J. Bryzek[3]

[1,] PVSF, Chairman, San Francisco, CA, USA
[2]Bioengineering Advisory Board, Santa Clara University, Santa Clara, CA, USA
[3]J. Bryzek, Chief Executive Officer and Director, Inc.. Oakland, CA, USA
kizdebski@pvsf.org

Abstract
Since introduction of iPhone in 2007, global smart phone market grew to about 7 billion subscribers, almost matching the world's population. Each of the phone has built-in at least one microphone and powerful computer with connection to Cloud. This enables embedding application based on real-time language-independent voice and speech processing (VSP).

We believe that the progress in VSP technologies reached the point enabling global commercialization, justifying founding of a dedicated startup company.

For this Round Table, we would like to brainstorm the feasibility of such startup by discussing potential applications, supporting VSP technology status, and their attractiveness for embedding in mobile devices.

Some applications which seem to be ready for commercialization include:

Social Networking
Social networking could be enhanced by addition of automatic real-time sharing of the person's emotional state during phone conversation, such as the mood (up, down), love, anger, excitement, joy, sadness, indifference, hesitation, hunger, pain and fear, etc., making phone interaction more personal and engaging. Global adoption could be quite rapid and it could represent one of the largest markets for VSP.

Unobtrusive Health Monitoring
About half of human population, 3.5 billion people, has substandard or no medical care. The emerging eHealth is attempting to embed health diagnostics in mobile devices, to reach a large fraction of global population.

Resent progress in VSP enables early detection of multiple diseases and various medical conditions encountered on the road to senescence. VSP is one of technologies creating a foundation for Unobtrusive Health Monitoring, defined as monitoring health without any action by the monitored person. Once deployed on mobile devices supplemented by Deep Data Mining and AI algorithms, it could enable health monitoring of the significant fraction of global population, a stepping stone towards healthcare abundance.

The range of health conditions which current generation of VSP could detect and monitor includes neurodegenerative disorders (Parkinson, dementia, autism, etc.), mental states, depression, vocal-cord nodules, Apnea, Asphyxia, Hypothroidsim, Hyperbilirubinemia, cleft palate, Ankyloglossia, respiratory distress syndrome, deafness and with time of general health status. Attempts were made to establish acoustic biopsies of the organs of voice and speech. Sound capture will help in monitoring malignancies, trauma, etc.

Personality Traits
Correct recognition of emotions in the human voice is fundamental to human interactions. Faulty detection of deception, of emotional states can lead to conflicts instead of resolution and peaceful solutions. Rapid and verifiable detections of various emotional states are of enormous social and commercial potential. Voice authentication and personality traits could simplify many aspects of human day-to-day interactions and enable multiple applications, such as preventing personality thefts, anti-bribery/anti-corruption compliance monitoring, detection of terrorism and risk assessment. Speech analysis of group's behavior and ethnic traits (behaviors) will be needed to be included in the analytics to determine group dynamics. Global alliances, comprising communications and rapid decision makings by people from diverse cultural, linguistic and emotional groups

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

need to be clearly identified as friends or enemies, and voice/speech signals are the best ways to be utilized on the large scale to draw significant decisions and conclusions of actions to follow.

One of the goals for Round Table is to check:

- Participants interest in becoming either forward-thinking contributors or advisors, capable of identifying global need of voice technology and voice applications.
- VSP technologies available for licensing to the new company, if formed.
- Potential funding sources for new company; it is envisioned the need for $2.5M initial round.

As a point of reference, Voyager Labs http://voyagerlabs.co/company/about-us/ is developing somewhat similar applications, although based on cognitive computing analyzing in real-time billions of publicly available unstructured data points. Company was formed in 2012 and secured $100M equity funding.

The concept of new company was conceived by Dr. Krzysztof Izdebski and Dr. Janusz Bryzek.

- Krzysztof is an internationally recognized voice-speech patho-physiologist, Associate Clinical Professor at the UCLA, School of Medicine, clinical consultant for Speech Pathology at Kaiser Permanente Medical Foundation, Advisory Board member of the Bioengineering Division of Engineering at Santa Clara University, founder and Chairman of Pacific Voice and Speech Foundation, and a Vice President of the World Voice Consortium. Several other voice/speech processing gurus are being contacted to complement the team full time.
- Janusz is a serial visionary entrepreneur with 11 Silicon Valley sensor startups.

SESSION IV:

VOICE AND SPEECH IN NEUROLOGY

DYSARTHRIC SPEECH ANALYSIS BY MEANS OF THE PRINCIPAL COMPONENTS OF THE SPECTROGRAM

A. Kacha[1], F. Grenez[2], Juan R. Orozco-Arroyave[3, 4], J. Schoentgen[2, 5]

[1]Laboratoire de Physique de Rayonnement et Applications, University of Jijel, Jijel, Algeria
[2]Laboratory BEAMS, Université Libre de Bruxelles, Brussels, Belgium
[3]Faculty of Engineering, University of Antioquia, Medellín, Colombia
[4]Pattern Recognition Lab, Friedrich-Alexander-Universität, Erlangen-Nürnberg, Germany
[5]National Fund for Scientific Research, Belgium
akacha@ulb.ac.be, fgrenez@ulb.ac.be, Rafael.orozco@udea.edu.co, jschoent@ulb.ac.be

Abstract: **Principal component analysis (PCA) of the spectrogram is proposed for dysarthric speech analysis. Principal component analysis involves a dimensionality reduction that represents the spectrogram in a low-dimensional subspace while retaining most of the variation of the original data. Each principal component (PC) of the spectrogram is a weighted sum of all frame spectra. The acoustic markers used to document articulation deficits in Parkinson speakers are the spectral mean of the first principal component and the two weighted sums of frame spectra associated with the positive and negative coefficients of the second principal component. The analysis is applied to a corpus comprising monologues produced by 50 control and 50 Parkinson speakers. Results show that the acoustic markers differ statistically significantly between Parkinson and control speakers.**

Keywords: Principal component analysis, dysarthric speech, spectrogram.

I. INTRODUCTION

Dysarthria is a speech disorder that may be caused by neurological diseases such as Parkinson disease, cerebral palsy, etc. Acoustic analysis of dysarthric speech aims at understanding the production of dysarthric speech and developing methods to detect the disease at an early stage or to quantify the degree of severity and to monitor its progress. Dysarthria in patients with Parkinson disease is referred to as hypokinetic. Previous studies of Parkinson speakers have shown deficits in vowel articulation [1].

Analyses of dysarthria may involve acoustic markers that document phonatory, prosodic and articulatory properties of speech [2]. The analysis is based on sustained vowels, isolated words or sentences mostly. Obtaining acoustic markers for the latter involves a manual segmentation. Studies of read texts or spontaneous speech are few [3-5] and may ask for a large number of acoustic cues [3] causing difficulties of interpretation and masking the relevance of individual cues in distinguishing between dysarthric and control speakers.

The long-term average spectrum (LTAS) of continuous speech has been used in several studies to report voice quality as well as other speaker properties or states. Most LTAS-based clinical assessments of voice quality have focused on patients with dysphonia caused by organic or functional diseases of the larynx. Only few studies have targeted dysarthric speech [6, 7].

In this study, principal component analysis (PCA) of the spectrogram is proposed for dysarthric speech analysis. The interpretation of the first principal component of frame-wise obtained spectra in terms of the LTAS has been discussed in [8]. Each principal component of a spectrogram is a weighted sum of all frame spectra. Thus, the LTAS, defined as the average of equally weighted per frame amplitude-spectra, may be related to the principal components of the spectrogram. Indeed, when the weights of the summed spectra of a PC of the spectrogram all have the same sign then that PC is a generalised LTAS. This is the case for the first principal component, for instance. When the weights of the summed spectra of a PC are positive and negative, then that PC is a difference between two generalised LTASs, one assigned to the positive weigths and the other to the negative weights. This is the case for the second principal component, for instance.

The remainder of the presentation is organized as follows. The corpus used to demonstrate the analysis method is described in Section II. Principal component analysis of the spectrogram is presented in Section III. Results are presented in Section IV. Finally, conclusions are given in Section V.

II. CORPUS

The corpus comprises Colombian Spanish monologues produced by 25 male and 25 female Parkinson speakers and 25 male and 25 female control speakers [9]. The age of the male and female Parkinson speakers respectively ranges from 33 to 81 years

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

(average 61.5 ± 11.6 years) and from 49 to 75 years (average 60.7 ± 7.2 years) and the age of the male and female control speakers respectively ranges from 31 to 86 years (average 60.3 ± 11.5 years) and from 49 to 76 years (average 61.4 ± 6.9 years). The duration since diagnosis for male and female Parkinson speakers ranges from 0.4 to 20 years (average 8.8 ± 5.8 years) and from one to 43 years (average 12.5 ± 11.5 years). The duration of the stimuli ranges from 16.8 to 164 s (average 48.4 ± 28.8 s) for the control speakers and from 14.1 to 110.9 s (average 45.7 ± 23.6 s) for Parkinson speakers.

The spectral analysis has been carried out on the voiced phonetic segments after removing the unvoiced segments by automatic voiced-unvoiced detection. The total lengths of the voiced stimuli range for the control speakers from 6.9 to 76.9 s (22.7±14.3 s) and for the Parkinson speakers from 5.3 to 58.5 s (20.8±10.9 s).

III. METHODS

Let $\boldsymbol{X}=[\boldsymbol{X}_1\ \boldsymbol{X}_2\ \ldots\ \boldsymbol{X}_M]$ be a KxM spectrogram matrix, where the column index of the matrix denotes the frame number (variables) and the row index stands for frequency points (observations). The amplitude-spectrogram may involve thousands of variables that are the successive spectral frames. Principal component analysis aims at representing the spectrogram by a few linear combinations of the column data, which report the salient properties of the full data matrix [10].

The first principal component is obtained as a linear combination of the vectors $\boldsymbol{X}_j$, j=1,. . ., M

$$\boldsymbol{z}_1=\sum_{j=1}^{M}u_{j1}\boldsymbol{X}_j \tag{1}$$

having maximum variance. The elements of the vector $\boldsymbol{u}_1=[u_{11}\ \ u_{21}\ \ \cdots\ \ u_{M1}]^T$ are the coefficients of the linear combination.

The k^{th} principal component $\mathbf{z}_k$, k=2, . . ., M is computed as a linear combination of the vectors $\boldsymbol{X}_j$, j=1,. . ., M

$$\boldsymbol{z}_k=\sum_{j=1}^{M}u_{jk}\boldsymbol{X}_j \tag{2}$$

having maximum variance and being uncorrelated with the principal components $\mathbf{z}_1$, $\mathbf{z}_2$, . . ., $\mathbf{z}_{k-1}$. The elements of the vector $\boldsymbol{u}_k=[u_{1k}\ \ u_{2k}\ \ \cdots\ \ u_{Mk}]^T$ are the coefficients of the kth linear combination.

The principal components of the amplitude-spectrogram are obtained via the eigendecomposition of the covariance matrix. Element (i,j) of that matrix is the covariance of the ith and jth spectral frames. Covariance matrix $\boldsymbol{C}$ is a MxM symmetric matrix that can be rewritten as follows.

$$\boldsymbol{C}=\boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^{\mathrm{T}} \tag{3}$$

In decomposition (3), $\boldsymbol{A}$ is a matrix whose columns $\boldsymbol{u}_k$, k=1, 2, . . ., M are eigenvectors of $\boldsymbol{C}$, and $\boldsymbol{D}$ is a diagonal matrix that reports eigenvalues λ_j in decreasing order.

The representation of the spectrogram in the principal component coordinates is $\boldsymbol{Z}=\boldsymbol{X}\boldsymbol{A}$. The kth principal component of the spectrogram is a linear combination of the M variables (i.e. spectral frames), where the coefficients of the linear combination are the components of the k^{th} eigenvector $\boldsymbol{u}_k$ of covariance matrix C.

$$\boldsymbol{z}_k=\boldsymbol{X}\boldsymbol{u}_k,\ k=1,\cdots,M \tag{4}$$

Using only the first few principal components as an alternative coordinate system may reduce the dimensionality of the spectrogram.

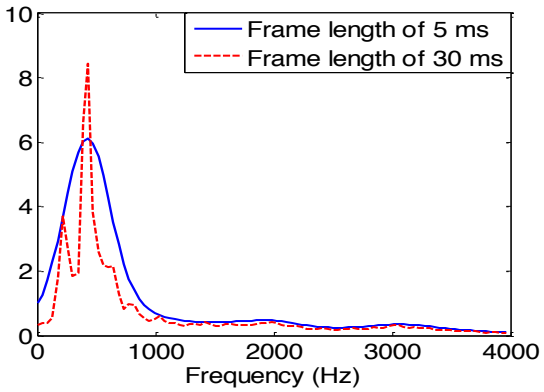
IV. RESULTS AND DISCUSSION

Principal component analysis has been carried out on the spectrogram of the monologue produced by a control speaker. To investigate the effect of the frame length, a short and long frame of 5 ms and 30 ms is used. The short frame enables to carry out a broadband analysis focusing on the transfer function of the vocal system (spectral slope of the voice source included), while a long frame carries out a narrowband analysis, which enables tracking harmonics. A Hamming window is used with an overlap of 50%. The first and second principal components of the spectrogram for two frame lengths are shown in Fig.1. For a frame length of 5 ms, the first and second principal components account for more than 90% of the total variability of spectra. When the frame length is set to 30 ms, the percentage of variability explained by the first and second principal components decreases to 63.9%. The description of the spectrogram by the first and second principal components only is more accurate if the frame length is fixed to 5 ms. Indeed, a short frame length produces less details along the frequency axis. The correlations between the spectral amplitudes are therefore larger in frequency and time and a larger percentage of the variability of the data is explained by the two first PCs.

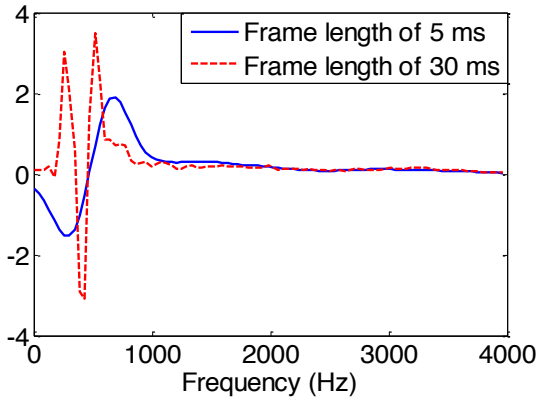
Then, principal component analysis of the spectrogram has been carried out on all the monologues produced by 50 control and 50 Parkinson speakers. Each monologue is long enough to average out the effects of individual phonetic segments.

Also, spontaneous speech may be more altered in Parkinson's disease than other speaking tasks. As a consequence, acoustic features obtained from the monologue may be more relevant for the assessment of articulatory deficits [8].

Because we wish to focus on articulatory deficits in dysarthric speech, the partials of the vocal source are discarded by carrying out a broadband fast Fourier analysis using a short analysis frame of 5 ms. The broad-band amplitude-spectrogram can be represented compactly by the two first principal components, which explain more than 90% of the total spectral variability. The first and second principal components averaged over all speakers are shown in Fig. 2. One observes that the principal components of the control speakers are positioned to the right of the Parkinson speakers.

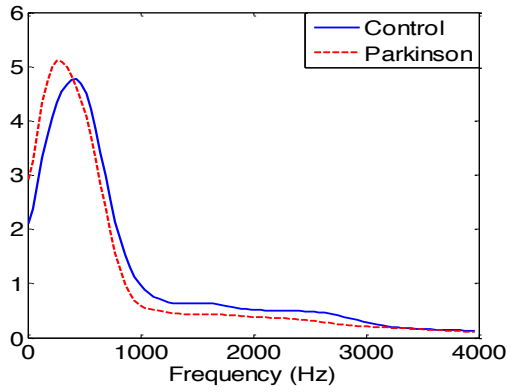


First principal component

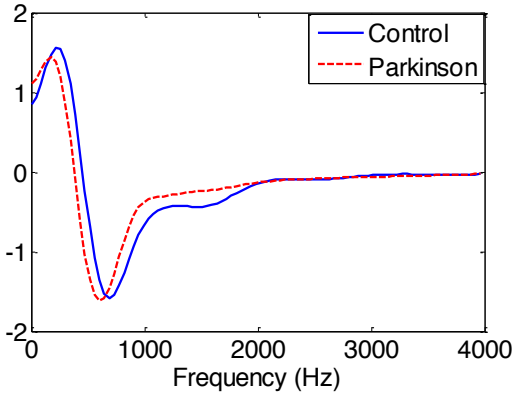


Second principal component

Fig. 1: First and second principal components of the amplitude-spectrogram of a monologue of a control speaker for frame lengths of 5 ms and 30 ms.



First principal component

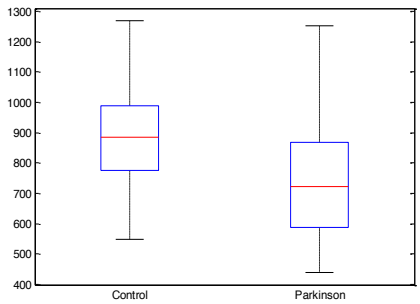


Second principal component

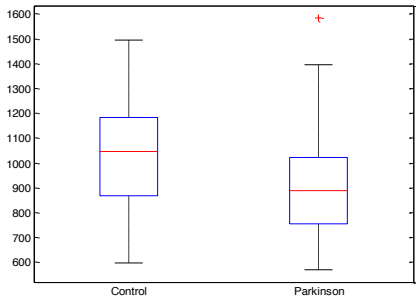
Fig. 2: Averaged first and second principal components of the amplitude-spectrograms of the monologues produced by all control and Parkinson speakers.

In contrast to the weights associated with PC1 (i.e. the elements of vector $\boldsymbol{u}_1$) that have been observed to be all positive, PC2 coefficients (i.e. the elements of vector $\boldsymbol{u}_2$) are positive as well as negative. The second principal component has therefore been decomposed into two weighted sums of frame spectra associated with positive and negative coefficients respectively.

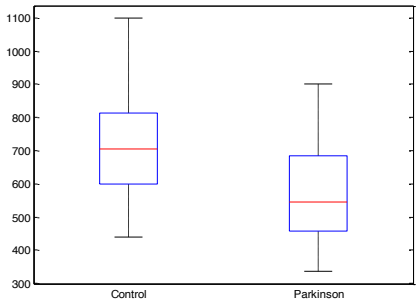
Hereafter, the spectral means of the first principal component of the spectrogram and of the weighted sums of frame spectra associated with positive and negative PC2 coefficients have been used as descriptors to report the differences between control and Parkinson speakers. The quartiles of the spectral means in the frequency range 0-4000 Hz are shown as boxplots in Fig. 3. Generally speaking, Parkinson speakers are characterized by smaller spectral means than control speakers. Two-tailed t-tests show that the averages of the three spectral means differ statistically significantly ($p<0.05$) for control and Parkinson speakers.



First principal component



Weighted sum of frame spectra with positive PC2 coefficients



Weighted sum of frame spectra with negative PC2 coefficients

Fig. 3: Boxplot of the spectral means (in Hz) of the first principal component and of the weighted sums of frame spectra with positive/negative PC2 coefficients of the monologues produced by control and Parkinson speakers.

V. Conclusion

Principal component analysis of the amplitude-spectrograms of speech may be used to document differences between control and Parkinson speakers. The first and second principal components account for a large percentage of the observed spectral variability when the analysis is broadband. The acoustic markers are the spectral means of the first principal component and of the weighted sums of frame spectra associated with positive and negative coefficients of the second principal component. Experimental results show statistical significant differences between Parkinson and control speakers.

Acknowledgments

The corpus has been provided by the research group GITA of the Universidad de Antioquia, Medellín, Colombia.

References

[1] S. Skodda, W. Gronheit, and U. Schlegel, "Intonation and speech rate in Parkinson's disease: General and dynamic aspects and responsiveness to levodopa admission,"*J. Voice*, vol. 25, pp. 199–205, 2011.

[2] R. D. Kent, "Acoustic studies of dysarthric speech : methods, progress, and potential," *J. Commun. Disord*, vol. 32, pp. 141-186, 1999.

[3] J. R. Orozco-Arroyave et al., "Automatic detection of Parkinson's disease in running speech spoken in three different languages," *J. Acoust. Soc. Am.*, 139(1), 481-500, 2016.

[4] J. Rusz, R. Cmejla, and T. Tykalova, "Imprecise vowel articulation as a potential early marker of Parkinson's disease: Effect of speaking task," J. Acoust. Soc. Am., vol. 134, pp. 2171-2181, 2013.

[5] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease," J. Acoust. Soc. Am. vol. 134, pp. 350-367.

[6] C. Dromey, "Spectral measures and perceptual ratings of hypokinetic dysarthria," J. Med. Speech Lang. Path., vol. 11, pp. 85-94, 2003.

[7] K. Tjaden., J. E. Sussman, G. Liu, and G. Wilding, "Long-term average spectral (LTAS) measures of dysarthria and their relationship to perceived severity," J. Med. Speech Lang. Path., vol. 18, 125-132, 2010.

[8] J. L. Blanco and J. Schoentgen, "Vocal tract setting in speakers with obstructive sleep apnea syndrome," in : Proc. Maveba 2013, Florence, Italy, Dec. 2013, pp. 211-214.

[9] J. R. Orozco-Arroyave et al., "New Speech Corpus Database for the Analysis of People Suffering From Parkinson's Disease," in Proc. Int. Conf. Lang. Resources and Evaluation (lrec), pp. 342-347, 2014.

[10] I. T. Jolliffe, Principal component analysis, 2nd ed., Springer, 2002.

USE OF ACOUSTIC LANDMARKS AND GMM-UBM BLEND IN THE AUTOMATIC DETECTION OF PARKINSON'S DISEASE

L. Moro-Velazquez[1*], J. I. Godino-Llorente [1, 3], J. A. Gómez-García[1], J. Villalba[2], S. Shattuck-Hufnagel[3], N. Dehak[2]

[1]Centro de Tecnología Biomédica, Universidad Politécnica de Madrid, Madrid, España
[2]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, USA
[3]Speech Communication Group, Research Laboratory of Electronics, MIT, Boston, USA.
laureano.moro@upm.es, igodino@ics.upm.es, jorge.gomez.garcia@upm.es, jvillal7@jhu.edu, sshuf@mit.edu, ndehak3@jhu.edu

Abstract: **New tools based on speech analysis can improve and accelerate diagnosis of Parkinson's Disease. In this work, the use some specific segments of speech, around the so called Acoustic Landmarks, are used with different families of features such as acoustic cues or Rasta-PLP and GMM-UBM-Blend classification methods to detect Parkinson's Disease. Results of 87% of accuracy are obtained.**

Burst segments provide the most relevant information when detecting Parkinson's Disease while GMM-UBM-Blend is revealed as a promising technique when using small databases and segmented speech.

Keywords: **Parkinson's Disease, GMM-UBM, Acoustic Landmarks, Rasta-PLP.**

I. INTRODUCTION

Diagnosis of Parkinson's Disease (PD) is a challenging task which might require several years, depending on the patient. New tools based on motor analysis such as speech analysis can provide the means to do a more rapid and robust diagnosis.

Literature reports multiple efforts to detect and assess PD using voice and speech. These works can be classified as phonatory, articulatory, prosodic and linguistic, depending on the type of material employed and the analyzed speech/voice features.

In the present work, which can be framed into the articulatory group, the detection of PD is performed employing some specific points of speech, called Acoustic Landmarks [1], which are detected along several speech tasks. Some acoustic measurements associated to these landmarks (acoustic cues), or Rasta-Perceptual Linear Predictive (Rasta-PLP) features calculated over several time windows around the landmarks, are used to detect the presence of PD, employing GMM and GMM-UBM Blend classification techniques in two different databases.

Acoustic Landmarks were first defined by Stevens in [1] as "*a discrete representation of the speech stream in terms of a sequence of segments, each of which is described by a set (or bundle) of binary distinctive features*". These landmarks can be determined following the procedure described in [2], in which their detection is mainly based upon the analysis of the energy changes in six frequency bands. In this study, three types of landmarks are considered: b-Lmk, which are related to bursts during articulation; g-Lmk, coinciding with the beginning or ending of vocal fold vibration; and s-Lmk which mark the transitions between vowels and sonorant consonants or vice-versa. Fig. 1 shows the spectrogram and acoustic landmarks of a normal voice during a diadochokinetik (DDK) test.

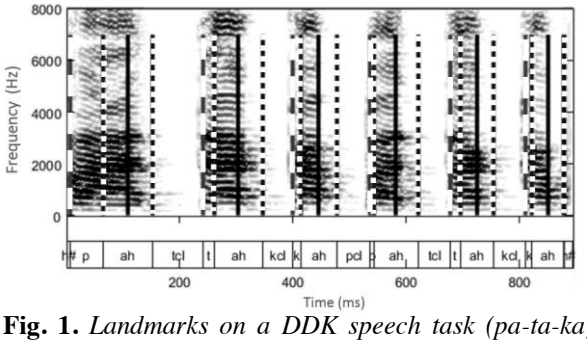


Fig. 1. *Landmarks on a DDK speech task (pa-ta-ka). Dashed lines represent b-lmk, dotted lines are related to g-lmk (pointing the beginning and end of vocal fold vibration). Additionally, black continuous lines mark the Vowel landmark, in the middle of the two g-lmk.*

The acoustic landmarks have already been employed in [3] for PD detection where these are used as a mean to characterize prosody. Other works like [4], [5] do not specifically utilize acoustic landmarks but employ particular segments of speech to characterize and detect parkinsonian speakers.

In this work, three different approaches for the automatic detection of PD are assessed. On each one, a different family of features characterizing speech and classification scheme are considered.

II. METHODS

Overview: The main objective of this study is to automatically detect PD using some articulatory-related features which are introduced in a classification scheme. The families of features can be acoustic cues, probability of a candidate (PoC) or Rasta-PLP coefficients. The classification schemes can be GMM + Logistic regression, GMM-Blend and GMM-UBM-Blend. Some combinations of families of features and classification schemes are performed aiming to obtain the highest accuracies.

* orcid.org/0000-0002-3033-7005

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

Features: On this study three families of features are used. The first one is integrated by the acoustic cues, defined in [2]. Each landmark type (b-Lmk, g-Lmk and s-Lmk) has several associated specific acoustic cues which consist on some measurements over the speech signal around the acoustic landmark. For b-Lmk, acoustic cues are abruptness (i.e., difference of energy level between two points separated by a certain time window) and silence (i.e., energy level on both sides of the landmark). For g-Lmk, acoustic cues are abruptness and vocalic level (again, energy level on both sides of the landmark). For s-Lmk, the acoustic cues used in this work are abruptness and energy statistics (as mean, maximum, minimum and tilt).

The second family of features is the PoC. As not all the landmarks detected by the algorithm are true landmarks, the acoustic cues of each candidate of being a landmark are introduced in a statistical model trained with the TIMIT database as explained in [2]. After this, it is possible to calculate the probability of a candidate of being a true landmark, obtaining the PoC values.

The third family of features is Rasta-PLP. On this work, these last features are extracted along the whole signal or are calculated only in three overlapped time windows located around each specific landmark, as represented in Fig. 2. In these last cases, the rest of the signal is discarded. Thus, these features are called Lmk-based Rasta-PLP and can be related to b-Lmk, g-Lmk or s-Lmk, depending on the landmark around which these features are extracted.

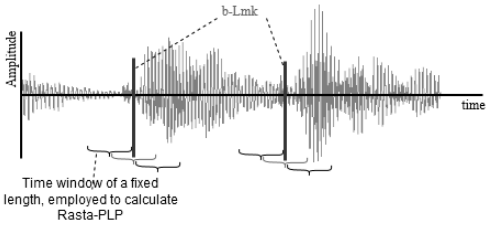


Fig. 2. *Selection of three time windows around b-Lmk to calculate Lmk-based Rasta-PLP*.

Databases: Three databases are used in this study: The first one is the GITA database [6], which includes speech from 50 parkinsonian and 50 control Colombian speakers. A DDK task (/pa-ta-ka/) and two sentences are selected in this work from all the available materials, being the two sentences: Sentence 1: *"Los libros nuevos no caben en la mesa de la oficina"*; and Sentence 2: *"Luisa Rey compra el colchón duro que tanto le gusta"*.This database is used to train and test different classification models as it is proposed in the methodology. The second database is the Neurovoz corpus which is employed for validation purposes and contains DDK tasks (/pa-ta-ka/) of 46 and 26 speakers in the parkinsonian and control groups respectively. The third database consists on the first corpus of the Albayzin database [7] which is used to create the UBMs to train the GMM-UBM models.

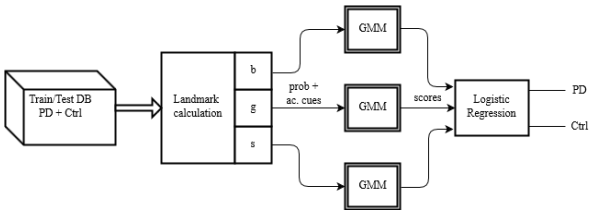


Fig. 3. *First approach for PD detection*.

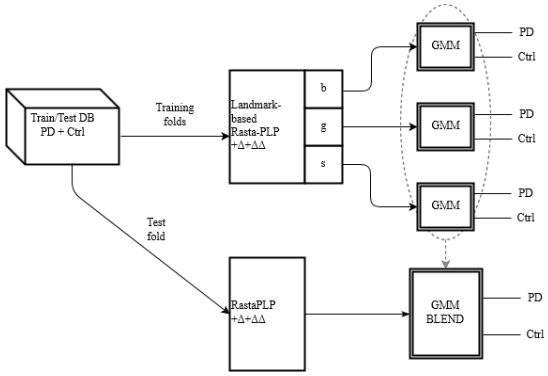


Fig. 4. *Second approach for PD detection*.

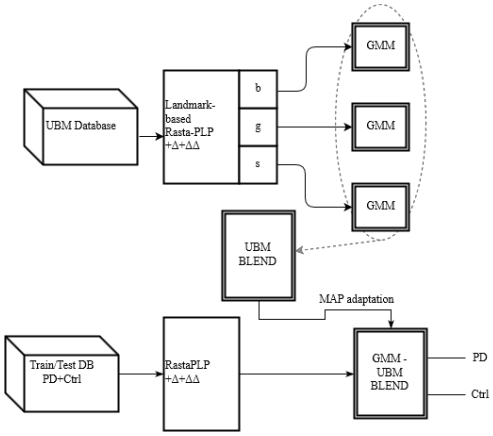


Fig. 5. *Third approach for PD detection*.

Methodology: Firstly, a preliminary analysis of acoustic cues and PoC using DDK utterances from GITA database is performed, to evaluate the statistical behavior of these families of features and their separability on the parkinsonian and control classes. Then, three different approaches are considered depending on the used features and the classification scheme. All the training-testing iterations on each approach follow a k-folds validation scheme with k=7. On the first approach one different GMM classifier is trained and tested for each acoustic landmark type, namely b-Lmk, g-Lmk and s-Lmk, using acoustic cues and PoC joined in a feature vector for each landmark point. Therefore, three global scores are obtained per speaker, one for each type of landmark. These three

global scores are fused following a logistic regression scheme in order to classify the speaker as parkinsonian or control using the equal error rate as threshold. The diagram of this stage is depicted in Fig. 3. In the second approach, features are Lmk-based Rasta-PLP plus derivatives (Δ+ΔΔ) obtained employing windows of 15 ms with 50% overlapping and. On this second approach, three different GMM are trained too, one for each landmark type. Then, the three resulting GMM are blended into a new GMM which is tested with using Rasta-PLP+Δ+ΔΔ features from the testing fold at each iteration of the cross-validation. The GMM blending consists on the creation of a model containing all the Gaussians of the three original models and the weightings of these pondered by a factor. On this study, this factor is always 1/3. Fig. 4 depicts this approach. On the third approach, Lmk-based Rasta-PLP+Δ+ΔΔ extracted from the Albayzin database are employed to obtain three different UBMs, one for each landmark type. Then, these three UBMs are blended into one and this is readjusted into a new GMM using MAP adaptation and Rasta-PLP features from the utterances included the training folds. A diagram of this approach is presented in Fig. 5. This third approach is repeated using Rasta-PLP+Δ+ΔΔ to characterize UBM database (and, therefore, avoiding segmentation and the GMM-UBM-Blend) in order to compare results obtained with and without the use of landmark-based segmentation. The three approaches are achieved using the DDK task from the GITA database. Additionally, the third approach is repeated using the two sentences from this database and the DDK utterances from Neurovoz database. In all three approaches, the number of Gaussians of the GMM models are varied in the range [4, 8, 16, 32, 64, 128] while the number of Rasta-PLP coefficients is 12.

III. Results

Results regarding the preliminary study and the three approaches are included in this section. Only results leading to the highest accuracy are included.

Fig. 6 shows the boxplots of some acoustic cues and PoC associated to the three types of landmarks. For the sake of simplicity, some of the acoustic cues are not referred.

The accuracy, confidence interval (CI), area under the curve (AUC), specificity and sensitivity obtained on the three different approaches are included in tables 1, 2 and 3.

Table 1. Best results on first approach

Lmk	*Accu. (%)*	*CI (%)*	*AUC*	*Spec.*	*Sens.*
b	**80**	**±8**	**0,83**	**0,82**	**0,78**
g	70	±9	0,79	0,70	0,70
s	75	±8	0,81	0,76	0,74
Fusion	77	±8	0,84	0,70	0,84

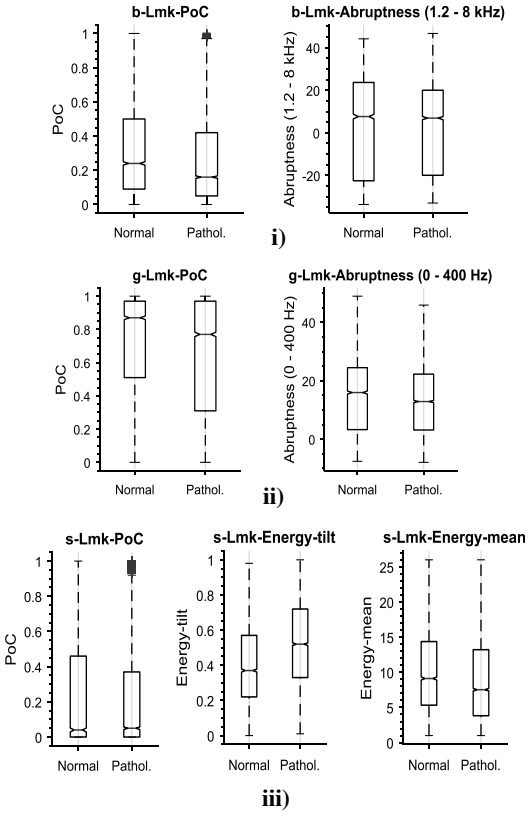


Fig. 6. *Boxplots of PoC and acoustic cues for b (i), g (ii) and s (iii) landmarks. All values, except by PoC are expressed in dB.*

Table 2. Best results on second approach

Lmk	*Accu. (%)*	*CI (%)*	*AUC*	*Spec.*	*Sens.*
none	76	±8	0,8	0,74	0,78
b	74	±9	0,8	0,74	0,74
g	75	±8	0,81	0,74	0,76
s	74	±9	0,82	0,68	0,8
All	**78**	**±8**	**0,85**	**0,78**	**0,78**

Results are referred to DDK task of GITA database in all cases unless otherwise specified. Specifically, other speech tasks and databases are used additionally in the third approach.

Table 3. Best results on third approach

Database	*Speech task*	*Lmk-based & GMM-Blend*	*Accu. (%)*	*CI (%)*	*AUC*	*Specif.*	*Sensit.*
GITA	DDK	Yes	82	±8	0,87	0,82	0,82
		No	75	±8	0,82	0,72	0,78
GITA	Sentence 1	Yes	82	±8	0,88	0,91	0,71
		No	80	±8	0,88	0,90	0,70
GITA	**Sentence 2**	**Yes**	**87**	**±7**	**0,91**	**0,92**	**0,82**
		No	78	±8	0,88	0,92	0,64
Neurovoz	DDK	Yes	82	±9	0,89	0,77	0,85
		No	78	±10	0,84	0,69	0,83

IV. Discussion

Preliminary results, as shown in Fig. 6, reveal that acoustic cues for the three types of landmarks have a different statistical distribution in the two classes, especially in the case of Energy tilt and Energy mean for s-Lmk and PoC for b-Lmk and g-Lmk. Although the used speech tasks in this case (DDK) do not include s-Landmarks (/pa-ta-ka/ only contains b-Lmk and g-Lmk), these are used for the rest of the study as in many occasions, s-Lmk candidates are detected, especially for parkinsonian speakers. This can be caused by the motor perturbation associated to articulation that many PD patients suffer in which some burst or plosive consonants become sonorant. This sign may be a consequence of the reduction of the articulation ranges. That is the reason why s-Lmk and its acoustic cues provide considerable outcomes, as it can be inferred from tables 1 and 2.

Regarding the first approach, acoustic cues + PoC extracted from b-Landmarks provide the best results, with 80% of accuracy. Fusion of scores of the three types of landmarks does not result in better accuracies as it can be inferred from Table 1. Table 2 shows the results for the second approach, where the use of Lmk-based Rasta-PLP+Δ+ΔΔ provides lightly lower accuracies (75% for g-Lmk) than in the case in which all speech frames are used (76%) while the GMM-Blend, considering the models of the three types of landmarks, provides the best results of this second approach (78%). Finally, Table 3 shows the results of the GMM-UBM-Blend approach using different types of speech materials in the train/test databases. The accuracy employing the DDK task is higher in this third approach than in the rest (reaching 82%) while best results are obtained using Sentence 2 (87%) where a relative improvement of 11% is achieved with respect to the non Lmk-based segmentation and non GMM-UBM-Blend scenario. On this approach, the obtained specificity and sensitivity repeating the methodology with the Neurovoz database are comparable to those obtained with GITA database. Therefore, this approach seems to be appropriate to be used in voice pathology detection schemes in which the databases are relatively small and some parts of the speech are more relevant for detection than others. In future works, new studies based on the use of specific segments such as plosive or fricative consonants should consider the GMM-UBM-Blend technique. It has been observed that the landmark detection techniques detect much more candidates than the true number of landmarks present in a sequence and, therefore, more precise techniques such as forced alignment [8] might be employed in the future to detect specific segments.

V. Conclusion

In this work, several approaches for the detection of PD using speech have been analyzed. From all the proposed schemes, the use of GMM-UBM-Blend provides the best results, 87% of accuracy, employing Lmk-based Rasta-PLP to train the UBM model and Rasta-PLP when performing the MAP adaptation. Results evidence that the use of GMM-UBM-Blend techniques with acoustic landmark segmentation in the UBM database provide better results than just GMM-UBM typical use. The employment of the acoustic landmark segmentation for the training of GMM models along with Rasta-PLP coefficients, provides relative improvements up to 11% respect non-segmentation scenarios and must be considered for future works.

Acknowledgements

This work was supported by the grants EEBB-I-17-12092 and BES-2013-062984 within project TEC2012-38630-C04-01, RR01/2011, PRX15/00385, *XV Ayudas Consejo Social-UPM* and *Ayudas EEBB para PDI-UPM*, with special thanks to the Fulbright Foundation.

References

[1] K. N. Stevens, “Toward a model for lexical access based on acoustic landmarks and distinctive features,” *J. Acoust. Soc. Am.*, 2002.

[2] C. Park et al., “Automatically determining acoustic landmark sequences using physiological constraints,” in *ICASSP*, 2008.

[3] Huici, H. D. et al., “Speech rate estimation in disordered speech based on spectral landmark detection”. *Biomedical Signal Processing and Control*, 2016.

[4] J. R. Orozco-Arroyave et al., “Voiced/unvoiced transitions in speech as a potential bio-marker to detect Parkinson’s disease,” in *INTERSPEECH*, 2015.

[5] Novotný, M. et al., “Automatic evaluation of articulatory disorders in Parkinson's disease.” IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 2014.

[6] J. R. Orozco-Arroyave et al., “New Spanish speech corpus database for the analysis of people suffering from Parkinson’s disease,” Proc. Ninth Int. Conf. Lang. Resour. Eval., 2014.

[7] A. Moreno, D. Poch et al., “Albayzín speech database: Design of the phonetic corpus,” Eurospeech 1993. Proc. 3rd Eur. Conf. Speech Commun. Technol., 1993.

[8] Moreno, P. J. et al., "A factor automaton approach for the forced alignment of long speech recordings." ICASSP 2009.

VARIABILITY OF THE FUNDAMENTAL FREQUENCY OF PARKINSONIAN VOICES IN READ SPEECH: A TRANSVERSAL STUDY

P. Rodríguez-Pérez[1], R. Fraile[1], M. García-Escrig[2,3], N. Sáenz-Lechón[1], J.M. Gutiérrez-Arriola[1], V. Osma-Ruiz[1]

[1] CITSEM, Universidad Politécnica de Madrid, Madrid, Spain
[2] Hospital de Sagunto, Sagunto (Valencia), Spain
[3] Department of Medicine, Universidad CEU Cardenal Herrera, Castellón, Spain

pablorodriper@gmail.com, rfraile@ics.upm.es, miguel.garcia1@uchceu.es, nslechon@ics.upm.es
jmga@ics.upm.es, vosma@ics.upm.es

Abstract: **A transversal study of the pitch variability of parkinsonian voices in read speech is presented. 30 Parkinson's disease (PD) patients and 6 healthy speakers were recorded while reading a text without voiceless phonemes. The following measures were obtained from the fundamental frequency contours: mean, minimum, maximum, and standard deviation. These measures, as well as a parameter describing the form of the modulation spectrum, were investigated for correlation with age and PD stage evaluated using the Hoehn and Yahr scale. Results indicate that the influence of PD on intonation can be masked by the effects of aging. This is the case for the male voices herein analyzed. The study of the modulation spectrum of the fundamental frequency can provide some insight into the ability of the speakers to plan the intonation of full phrases. For the female population some significant correlations are found between parameters obtained from this modulation spectrum and the PD stage.**

Keywords: Parkinson's disease, Voice analysis, Fundamental frequency, Modulation spectrum

I. INTRODUCTION

Parkison's disease (PD) is the second most usual neurodegenerative disease, after Alzheimer's disease [1], and one of the most usual movement disorders in patients above 50 years old, with only about 4% of PD patients having developed clinical signs of the disease before that age [2]. Between 70% and 89% of PD patients report vocal difficulties [3], being hypophonia one of the most widely recognized [4].

Apart from hypophonia, disordered prosody probably is the most relevant speech impairment for PD patients [5]. This includes monotony [4], speech rate abnormalities [6], difficulties in initiating speech and finding words [4], and syllable repetition [7]. Regarding intonation, the key difference between PD patients and age-matched healthy speakers seems to be the narrowing of the pitch range in PD patients [8]: lowering of the highest fundamental frequency (f_o) in both sexes, and elevation of the lowest f_o in males.

In this paper, we present a transversal study in which the pitch variability corresponding to a text read by 30 PD patients and 6 healthy speakers is analyzed. Specifically, the following measurements the evolution of the fundamental frequency are calculated: mean, minimum, maximum, and standard deviation. In addition, the spectrum of the fundamental frequency evolution is calculated and described by a ratio of its energy for modulation frequencies below 3 Hz to its energy in the 0-50 Hz interval. All these measures are investigated for correlation with age and PD stage, the latter being evaluated by means of the Hoehn and Yahr (H&Y) scale [9]. The analysis of the modulation frequency of pitch is novel, since pitch variations have been mainly described in terms of range until now [8], but not in terms of the velocity of variation (i.e. modulation frecuency).

II. MATERIALS

30 outpatients (19 men, 11 women) of the Neurology Service at the Hospital de Sagunto were recorded between April and November 2015. The recording protocol was approved by the ethics committee of the Hospital and all participating patients signed informed consent before being recorded. They were recorded in a quiet room within the hospital after seeing the neurologist. Before recording, the neurologist collected a data sheet for each patient, including information on age, sex, years since PD was diagnosed for the first time, and illness evolution stage according to the H&Y scale.

The equipment used for recording consisted of a microphone, a mixer and a personal computer (PC). A lavalier microphone was chosen in order to maintain the distance between mouth and microphone as fixed as possible, while avoiding the stress that head mounted microphones might cause in the patients.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

Specifically, a Fonestar FCM-410 was selected due to its bandwidth: 30 Hz to 18,000 Hz. A Fonestar SM-303SC mixer was used for amplifying the microphone signal and directing it to the USB port of the PC. The opensource software Audacity was run in the PC to manage analogue-to-digital conversion. This was performed at 44,100 samples per second and 16 bits per sample.

All patients were requested to read a phrase containing only voiced phonemes, including the five Spanish vowels /*a*, *e*, *i*, *o*, *u*/. They were asked to phonate at comfortable pitch, pace and intensity. 6 volunteers (4 men, 2 women) with ages in the same range as those of patients were recorded with the same equipment in similar conditions (quiet rooms in either home or university environment). These speakers were assigned the H&Y label 0.

III. METHODS

The recorded voice signals were band-pass filtered to discard all spectral energy outside the microphone bandwidth (30 Hz to 18,000 Hz). This filtering was performed using the discrete Fourier transform (DFT) with previous zero padding. The fundamental frequency contour of each filtered signal was estimated using the YIN algorithm [10]. Afterwards, all estimates were manually revised and corrected by visually comparing them to the first harmonic of the spectrogram.

The YIN algorithm provides the estimated contour of the fundamental frequency $f_o[n]$ at a sampling rate equal to the sampling rate of the signal divided by 32 (approx. 1,378 Hz). A null value was assigned to the fundamental frequency for unvoiced samples. The following parameters were extracted from $f_o[n]$:

- Mean value (μ_{fo}):

$$\mu_{fo} = \frac{1}{N}\sum_n f_o[n] \tag{1}$$

where N is the number of non-null samples of $f_o[n]$:

$$N = \sum_n v[n]\,;\;\; v[n] = \begin{cases} 1 \text{ if } f_o[n] \neq 0 \\ 0 \text{ if } f_o[n] = 0 \end{cases} \tag{2}$$

- Minimum (f_{omin}):

$$f_{omin} = \min_{v[n]=1} \{f_o[n]\} \tag{3}$$

- Maximum (f_{omax}):

$$f_{omax} = \min_{v[n]=1} \{f_o[n]\} \tag{4}$$

- Standard deviation (σ_{fo}):

$$\sigma_{fo} = \sqrt{\frac{\sum_n v[n]\left(f_o[n] - \mu_{fo}\right)^2}{N}} \tag{5}$$

- Normalised standard deviation ($\frac{\sigma_{fo}}{\mu_{fo}}$).

The modulation spectrum of each fundamental frequency contour was calculated as the DFT of its autocorrelation function $\rho_{fo}[m]$:

$$\rho_{fo}[m] = \frac{\sum_n v[n]v[n+m]\widehat{f_o}[n]\widehat{f_o}[n+m]}{\sigma_f^2 N} \tag{6}$$

where $\widehat{f_o}[n] = f_o[n] - \mu_{fo}$.

The square modulus of the modulation spectrum was processed to obtain the ratio of its integral between 0 and 3 Hz to the integral between 0 and 50 Hz (low frequency energy ratio – LFER).

Correlations between the aforementioned parameters and patients' age and H&Y labels were evaluated using the Spearman coefficient ρ_s. Due to the limited number of samples, this non-parametric measure of correlation was preferred to the Pearson coefficient. The *p*-values of the measured correlations were also evaluated using a non-parametric approach based on analyzing correlations in random permutations of data [11, chap.5].

IV. RESULTS

A. Fundamental frequency

No significant correlations were found between μ_{fo} and H&Y labels. A moderate correlation was found between μ_{fo} and age for males (ρ_s=0.45; p<0.05). This trend is more related to an increase in the highest pitch values f_{omax} (ρ_s=0.54; p<0.01) than to an overall positive shift in the frequency range, since no significant correlations were found for f_{omin}. On the contrary, a significant negative correlation was found between f_{omax} and H&Y labels (ρ_s=-0.52; p<0.05) for females (Fig. 1).

B. Pitch range

A significant correlation of moderate value was found between pitch range σ_{fo} and age for men (ρ_s=0.54; p<0.01), while no significant correlations were found between pitch range and H&Y labels for men. In contrast, significant correlations were found for women between H&Y labels and pitch range

(Fig. 2), both absolute σ_{fo} (ρ_s=-0.66; p<0.01) and relative $\frac{\sigma_{fo}}{\mu_{fo}}$ (ρ_s=-0.55; p<0.05).

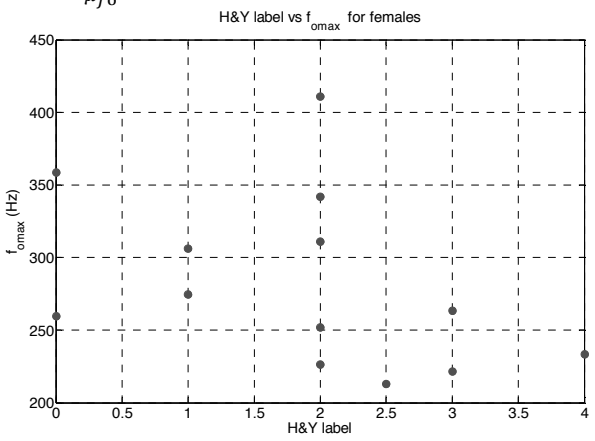


Fig. 1. Scatter plot of f_{omax} vs H&Y labels for females.

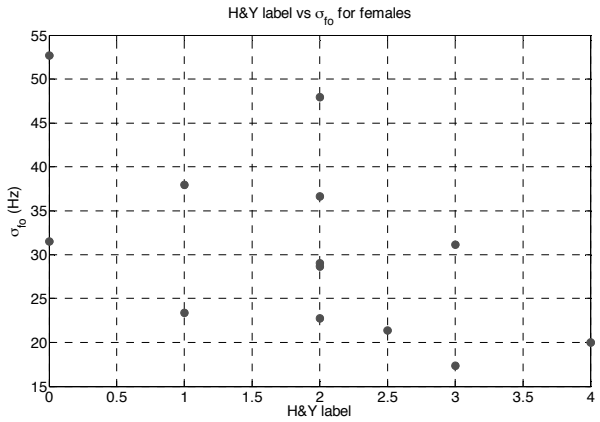


Fig. 2. Scatter plot of σ_{fo} vs H&Y labels for females.

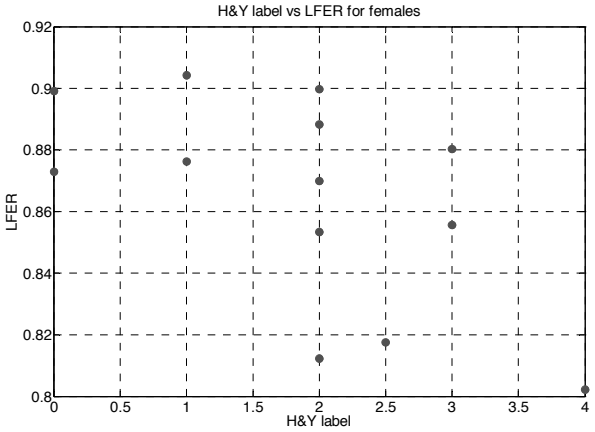


Fig. 3. Scatter plot of LFER vs H&Y labels for females.

C. Modulation spectrum

As described before, the LFER measures the relevance of f_o modulations slower than 3 Hz with respect to all modulations in the 0-50 Hz band. A significant correlation between this parameter and H&Y labels was found for women (ρ_s=-0.54; p<0.05) (Fig. 3).

V. DISCUSSION

Previously published results have indicated that the main effect of PD on voice intonation is the narrowing of the pitch range (lowering of the highest f_o in both sexes and elevation of the lowest f_o in males) [8]. However, since PD mainly affects the population over 50 years old, this effect is likely to interact with the effects of aging on voice. For males, pitch tends to increase as a function of age [12]. This explains the correlations between μ_{fo} and age (ρ_s=0.45; p<0.05) and between f_{omax} and age (ρ_s=0.54; p<0.01). Thus, the effect of aging seems to dominate over the effect of PD for this population. In contrast, the effect of PD in lowering f_{omax} seems to be dominant for women, since correlation between f_{omax} and H&Y label was greater in absolute value (ρ_s=-0.52; p<0.05) and more significant than between f_{omax} and age.

Regarding σ_{fo}, it tends to increase with age in the case of men [13], which is consistent with the correlation reported before (ρ_s=0.54; p<0.01). No effect of PD was detected for this population either. In the case of women, however, a significant reduction of σ_{fo} with H&Y label was measured (ρ_s=-0.66; p<0.01), which is opposed to the general influence of age on σ_{fo} [13]. Consequently, PD seems to be the dominant factor affecting both f_{omax} and σ_{fo} for this female population.

Significantly, the reduction of pitch range for the female population is accompanied by an increase in the average modulation rate of f_o, or decrease in the LFER parameter defined before. This implies that the fundamental frequency changes less but faster as PD progresses. This may be related with a reduction in the ability to plan and execute intonation for longer time intervals (phrases lasting a few seconds).

Fig. 4 illustrates this effect. It shows the fundamental frequency contours corresponding to two female patients with H&Y labels 1 and 4, and LFER values equal to 0.90 and 0.80, respectively. In both cases, four segments can be clearly identified in the contour. While for the upper plot (LFER = 0.90) there is a smooth transition in the fundamental frequency between the first two segments, the same segments seem to be more independent in the lower plot (LFER = 0.80). Similarly, for the last segment, the dynamics

of the fundamental frequency are fairly well described by a curve with frequency components below 3 Hz for the upper plot. This does not happen for the lower plot, where intonation does not seem to have the long-term component that is apparent in the upper plot.

VI. Conclusions

The analysis of the fundamental frequency at which 30 PD patients plus 6 controls read a text with only voiced phonemes in Spanish indicates that the influence of PD on intonation can be masked by the effects of aging. This seems to be the case for the male population herein studied. In contrast, the dependence of f_{omax} and σ_{fo} with the PD progression for the female population is consistent with previous results reported in the literature.

The dynamics of the fundamental frequency has been studied not only in terms of its range, but also in terms of its modulation frequencies. According to the results shown in this paper, this analysis is potentially significant for studying the effect of PD on prosody, since it allows detecting the relevance of the slow components of pitch evolution, which are related to the ability to plan intonation in the long term.

ACKNOWLEDGMENTS

This work has been funded by the Spanish government through project grants TEC2012-38630-C04-01 and MAT2015-64139-C4-3-R

REFERENCES

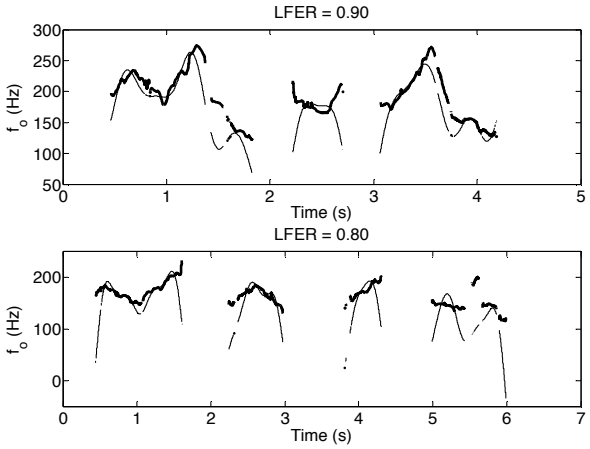


Fig. 4. Fundamental frequency contours (thick lines) and their components with modulation frequencies up to 3 Hz (thin lines) for two female voices.

[1] K. Wirdefeldt, H.O. Adami, P. Cole, D. Trichopoulos, and J. Mandel, "Epidemiology and etiology of Parkinson's disease: A review of the evidence," *Eur J Epidemiol*, vol. 26, 2011.

[2] L.M. De Lau, and M.M. Breteler, "Epidemiology of Parkinson's disease," *Lancet Neurol*, vol. 5, pp. 525-535, 2006.

[3] A.L. Merati, Y.D. Heman-Ackah, M. Abaza, K.W. Altman, L. Sulica, and S. Belamowicz, S. "Common movement disorders affecting the larynx: A report from the neurolaryngology committee of the AAO-HNS," *Otolaryng Head Neck*, vol. 133, pp. 654-665, 2005.

[4] J. Jankovic, "Parkinson's disease: Clinical features and diagnosis," *J Neurol Neurosur Ps*, vol. 79, pp. 368-376, 2008.

[5] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease," *J Acoust Soc Am*, vol. 129, pp. 350-367, 2011.

[6] A. Bandini, F. Giovannelli, S. Orlandi, S.D. Barbagallo, M. Cincotta, P. Vanni, R. Chiaramonti, A. Borgheresi, G. Zaccara, and C. Manfredi, "Automatic identification of dysprosody in idiopathic Parkinson's disease," *Biomed Signal Proces*, vol. 17, pp. 47-54, 2015.

[7] S. Skodda, "Steadiness of syllable repetition in early motor stages of Parkinson's disease," *Biomed Signal Proces*, vol. 17, pp. 55-59, 2015.

[8] Y. Ikui, H. Nakamura, D. Sano, H. Hyakusoku, H. Kishida, Y. Kudo, H. Joki, S. Koyano, A. Yamauchi, S- Takano, N. Tayama, H. Hirose, N. Oridate, and N. Tayama, "An aerodynamic study of phonations in patients with Parkinson Disease (PD)," *J Voice*, vol. 29, pp. 273-280, 2015.

[9] M.M. Hoehn, and M.D. Yahr, "Parkinsonism: Onset, progression, and mortality," *Neurology*, vol. 17, pp. 427–442, 1967.

[10] A. de Cheveigné, and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J Acoust Soc Am*, vol. 111, pp. 1917-1930, 2002.

[11] J.J. Higgins, *Introduction to Modern Nonparametric Statistics*, Brooks Cole, 2004.

[12] S. Schötz, and C. Müller, "A study of acoustic correlates of speaker age," In *Speaker Classification II: Selected Projects*, C. Müller Ed. Berlin: Springer, 2007, pp. 1-9.

[13] S. Schötz, "Acoustic analysis of adult speaker age," In *Speaker Classification I: Fundamentals, Features, and Methods*, C. Müller Ed. Berlin: Springer, 2007, pp. 88-107.

ARTICULATION DYNAMICS IN PARKINSON DYSARTHRIA

P. Gómez[1], J. Mekyska[2], A. Gómez[1], D. Palacios[1], V. Rodellar[1], A. Álvarez[1]
[1]NeuVox Lab, Center for Biomedical Technology, Universidad Politécnica de Madrid
Campus de Montegancedo, s/n, 28223, Pozuelo de Alarcón, Madrid, Spain
pedro@fi.upm.es
[2]Department of Telecommunications, Brno University of Technology
Technicka 12, 61600 Brno, Czech Republic

Abstract: **The Vowel Space Area (VSA) and the Formant Centralization Ratio (FCR) have been proposed to describe dysarthria in Parkinson Disease (PD) as well as in other neuromotor diseases affecting speech. These features are based in global estimations of the positions of the first two formants in the representation of a vowel triangle. The aim of the paper is to give a description of speech articulation dynamics as a probability density function of the kinematic features derived from the evolution of formants in the time domain. The statistical distribution of the dynamic behaviour of articulation features can be used to estimate differences between speech features from subjects with Parkinson dysarthria relative to normative subjects. Utterances of vowels [a:, i:, u:] from a subset of 16 subjects with PD (8 males and 8 females), confronted to a subset of 16 normative subjects (8 males and 8 females) have shown that the statistical distributions of dynamic articulation features can be differentiated using information theory based estimations such as Kullback-Leibler's Divergence (KLD). These estimations allow establishing relevant statistical differences between PD and normative subjects both for males and females, well over the differentiation capability of VSA and FCR.**

Keywords: **Neuromotor diseases, speech processing, articulation biomechanics, Kullback-Leibler Divergence, speech kinematics.**

I. INTRODUCTION

Parkinson Disease (PD) is a sickness produced by a deficit of the neurotransmitter dopamine in basal ganglia, resulting in hampered neuromotor activity. As a consequence, it also interferes with speech capability in different ways, which have been extensively well documented [8]. Rough and asthenic phonation, monotonicity, mono-loudness, freezing, velo-pharyngeal incompetence, and low tone, are some of the observed alterations of speech coined under the term of hypokinetic dysarthria [9]. Illness progress is evaluated by neurologists using scales as Hoehn &Yahr [6] or UPDRS [3], although these scales have not been specifically designed for speech or phonation assessment. PD articulation has been characterized using features as Vowel Space Area (VSA) or Formant Centralization Ratio (FCR) [10]. These measurements are of static nature, because they estimate the state of the articulation limits as an average of formant frequency limits. Having into account that PD affects strongly the dynamics of normal movement, it could be possible that a description of hampered articulation, supported by features estimated from speech in terms of the dynamic changes experimented by the resonant frequencies of the vocal tract could give a more vivid description of articulation behaviour. The aim of the present study is to evaluate to which extent dynamic features can be used in the multimodal study of PD speech production. Initially, dynamic estimates of formant activity, as the absolute kinematic velocity (AKV) which are highly correlated with the superficial myoelectric activity of certain facial muscles [4], seem to be the adequate candidates for such study. The structure of the present paper is as follows: the biomechanical foundations explaining distortion of vowel articulation in terms of formant dynamics are exposed in section II. Section III is devoted to describe the fundamentals of the experimental setup (materials and methods). The results derived from the present work are shown and discussed in section IV. Conclusions are given in section V.

II. BIOMECHANICAL FOUNDATIONS

The present study is focussed on the dynamic tracking of the kinematic activity of the jaw-tongue reference point (JTRP), which may be defined as a hypothetical point in the sagittal plane (x: caudal-rostral; y: dorsal-ventral). As seen in Fig. 1 this is a hypothetical point $\{x_r, y_r\}$ where the sum of forces is null (masseter: f_m, stylo-glossus and genio-hyoglossus: f_{sg} and f_{gh}, genio-glossus: f_{gi}, and the gravity: f_w). The AKM is integrated by the jaw (J) and tongue (T) and the facial tissues attached to them. The dynamics of this system [5] may be approximated by a third-order lever fixed at the skull in (F), articulating movements on the sagittal plane (x, y).

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

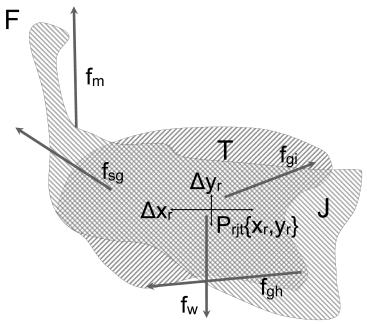


Figure 1. Jaw-Tongue Articulation Kinematic Model (AKM) considered in the study.

The position of the JTRP will change in time under the action of the forces mentioned, modifying the resonant properties of the oral cavity, and producing dynamic changes in formants [2]. The work hypothesis considers that the changes of the first two formants f_1 and f_2 can be related to the AKM dynamics as by:

$$\begin{bmatrix} v_x \\ v_y \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \begin{bmatrix} \frac{df_1(t)}{dt} \\ \frac{df_2(t)}{dt} \end{bmatrix} \quad (1)$$

where w_{ij} are the parameters relating JTRP kinematics described by the velocity estimates of the caudal-rostral (v_x) and dorsal-ventral velocities (v_y), with formant dynamics. It is also hypothesized that v_x will be mostly related to changes in the second formant f_2 (back-front), and that v_y will be related to the dynamics of the first formant f_1 (up-down), or in other words $w_{11}\approx 0$ and $w_{22}\approx 0$. Therefore, the AKV of the reference point (RP) may be stated as:

$$|v_{RP}(t)| = \sqrt{\left(w_{12}\frac{df_1(t)}{dt}\right)^2 + \left(w_{21}\frac{df_2(t)}{dt}\right)^2} \quad (2)$$

Reliable estimates for w_{12} and w_{21} may be obtained from articulations involving changes in the positions of the reference point showing predictable dynamic changes. A very relevant feature to describe articulation dynamics can be defined from the probability distribution of the AKV in (2), directly estimated as its normalized amplitude histogram over bins between 0 and 50 cm.s^{-1} as:

$$p(|v_{RP}|) = \frac{hist(|v_{RP}|)}{\sum hist(|v_{RP}|)} \quad (3)$$

where $hist(|v_{RP}|)$ is the histogram in amplitude counts of the AKV. This feature has proven to be quite relevant in separating dysarthric from normative speech as will be explained in the sequel.

III. MATERIALS AND METHODS

The database of normative and pathological speech used is a part of the Parkinsonian Speech Database (PARCZ) recorded at St. Anne's University Hospital in Brno, the Czech Republic [8], consisting of four sets of 5 Czech vowels ([a:, e:, i:, o:, u:]) pronounced in 4 different ways: short and long vowels uttered in a natural way, long vowels uttered with maximum loudness, and long vowels pronounced with minimum loudness, but not whispering. The recordings selected corresponded to utterances by four subsets of speakers corresponding to eight normative females (NF; average age: 62.25 y; std age: 3.81 y), eight normative males (NM; av.: 63.63 y; std: 7.15 y), eight PD females (PF; av.: 69.25 y; std.: 7.11 y) and eight PD males (PM; av.: 64.88 y; std.: 8.51 y), see Table 1. Recordings of the three vertex vowels at maximum loudness [a: i: u:], sampled at 16 kHz and 16 bits were selected from the database to estimate the logarithm of the VSA (lnVSA) and the FCR. An example of one of these utterances showing the first two formants extracted from LPC spectral estimation is given in Fig. 2 (at the end of the paper). An example of the normalized histogram and cumulative distribution for the same sequence shown in Fig. 2 is given in Fig. 3.

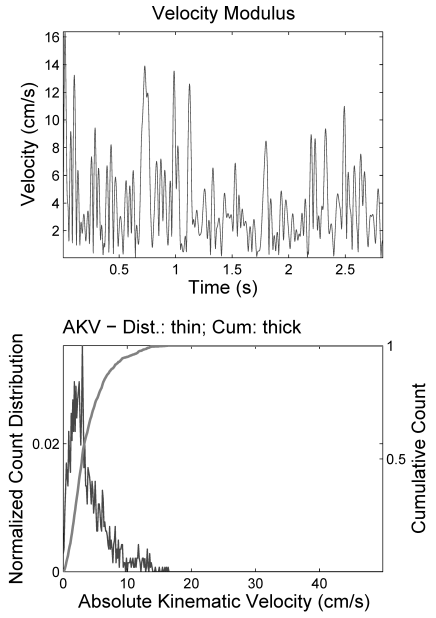


Figure 3. Articulation Kinematic Velocity for the sequence shown in Fig. 2. Top: time series. Bottom: normalized histogram (extending to 17 cm.s^{-1}) in thin line, and its respective cumulative distribution in thick line.

It may be seen that the most active events (larger AKV) are aligned with vowel insertions (start of phonation requiring proprioceptive adjustments, near the origin, around 1.1 s and 2.3 s) or during imperfect vowel emission (between 0.7-1.0 s). The AKV distribution shows a χ^2 behaviour (two degrees of freedom). Its similarity to Maxwell-Boltzmann distributions allows to establish a parallelism with thermodynamic concepts, giving sense to the term "emotional temperature" used by some researchers in the field of neurological deterioration, as in Alzheimer Disease speech studies [7]. The normalized histograms may be interpreted as probability distributions, and these can be applied to estimate the difference in terms of Information Theory [1] between two probability distributions in terms of Kullback-Leibler's Divergence (KLD) as:

$$(4) \qquad D_{KLij}\left\{p_i\left(|v_{RP}|\right), p_j\left(|v_{RP}|\right)\right\} = -\int_{\zeta=0}^{\infty} p_i(\zeta)\log\left[\frac{p_i(\zeta)}{p_j(\zeta)}\right]d\zeta$$

The AKV estimates from (2) and their normalized histograms by velocity bins between 0 and 50 cm.s^{-1} were evaluated from (3). Four sets of normalized histograms were produced respectively for the NM: $\{p_{NM}\}$; NF $\{p_{NF}\}$; PM: $\{p_{PM}\}$; PF: $\{p_{PF}\}$. The KLD between each subject in the pathologic sets PM and PN was estimated with respect to the averages of their respective normative sets, NM and NF.

IV. RESULTS AND DISCUSSION

The results of evaluating lnVSA, FCR and KLD for the PD patients are given in Table 1.

Table 1. Subject set description, static and kinematic estimates. Nxxxx: Normative subjects; Pxxxx: pathologic subjects. UPDRS refers to section III of the rating scale.

Subject	Gender	Age	UPDRS	lnVSA	FCR	KLD
N1003	F	63		13.09	0.92	47.10
N1004	F	65		12.72	0.99	29.89
N1006	F	64		13.30	0.90	18.34
N1007	F	59		13.40	0.81	38.99
N1012	F	67		12.85	0.95	64.54
N1017	F	61		13.30	0.89	35.53
N1018	F	55		13.21	0.91	22.29
N1019	F	64		13.17	0.84	25.30
P1006	F	59	24	12.84	0.90	38.69
P1007	F	76	55	12.85	0.90	76.31
P1008	F	78	23	13.01	0.85	37.87
P1020	F	64	8	12.82	1.03	100.14
P1021	F	65	5	13.33	0.87	63.67
P1022	F	72	6	12.96	0.99	67.75
P1025	F	64	8	13.09	0.85	57.25
P1026	F	76	12	13.00	0.93	44.42
N2001	M	59		12.49	0.83	24.98
N2002	M	68		12.60	0.95	52.77
N2008	M	70		12.74	0.95	19.50
N2009	M	68		12.48	0.93	40.06
N2010	M	73		12.14	1.02	23.37
N2011	M	55		12.62	0.89	34.80
N2013	M	54		12.61	0.97	41.22
N2014	M	62		12.04	1.04	15.88
P2005	M	46	25	12.45	1.29	40.32
P2009	M	66	14	12.46	0.92	53.92
P2010	M	66	39	12.22	1.00	121.42
P2012	M	71	35	12.14	1.03	62.15
P2017	M	71	35	12.88	1.43	34.83
P2018	M	63	19	12.08	1.03	55.47
P2019	M	63	32	12.24	0.91	45.06
P2023	M	73	12	12.14	1.00	77.03

It may be seen that although lnVSA and FCR show some differences between the normative and pathologic sets, these do not show as remarkable differences as in the case of KLD. To appreciate the relevance of these differences two-tail t-tests have been evaluated between each pathological subset and its normative counterpart. The results of the tests based in the identity of the means (H0) and different variances are shown in Table 2.

Table 2. T-tests on the results between normative and pathologic sets.

Feature/Subset	p-value	H0
VSA/Females	0.252	Not rejected
VSA/Males	0.451	Not rejected
FCR/Females	0.885	Not rejected
FCR/Males	0.495	Not rejected
KLD/Females	0.016	Rejected
KLD/Males	0.020	Rejected

It may be seen that neither lnVSA nor FCR are able of distinguishing the normative and the pathologic sets under a significance level of 0.05. On its turn KLD is able of differentiating pathologic cases from normative ones in both gender sets with a clear significance.

V. CONCLUSIONS

These results would avail the importance of dynamic features derived from kinematic variables, as a complement to static features. It is also clear that the formulation of speech dynamics in terms of probability density functions of the kinematic variables allow the use of Information Theory principles to differentiate between dysarthric and normative speech. One of the inconveniences of KLD is its asymmetry. For this reason, other similar metrics are sought with a more balanced behavior. Besides, given the reduced number

of subjects included in the study this methodology is to be tested against a larger database.

REFERENCES

[1] Cover, T. M. and Thomas, J. A., *Elements of information theory*, Wiley, New York, 2006.

[2] Dromey, C., Jang, G. O., Hollis, K. "Assessing correlations between lingual movements and formants", *Speech Communication* Vol. 55, No. 2, 2013, pp. 315-328.

[3] Goetz, C. G., et al., "Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Process, Format, and Clinimetric Testing Plan", *Movement Disorders* Vol. 22, No. 1, 2007, pp. 41-47.

[4] Gómez, P., et al., "Relating Facial Myoelectric Activity to Speech Formants", Lecture Notes on Computer Science Vol. 10338, 2017, pp. 520-530.

[5] Hannam, A. G., Stavness, I., Lloyd, J. E., and Fels, S., "A dynamic model of jaw and hyoid biomechanics during chewing", *J. of Biomechanics* Vol. 41, No. 5, 2008, pp. 1069-1076.

[6] Hoehn, M. M., and Yahr, M. D., "Parkinsonism: onset, progression, and mortality", *Neurology* Vol. 17, No. 6, 1967, 427-442.

[7] Lopez-de-Ipiña, K., et al., "On automatic diagnosis of Alzheimer's disease based on spontaneous speech analysis and emotional temperature", *Cognitive Computation*, Vol. 7, No. 1, 2015, pp. 44-55.

[8] Mekyska, J., et al., "Robust and complex approach of pathological speech signal analysis", *Neurocomputing* Vol. 167, 2015, 94-111.

[9] Sapir, S., "Multiple factors are involved in the dysarthria associated with Parkinson's disease: a review with implications for clinical practice and research", *Journal of Speech, Language, and Hearing Research*, Vol. 57, No. 4, 2014, 1330-1343.

[10] Sapir, S., "Ramig, L. O., Spielman, J. L. and Fox, C., Formant Centralization Ratio: A Proposal for a New Acoustic Measure of Dysarthric Speech", *Journal of Speech, Language and Hearing Research*, Vol. 53, No. 1, 2010, 114-125.

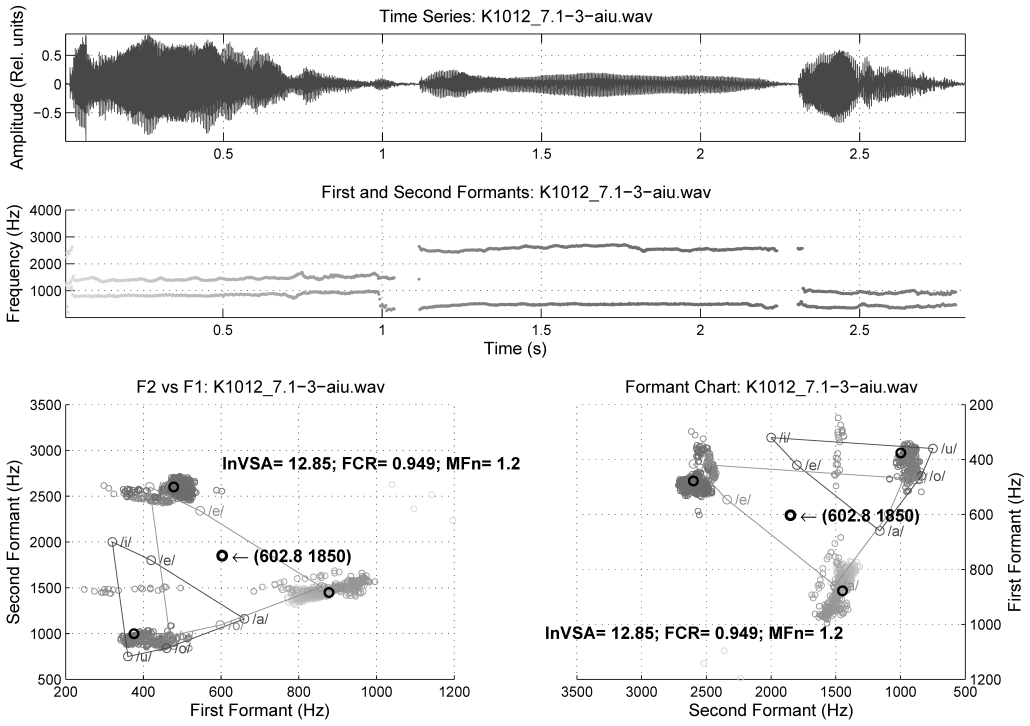


Figure 2. Example of the first two formants extraction from a sequence [a:, i:, u:]. Top: speech signal. Middle: first two formants from LPC spectral estimation. Bottom left and right: formant projection on the vowel triangle. Black circles give the vowel centroids and the vowel triangle centre of gravity to evaluate the lnVSA and the FCR, which are shown superimposed on the vowel triangle.

SESSION V:

VOCAL FOLDS DYNAMICS I

TURBULENCE INTENSITY MEASUREMENT DOWNSTREAM OF THE SELF-OSCILLATING VOCAL FOLDS MODEL

V. Radolf[1], J. Horáček[1], P. Antoš[1]
[1]Institute of Thermomechanics, AS CR, Prague, Czech Republic
radolf@it.cas.cz, jaromirh@it.cas.cz, antos@it.cas.cz

Abstract: **The air-flow fluctuations of the longitudinal velocity were measured by means of hot-wire anemometry downstream of the synthetic three-layer, self-oscillating, life-size vocal fold model. The mean airflow rate and the mean subglottic pressure were kept within physiologically relevant values for a normal human voice production. Resulted magnitudes of the turbulent fluctuation velocity component varied in the range 0.25 – 2.1 m/s and are comparable with those found in literature.**

Keywords : **Glottal jet, hot-wire anemometry, turbulence, vocal folds replica**

I. INTRODUCTION

Vocal folds vibrations are the main precondition for a voice production. The vocal folds, excited by the airflow, generate a primary sound which propagates in the airways of the vocal tract modifying its spectrum and producing the final sound radiated from the mouth. In the analysis [1] performed on a set of data obtained from many dysphonic patients, the authors detected glottal air leakage, resulting in turbulent noise, giving the perceptual impression of breathiness. According to computation modelling in [2], turbulent noise played important roles in the presence of irregular vocal fold vibrations. The purpose of the current study is to measure air-flow velocity fluctuations, which can contribute to the final acoustic signal, downstream of the vocal folds replica.

II. METHODS

Measurements presented in this study were performed with a 1:1 scaled three layer vocal folds (VF) model. Silicon wedge, modelling a vocal fold body, was added inside the vocal fold reducing the space of the liquid layer modelling the lamina propria layer positioned under the thin silicon cover, see [3]. No vocal tract model was included.

The vocal folds were excited by airflow coming from a regulated central pressure supply and the mean airflow rate Q and the mean subglottic pressure P_{sub} were kept in the ranges Q=0.03-0.25 l/s and P_{sub}=0.26-0.8 kPa, i.e. within physiologically relevant values for a normal human voice production, see e.g. [4].

The mean airflow rate was measured by the float flowmeter EMKO type DF3-09K5. The air was flowing through the model of the human lungs to the trachea modelled by a metal tube prolonged by a plexiglas tube (total length 23 cm and inner diameter 18 mm). The mean air subglottal pressure was registered by the digital manometer Greisinger Electronic GDH07AN at the entrance of the airflow to the vocal folds. At the same place fluctuations of the subglottal pressure were measured by the B&K 4138 miniature microphone (range 6.5 Hz - 140 kHz). The vocal fold velocity in the vertical direction was measured by the laser vibrometer Polytec OFV-505.

The air-flow fluctuations of the longitudinal velocity (the turbulence intensity level I_u) in the jet behind the VF were measured by means of hot-wire (HW) anemometry system Dantec Streamline operated in CTA mode. A straight single miniature hot-wire probe was used for HW measurement. The probe was placed 5 mm downstream of the top edge of the vocal folds in its rest position, see Fig. 1. During the VF oscillations this edge moved closer to the probe because of pressure drop at the glottis.

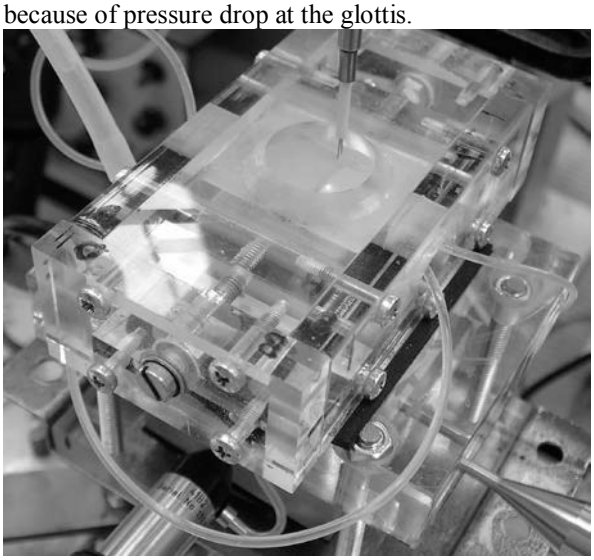


Fig. 1. Detail of the vocal fold model and the hot-wire probe.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

A sensor of the probe has tungsten wire of the diameter 0.005 mm and the length 1.25 mm. Operating wire temperature during measurement was of 493 K. The output anemometer signal was digitalized using the A/D transducer (National Instruments data acquisition system) and recorded in the PC using the LabVIEW scripts (sampling frequency 250 kHz, 16 bit).

The hot-wire probe was calibrated in the calibration rig with variable flow velocity in the range 0.5−50 m/s at operating temperatures within the range of 450−510 K. Measurement uncertainty of the velocity of this range of magnitude in the glottal jet is estimated about 2 percent due to certain flow-temperature variation.

The high-speed CCD camera NanoSense MkIII (maximum resolution 1280x1024 pixels) with a camera zoom lens Nikon AF micro Nikkor 60 mm was included in the measurement set up for analyses of the vocal folds vibration. The recordings were synchronized with the measurement of the pressures.

All measured pressure signals were simultaneously sampled by the frequency of 16.384 kHz and registered by the measurement system Brüel & Kjaer PULSE type 3560 C with Input/Output Controller Modules type 7537A and 3109 controlled by a personal computer equiped by the SW PULSE LabShop Version 10.

III. RESULTS

Examples of the measured time signals for the glottal exit jet velocity V_{jet}, glottal opening GO and sum of the mean and dynamic subglottic pressure are demonstrated in Fig. 2 for mean air-flow rate Q = 0.2 l/s. Mean P_{sub} = 0.7 kPa, maximum GO was 3.2 mm, fundamental frequency of self-sustained vocal folds vibration F_0= 84 Hz and the closed quotient CQ = 20%.

Peaks of the subglottic pressure correlate with the beginning of glottis opening phase. Existence of a short pulse of variable intensity can be detected in the V_{jet} signal about 1.3-1.4 ms delayed after the maximum of the subglottic pressure when glottis opening starts. These narrow peaks of airflow velocity in the beginning phase of the glottis opening may result from a quick rotation of the jet before stabilization of the jet position at maximum GO. Main glottal exit jet maximum about 20 m/s correlates with the maximum of glottal opening.

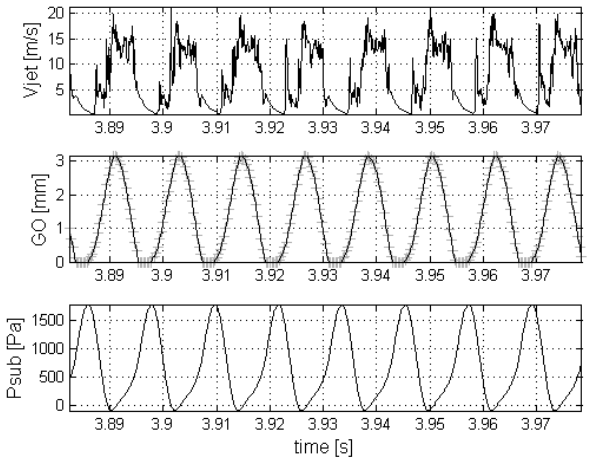


Fig. 2. Time signals for glottal exit jet velocity (top), glottal opening (middle) and subglottic pressure (bottom) for the flow rate Q = 0.2 l/s.

Subglottic pressure signal was periodic with relative jitter less than 0.45 % for all measured flow rates. Mean subglottic pressure increased almost linearly with the mean air-flow rate as depicted in Fig. 3.

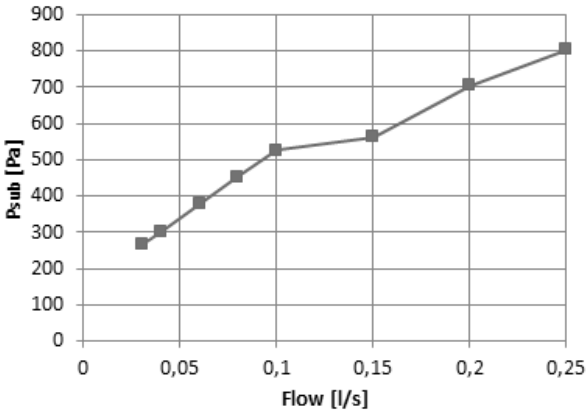


Fig. 3. Mean subglottic pressure in dependence on the mean airflow rate below glottis.

Fundamental frequencies f_0 varied from 78 Hz to 85 Hz. Closed quotient varied in the range 18-27 %. Mean values of f_0 and CQ calculated in time domain from the P_{sub} and GO signals for all preset values of the flow rate are shown in Fig. 4. These two dependences have approximately inverse character.

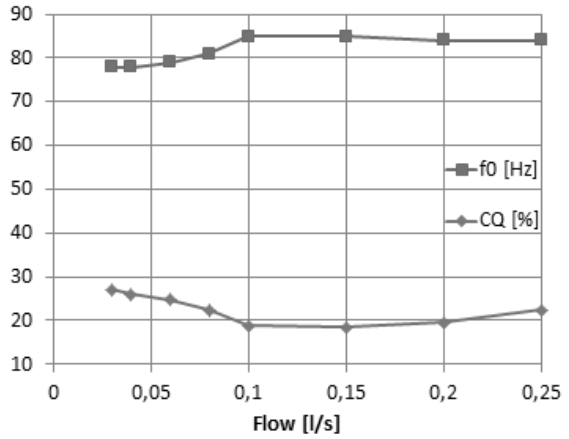


Fig. 4. Fundamental frequency and closed quotient in dependence on the mean airflow rate below glottis.

Turbulent fluctuations of the airflow velocity were estimated in frequency domain by a process analogous to ensemble averaging (also called phase averaging) method. Power spectral density *PSD* of the velocity signal (of the total length 6 s) was obtained by averaging FFT spectra of the raw signal multiplied by Hann-windows of the length 1 second with 75 % overlap; see Fig. 5. The thin spikes corresponding to the energy of periodic components, given by the fundamental frequency f_0 and higher harmonics, were cut off and the resulted *PSD* curve (thick line) was integrated in the range 10 Hz – 20 kHz. Square root of the resulting value gives an estimation of the turbulent fluctuation velocity component. As can be seen in Fig. 6, this component increased linearly with the flow rate, whereas the mean and the effective (root mean square) values of the raw velocity signal show a quadratic dependence on the mean airflow rate Q. Note that maxima of the raw velocity were from 5 m/s up to 47 m/s for the lowest and the highest flow rate, respectively.

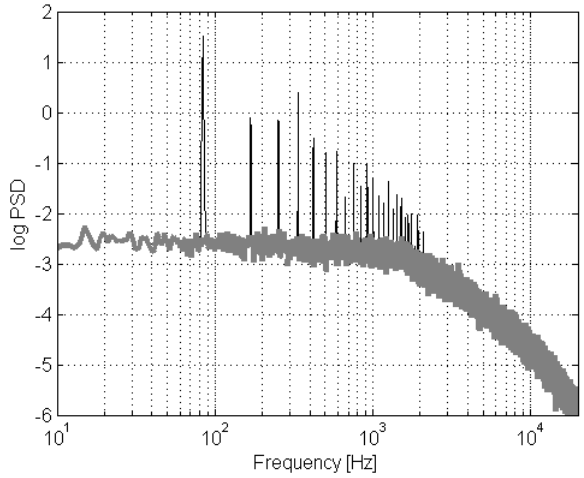


Fig. 5. Power spectral density of the glottal exit jet velocity signal for the flow rate Q = 0.2 l/s

Ratio of the turbulent fluctuation velocity component to the mean velocity of the raw signal is the turbulence intensity level I_u (intensity of fluctuations of longitudinal velocity). This ratio gives the values from 15 % for the highest flow rate up to 44 % for the lowest values of Q.

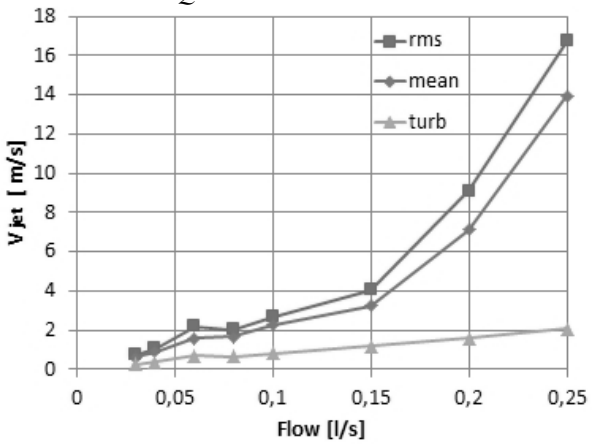


Fig. 6. Effective (root mean square) value, mean value and turbulent fluctuation component of the glottal exit jet velocity signal in dependence on the mean airflow rate below glottis.

For a comparison of the measured airflow velocities in the jet with the velocities of the self-oscillating vocal folds, Fig. 7 shows the maxima and minima of the velocity measured on the right vocal fold by the laser vibrometer in the vertical direction, i.e. in the direction of the airflow jet. The motion of the vocal folds in the opening phase is of about 30 times slower than the motion of the air in the glottal jet, and for the mean flow rates less than 0.15 l/s, the motion of the vocal folds in the vertical direction in the closing phase of the glottis is as fast as in the opening phase.

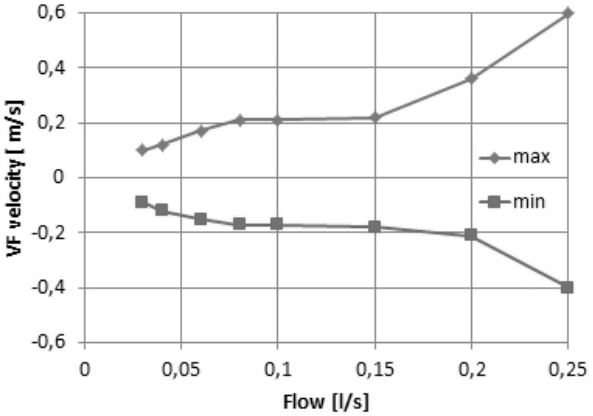


Fig. 7. Maxima and minima of the vocal fold velocity in the direction of the airflow jet in dependence on the mean airflow rate below glottis.

IV. Discussion

A single hot-wire technique is not able to detect a backflow, that could theoretically occur by vocal folds closing or opening. This feature may distort a magnitude of evaluated intensity of fluctuations of longitudinal velocity. Probability of such distortion generally rises for the intensity of fluctuations higher than 30 %.

Resulted magnitudes of the turbulent fluctuation velocity component (0.25 – 2.1 m/s) are a bit smaller than that measured downstream of excised canine larynges (0.6 – 3.5 m/s) and published in [5]. The authors of [5], however, used different technique to obtain the turbulent velocity component. Moreover, both the mean subglottal pressures and flow rates were about 3-4 times higher than in our experiments performed with artificial vocal folds.

V. Conclusion

The air-flow fluctuations of the longitudinal velocity were measured by means of hot-wire anemometry. These fluctuations correspond to the turbulence intensity level if the flow is isotropic. The mean airflow rate and the mean subglottic pressure were kept within physiologically relevant values for a normal human voice production. Resulted magnitudes of the turbulent fluctuation velocity component are comparable with those found in literature as well as with the peak to peak values of velocities of the vocal fold vibrations in the inferior – superior direction.

Acknowledgement

The study was supported by a grant from the Czech Science Foundation: No. 16-01246S "Computational and experimental modelling of self-induced vibrations of vocal folds and influence of their impairments on human voice".

REFERENCES

[1] P.H. Dejonckere, and J. Lebacq, "Acoustic, Perceptual, Aerodynamic and Anatomical Correlations in Voice Pathology," *ORL,* vol. 58, pp. 326-332, 1996.

[2] J.J. Jiang, and Y. Zhang, "Chaotic vibration induced by turbulent noise in a two-mass model of vocal folds," *J. Acoust. Soc. Am.* vol. 112 (5), pp. 2127-2133, 2002.

[3] J. Horáček, V. Bula, J. Košina, and V. Radolf, "Phonation characteristics of self-oscillating vocal folds replica with and without the model of the human vocal tract," in I. Zolotarev, V. Radolf (eds.) Engineering Mechanics 2016. Praha: Ústav termomechaniky AV ČR, v.v.i, pp. 214-217, 2016.

[4] R.J. Baken, and R. Orlikoff, *Clinical Measurement of Speech and Voice, 2*nd ed., Singular, 2000.

[5] F. Alipour, and R.C. Scherer, "Characterizing glottal jet turbulence," *J. Acoust. Soc. Am.* vol. 119 (2), pp. 1063-1073, 2006.

THE DYNAMICS OF VOCAL ONSET

J. Lebacq[1], P. H. DeJonckere [2]

[1]Neurosciences Institute, University of Louvain, Brussels, Belgium
[2]Federal Agency for Occupational Risks, Brussels & Department of Neurosciences KULeuven, University of Leuven, Leuven, Belgium
philippe.dejonckere@kuleuven.be; jean.lebacq@uclouvain.be

Abstract: **Vocal onset is the process occurring between the first detectable glottal movement and the steady state vibration of the vocal folds. High speed film and single line scan, photo-, electro- and flowglottography combined with sound analysis have been used to provide a detailed qualitative and quantitative insight into the phenomenon. Thirty five vocal onsets of different types were analysed, in various loudness and pitch conditions. Vocal fold vibration can start either from a closed glottis (hard onset) or from an open glottis (soft, c.q. breathy onset). In a soft onset, the amplitude of oscillations progressively increases over 2 to more than 30 cycles, before the first clear closed plateau is achieved: it is not possible to define whether the first movement is towards medial or lateral. The ratio "intraglottal pressure during the opening phase / intraglottal pressure during the closing phase" increases during the first free oscillations of the vocal folds (soft onset). Likewise, during these first free oscillations, when all signals are sinusoidal, the phase lag of the glottal area trace relative to the intraglottal pressure trace progressively increases from nearly 0° to nearly 90°.**

Keywords : **vocal onset, glottal attack, intraglottal pressure.**

I. INTRODUCTION

Vocal onset is the process occurring between the first detectable glottal movement and the steady state vibration of the vocal folds. The onset of vocal fold vibration is a dynamic phenomenon with a progressive adjustment of acting forces until a steady state is reached. Three broad categories of vocal onsets (or 'attack') are generally recognized: soft (or 'coordinated'), hard and breathy (or 'aspirate') [1]. To some extent, the voice onset mirrors the voice offset, in which damped oscillations of the vocal folds can be observed, unless the voicing is interrupted by a glottal closure [2]. Combined physiological and imaging techniques can provide a detailed qualitative and quantitative insight into the main aspects of the phenomenon:

° Characteristics of soft, breathy and hard onset
° Time course of oscillation amplitude
° Mechanics of the driving force
° Frequency changes during the first cycles

II. METHODS

The relevant parameters for investigating vocal onset are: acoustic wave, glottal area, transglottal flow, intraglottal pressure and vocal fold contact surface. High speed films and videokymographic recordings provide a global view of the phenomenon, but photoglottography gives the most accurate measure of glottal area. The combined transglottal airflow trace and the glottal area trace allow computing the waveform of the instantaneous intraglottal pressure [3]:

air particle velocity = flow/area, and
intraglottal pressure = - constant x velocity^2.

Thirty five vocal onsets of the different types were analysed in various loudness and pitch conditions. The subject was a healthy trained male vocalist.

Imaging

The pictures below have been taken with the Kay HSV (High Speed Video) model 9700 camera. The larynx is illuminated with a 300 Watt Xenon lamp. The digital HSV signal is sent at 384 Mb/s and a time window of 2 seconds is recorded at 2000 frames per second. Single line scanning of vocal fold (VF) vibrations (videokymography; VKG) is an imaging method based on a special digital camera, fixed onto a rigid 90 ° endoscope. In the high-speed mode, the video camera delivers images from a single line selected in the whole image, at the rate of approximately 7875/7812.5 line-images/s and 720 x 1/768 x 1 pixels resolution, depending on the video format [4]. The selected line is at the level of the mid-portion of the vibrating folds. The resulting high-speed image displays the vibratory pattern of the small selected part of the VF cycle by cycle.

It is also possible to extract several single line scans from a high speed video film [5;6].

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

Glottal area

The glottal area was derived from a photometric record, obtained by transilluminating the trachea. The light flux was detected by a photovoltaic transducer in the pharynx. The transducer, a BP104 Silicon Photodiode (Vishay Precision Group, Malvern, PA), was glued onto a small laryngoscopic mirror (Nr. 3), the handle of which was introduced—together with the sensor lead—through the hermetically sealed hole normally intended for the handpiece of a Rothenberg mask (Fig.1) [3].

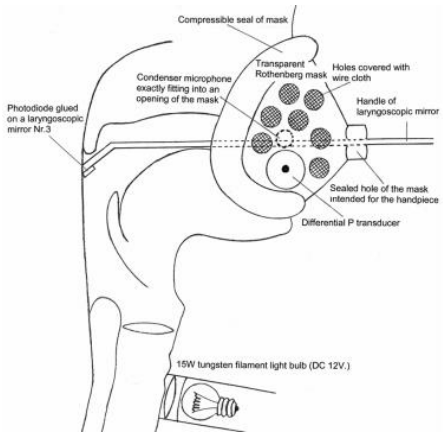


Fig. 1: Combined flow- and photoglottography

The current produced by the photodiode was preamplified by a current-to-voltage converter with a linear response up to 2 kHz. Calibration was described earlier [3].

Transglottal flow

The glottal flow waveform (flowglottogram) was recorded using a Rothenberg mask and the MSIF2 inverse filtering system of Glottal Enterprises (Syracuse, NY). The mask is equipped with a compressible seal and is firmly pressed against the face of the subject to avoid any air leakage. Again, the calibration procedure was described earlier [3]

Translaryngeal electrical impedance

Electroglottography (EGG) [7] measures the transversal transglottic electrical impedance using an AC current at a frequency above 100 kHz and monitors changes in the contact surface of the VF. However, the ability to detect very small transglottic impedance variations (essential in this context) depends on the design of the electronic circuit. Improved devices can show small sinusoidal EGG cycles before a true contact occurs over the full length of the VF [8].

Acoustic signal

A small condenser microphone (Ø 5.6 mm) was fixed laterally inside the Rothenberg mask (Figure 1): it exactly fits in an opening of the mask on the side opposite to the pressure transducer. Processing of the voice samples for SPL analysis was achieved using the *Praat* software (www.praat.org). The microphone sound levels were calibrated with a Wärtsilä 7178 sound level meter in a position corresponding to a direct measurement at 10 cm from the lips.

All signals were recorded using a 4-channels Pico Scope 3403D module (Pico Technology Ltd, St Neots, England, UK) and stored in a PC for later analysis.

III. RESULTS & DISCUSSION

(1) Characteristics of soft, breathy and hard onsets

VF vibration can start either from a closed glottis (hard onset) or from an open glottis (soft, c.q. breathy onset). The most frequently observed type of voice onset in spontaneous speech is the soft onset. A typical example of soft (somewhat breathy) onset at 120 Hz is shown in Fig. 2: the VF oscillation is initiated from a spindle shaped glottis (Fig. 3).

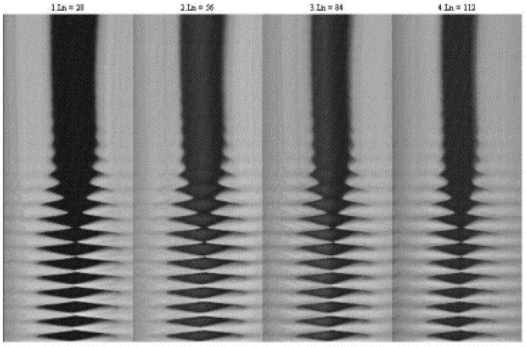


Fig. 2: VKG at 4 levels of the glottis. Time is progressing from top to bottom. /a:/ normal subject.

The amplitude of the oscillations very progressively increases until a first contact occurs between the vocal fold edges (Fig. 2). It is not possible to define whether the first movement is towards medial or lateral. Once a first, very short contact has occurred on the midline, the duration of the closed phase progressively increases.

Fig. 3: Spindle-shaped glottis just before a soft onset

This is more visible in a single line scan at the midpoint of the glottal length (Fig. 4a). The number of cycles of a vocal onset may vary largely. In Fig. 4b (hard onset) the glottis is closed when the first oscillation appears; here again, the duration of the closed phase progressively increases. In cases of hard onset, period irregularities are frequently observed in the first cycles.

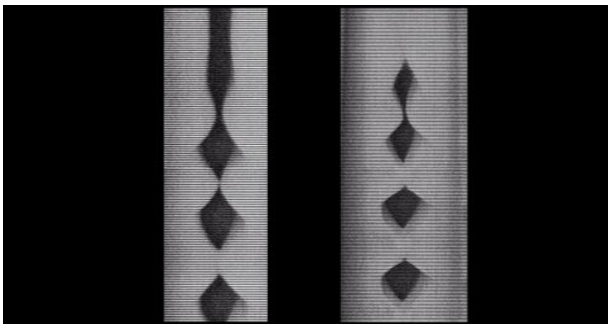


Fig. 4: VKG at midpoint of VF length. Left (a): soft onset; right (b): hard onset.

(2) Time course of oscillation amplitude

In a soft onset, the amplitude of oscillation measured on the photoglottographic signal progressively increases over 2 to more than 30 cycles, before the first clear closed plateau is reached. Plots of the increase of amplitude of glottal area peaks usually show a sigmoid pattern (Fig.5). The pattern is similar for the flow peaks (flowglottogram).

In hard onsets, the amplitude of oscillations also progressively increases, but in general, the number of cycles is smaller. The differences between a soft and a hard onset appear clearly in Figs. 6 & 7. In a soft onset, the first oscillation is detected in the flow trace, immediately followed by the area trace. Changes in electrical impedance (vocal fold contact) occur later (Fig. 6).

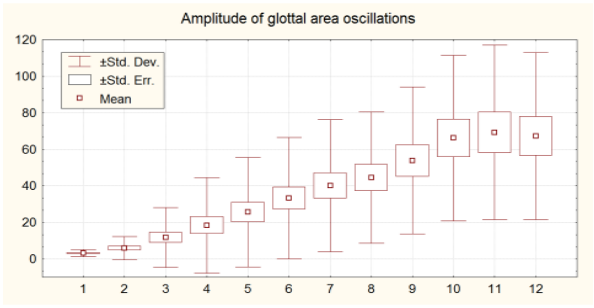


Fig. 5: Amplitude of glottal area oscillations during the first cycles (n = 35) (Arbitrary units,linear).

The first cycles observed in the EGG-signal are sinusoidal, while later on, the shape becomes more differentiated. In a hard onset, the electrical impedance changes first, the downwards movement (impedance decrease) indicating a glottal opening) (Fig. 7). As soon as the first cycle, a closed plateau is present.

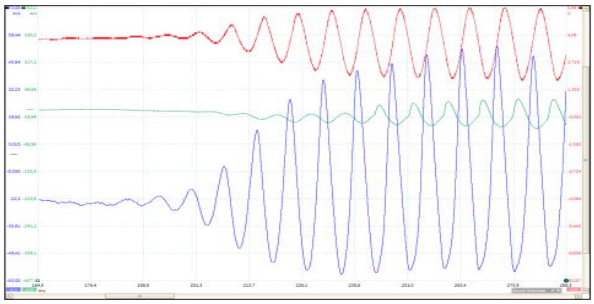


Fig. 6

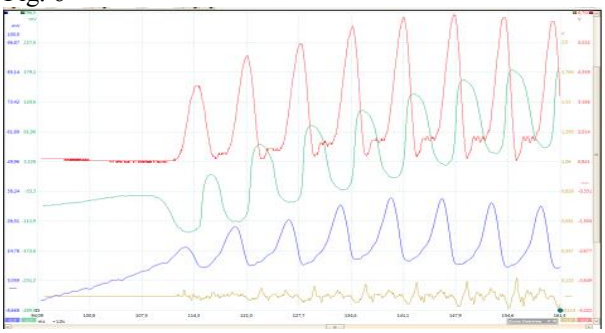


Fig. 7

Figs. 6 & 7: Soft and hard onsets. From top to bottom : flow-, electro- and photoglottograms. In Fig. 7 the sound oscillogram is added (arbitrary units).

It is also notheworthy that, as the amplitude of the oscillations rises, the flow curves become increasingly skewed to the right, up to the value observed during sustained phonation. This is in line with previous observations [3].

(3) Mechanics of the driving force

The intraglottal pressure waveform can be computed from the combined glottal area and airflow records, according to the Bernoulli energy law [3]. The ratio of intraglottal pressure during the opening phase to intraglottal pressure during the closing phase needs to be > 1, so that over one whole cycle, during the first free oscillations of the vocal folds, the pressure performs net work. Fig. 8 shows an example of the increase of the ratio during a soft voice onset with six cycles of free oscillation of the vocal folds before the first closed plateau occurs. This is in line with previous findings obtained in steady state phonation in various conditions of loudness [3]: at soft intensity (minimal closed plateau), the intraglottal pressure ratio is near 1, while it increases to around 6 in loud voicing, when the closed quotient (closed time/period) exceeds 0.5.

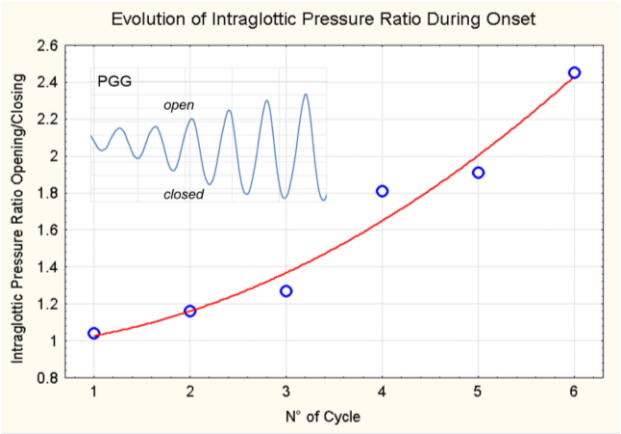


Fig. 8: Example of evolution of intraglottic pressure ratio during the first six cycles in a soft onset.

In steady state vibration, the tissue displacement is expected to show a phase lag of $\pi/2$ radians (in ideal conditions, without friction) with respect to the driving force. Actually, as soon as a closed plateau occurs, all signals undergo substantial distortion from their original sinusoidal shape, masking the phase difference. However, during a soft onset, it can be observed over a few cycles, with a progressive increase of the phase lag from about 0° to about 90° (Fig.9)

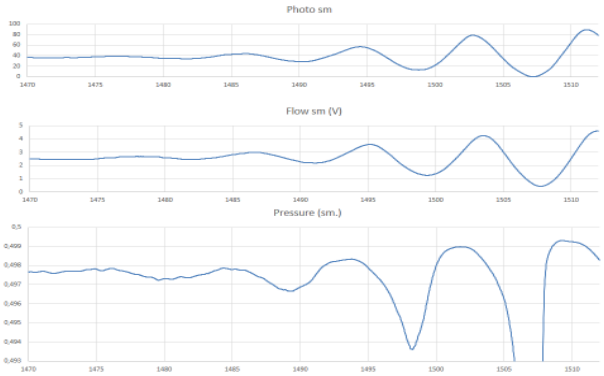


Fig. 9: Top to bottom: Glottal area, flow and computed intraglottal pressure. First cycles of a soft onset.

(4) Evolution of frequency during the first cycles

Fig.10 shows the evolution of cycle duration over the first 17 cycles of a breathy onset, before a closed plateau is reached. There is a clear trend to a slight progressive decrease of the fundamental frequency of the vocal fold oscillation. This seems to point out that, when the vibrating mass is limited to a very thin tissue strip along the vocal fold edge, the vibration frequency is higher than when a more substantial part of the vocal fold mass is involved.

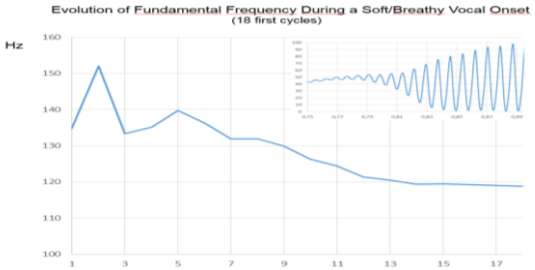


Fig.10: Evolution of Fo over 17 first cycles (soft onset)

IV. CONCLUSION

Similarly to the damping of the vocal fold oscillations at vocal offset, the characteristics of the onset phase of phonation are closely related to the mechanical properties of the vibrating tissues and they escape traditional videostroboscopic laryngoscopy. For clinical applications, an important limitation is the standardisation of the conditions of vocal emission. It may however be assumed that as soon as some pathology enhances stiffness or viscosity, it must affect the vocal onset dynamics.

REFERENCES

[1] R.J. Baken, R.F. Orlikoff, *Clinical measurement of speech and voice*. 2nd ed. San Diego: Singular Publishing Group, 2000.

[2] P.H. DeJonckere J. Lebacq, "Damping of vocal fold oscillation at voice offset". *Biomedical Signal Processing and Control*, vol. 37, pp. 92-99, 2017.

[3] P.H. DeJonckere, J. Lebacq, I. Titze, Dynamics of the driving force during the normal vocal fold vibration cycle. *JVoice* in press, 2017.

[4] J. Svec, H.K. Schutte, Videokymography: high speed line scanning of vocal fold vibration. *J. Voice* vol. 10 : pp. 201-205, 1996.

[5] P.H. DeJonckere, H. Versnel,"High-speed imaging of vocal fold vibration: analysis by four synchronous single-line scans of onset, offset and register break." In *Proc. XVIIIth IFOS World Congress*,D. Passali, Ed. Rome, 2005, pp. 1 – 8.

[6] P.H. DeJonckere, J. Lebacq, L. Bocchi, S. Orlandi C. Manfredi, Automated tracking of quantitative parameters from single line scanning of vocal folds: a case study of the 'messa di voce' exercise, *Logop. Phoniatr. Vocol*. vol. 40: pp. 44–54, 2015.

[7] P.H. DeJonckere, Instrumental methods for assessment of laryngeal phonatory function. In *European Manual of Medicine. Phoniatrics* Vol I. am A. am Zehnhoff-Dinnesen & al. (Eds.) Springer-Verlag Berlin Heidelberg. In press 2017.

[8] J.N. Sarvaiya, P.C. Pandey,V.K. Pandey, An impedance detector for glottography. *IETE Journal of Research* vol. 55: pp. 100-105, 2011.

AN EFFECT OF SOURCE-FILTER INTERACTION ON AMPLITUDES OF SOURCE SPECTRUM PARTIALS

J. Sundberg

KTH, Department of Speech Music Hearing, School of Computer Science and Communication, Stockholm, Sweden
and University College of Music Education Stockholm, Sweden
Email jsu@kth.se

Abstract: **The timbral properties of the voice are partly determined by the voice source, i.e., the pulsating glottal airflow, the properties of which are controlled by the combination of subglottal pressure, glottal adduction and other laryngeal adjustments. Its waveform, the flow glottogram, mainly reflects the amplitudes of the lowest partials. Due to source-filter interaction the lowest formants can affect the periodicity of vocal fold vibration, particularly when the first or second formant coincides with a partial. The aim of the present experimental study was to study associated spectrum effects.**
Glide tones performed by male singers on /ae/ or /a/ were analyzed by inverse filtering, using ripple-free closed phase as criterion. Partials coinciding with the first formant were observed to have amplitudes causing a dip in the source spectrum envelope.
The sound level of a vowel is determined mainly by the strongest spectrum partial, typically the partial closest to the first formant. Glide tones obtained from the formant synthesizer MADDE, which is void of source-filter interaction, showed a much stronger sound level variation with fundamental frequency than the singer subjects. The findings thus seem relevant to the understanding of voice range profiles which show sound level versus fundamental frequency.

I. INTRODUCTION

The timbral properties of the voice are partly determined the vocal tract resonances, or the formants, and partly by the voice source, i.e., the pulsating glottal airflow. The latter is controlled by the combination of subglottal pressure, glottal adduction and length and stiffness of the vocal folds. The effects of subglottal pressure on the glottal flow waveform have been recently analyzed in opera baritone singers and untrained voices. The results showed that a number of flow glottogram parameters could be expressed as functions the maximum flow declination rate MFDR, i.e., the negative peak amplitude of the derivative of the flow glottogram. Thus the AC amplitude, the closed quotient Q_{Closed}, the amplitude quotient AQ, (i.e. the ratio between the AC amplitude and negative peak of the derivative of the flow glottogram, the normalised AQ NAQ (i.e., AQ normalised with respect to period), the level difference between the first and the second voice source partials H1-H2 could all be approximated with equations [1]. The lowest spectrum partials of vocal sounds are mostly strongest and hence they determine the waveform characteristics almost entirely. This means that the equations mentioned provide information mainly about the lower voice source spectrum partials.
Previous research has found that the glottal airflow, and hence the source spectrum can be affected by the vocal tract resonances, the formants. Such source-filter interaction should be particularly strong when the first or second formant (F1, F2) coincides with one of the lowest spectrum partials [2, 3]. Titze and associates have developed this idea extensively and formulated theories that explain the effects on fundamental frequency (F0) control and periodicity [see e.g., 4].
As the flow glottogram mainly reflects the amplitudes of the lowest spectrum partials, even strong effects of source-filter interaction on single source spectrum partials can be difficult or impossible to detect by just examining the shape of the waveform. Yet, such effects can be relevant for voice analysis. The aim of the present study was to study experimentally effects of source-filter interaction on the amplitudes of voice source spectrum partials.

II. METHODS

As shown in previous research, source-filter interaction is likely to occur when a formant frequency coincides with the frequency of one of the lower spectrum partials [4]. Therefore, the phenomenon can best be studied from glide tones performed on a constant vowel.
Inverse filtering is a method for examining the voice source. It eliminates the effect of the vocal tract transfer function on the radiated sound [5]. However, the method becomes unreliable and difficult to use when F0 approaches F1 [6]. Therefore, it is

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

advantageous to analyse the effects of source-filter interaction in adult male subjects.

The effects of source-filter interaction on vocal fold vibration characteristics are well documented. Mostly they concern sudden shifts in F0 and/or in the periodicity of vocal fold vibration. Such effects cannot be accepted in classically trained singer voices. Therefore, to study effects of source-filter interaction, singers would be particularly relevant as subjects.

Three baritone singers served as subjects. They performed pitch glides covering their entire comfortable range, from 110 Hz to 350 Hz, approximately. The glides were performed on the vowels /a/ or /ae/. The subjects were asked to keep the vowel as constant as possible throughout the glide.

The sound was picked up by an omnidirectional condenser microphone (OM1, Line Audio design, Rinkaby, Sweden) and fed to the computer via a Focusrite Scarlett 2i2 (High Wycombe, UK) external soundcard. The microphone was held about 1 cm left of the corner of the mouth so as to avoid disturbances from room reflections, which is particularly important for inverse filtering analysis (Svante Granqvist, personal communication).

Inverse filtering was performed using the Sopran software (Svante Granqvist, KTH), see Figure 1. The frequencies and bandwidths of the inverse formants are set manually, and the associated effects on the waveform and the spectrum are instantly displayed. Filter settings were adjusted on the criterion of a ripple-free closed phase. The moments at which a spectrum partial passed the F1 value, which was used for the inverse filtering, was determined. The amplitudes of source spectrum partials below 1000Hz were measured at this point in time, using the Spectrum subroutine of the Sopran software with a 50 Hz analysis bandwidth. The amplitudes of the same partials were determined also at moments where two adjacent spectrum partials were located symmetrically around F1.

III. RESULTS

The graphs in Fig. 2 show sound level versus frequency for the source spectrum partials at different F0 values that were produced during the glide tones. The value of F1 used for the inverse filtering is represented by the gray bars and the various symbols refer to the indicated F0 values. In all cases a marked spectrum envelope dip can be seen for F0 values that produced a partial that coincided with F1 (solid curves). No dip can be seen for cases where the F0 values did not fulfil this condition (dotted curves). The dips are deeper for lower than for higher F0. They would be caused by source-filter interaction.

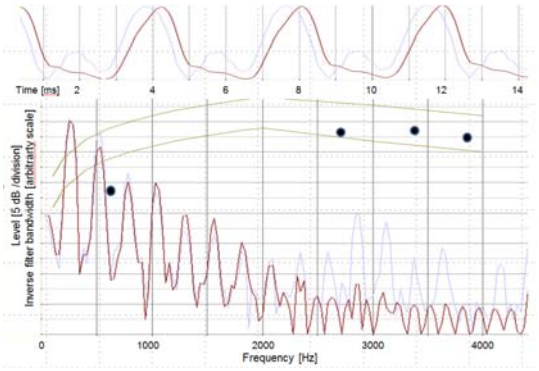


Fig. 1. Inverse filter display of Sopran. In the upper panel the gray and black curves show the waveform before and after inverse filtering. In the lower panel, the gray and black spectra show the associated spectra, the black dots the frequencies and bandwidths of the inverse filters and the gray curve typical formant bandwidth values.

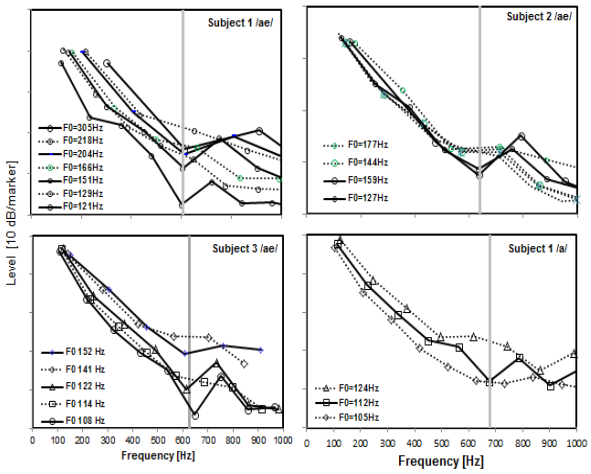


Fig. 2. Levels of the source spectrum partials below 1000 Hz, plotted as function of their frequencies observed for the indicated subjects performing the F0 glides on the indicated vowels. The gray bars show the F1 values used for the inverse filtering. The symbols refer to the F0 values listed in the left lower corner of each panel.

Phonetograms, or voice range profile show overall sound pressure level (SPL) versus F0. As mentioned, SPL is typically determined by the amplitude of the strongest partial in the radiated spectrum, which mostly is the partial that lies closest to F1 in frequency. Therefore, SPL should increase sharply when a partial approaches F1 in an ascending F0 glide and then sharply decrease, as the partial moves away from F1 after having passed it.

The curves in Fig. 3 illustrates this. The dashed curves were derived from synthesizing F0 glides on the Madde synthesizer (Svante Granqvist, KTH), using the formant and bandwidth values that were used for two of the glide tones shown in Fig. 2. In both cases the curves showed a peak when a partial passed F1. The increase with F0 for the valleys between these formant peaks was about 4 dB/octave, which was similar to the subjects' data, and approximately 6 dB/octave for the peaks. When the second partial approached F1 a marked peak occurred, the level increase from valley to peak being almost 10 dB. In the case of subject 2 (left panel) extra peaks occurred also when a partial passed F2 (1122Hz).

The solid curves in the graphs show the corresponding data for the two subjects. The increase caused when a partial passed F1 were much smaller than in the synthesis. In the case of subject 3 (right panel) shallow valleys appeared in the middle of the formant peaks for the synthesis.

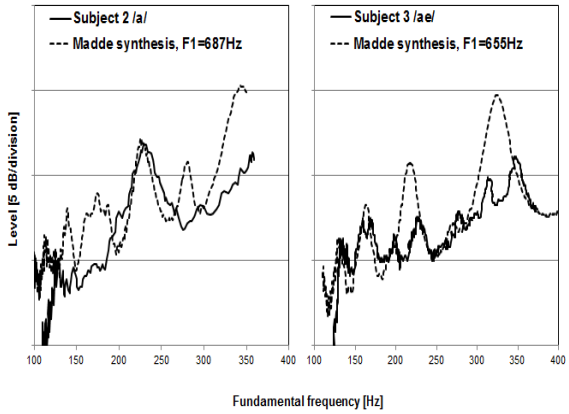


Fig. 3. Solid curves: Overall sound level versus fundamental frequency for glide tones produced by the indicated subjects on the indicated vowels. Dashed curves: Synthesised versions of the same glides obtained from the MADDE software, which is void of source-filter interaction.

IV. Discussion

The reliability of the findings reported here is entirely dependent on a correct setting of F1 in the inverse filtering. The instant display of the waveform and spectrum of the inverse filtered signal greatly facilitates this setting; even a small mistuning of F1 results in a clearly visible ripple in the closed phase which is combined with an uneven source spectrum envelope.

The observed attenuation of a source spectrum partial that coincides with F1 could have resulted also from a mistuning of F1 in the inverse filtering. If so, spectra obtained for adjacent F0 values would have shown a ripple during the closed phase. However, the closed phase remained ripple-free throughout the pitch glide in response to an unchanged inverse filter F1. This shows that the subjects kept the same articulation, which can probably be ascribed to the fact that they were experienced singers; less experienced singers would tend to raise the larynx with rising pitch, thus causing the formant frequencies to gradually increase with F0.

Source-filter interaction has been investigated quite extensively in the past. Mostly it has been shown to cause disturbance of the periodicity of vocal fold vibration, as mentioned. No such disturbances were observed in the present material. The reason again would be that the subjects were experienced singers. The methods available to singers for avoiding such disturbances remain an open question.

A less commonly observed effect of source-filter interaction is that the coincidence of a partial with F1 decreases the amplitude of that same partial. A somewhat related observation was recently made by Maxfield and associates [8]. They found that the intensity of a partial in the radiated sound failed to increase and decrease at nearly the same rate, as the partial passed a formant in a glide tone, thus creating an asymmetric peak in the intensity-time contour. They interpreted this finding as evidence of a source-filter interaction. Our data did not show intensity asymmetry around F1, but rather a quite substantial attenuation of the SPL peak that normally would be expected to accompany the coincidence of a partial and F1.

In the past some attempts have been made to derive the vocal tract transfer function from the amplitude modulation of partials in the radiated spectrum during vibrato singing [9,10]. In such singing F0 is modulated at a frequency in the vicinity of 5 Hz. This generates an amplitude modulation of the spectrum partials that increases with the partial's increasing proximity to a formant. In these studies, no phenomenon similar were reported. However, the amplitude modulation is dependent on the formant bandwidths. In our inverse filtering we used bandwidths that eliminated the ripple during the closed phase, while in those studies the bandwidths were derived from the observed amplitude modulation.

SPL typically is almost entirely determined by the amplitude of the strongest spectrum partial, as mentioned. The attenuation of the partial coinciding with F1 therefore produced a substantial reduction of the SPL peak observed in the synthesised glide tone, as demonstrated by the synthesis experiment. A similar phenomenon has been noted also in analyses voice range profiles (Peter Pabon, personal communication).

V. Conclusion

Inverse filtering glide tones produced by experienced singers on the vowel /a/ and /ae/ has shown that the amplitude of a source spectrum partial is attenuated when it coincides with F1. The effect can be assumed to be caused by source-filter interaction. The finding should be relevant to the understanding of voice range profiles.

REFERENCES

[1] J. Sundberg, "Flow glottogram and subglottal pressure relationship in singers and untrained voices". *J Voice*, early online May 2017.
[2] G. Fant, "Some problems in voice source analysis". *Speech Commun.* 13, 7–22, 1993.
[3] M. Rothenberg, "Acoustic interaction between the glottal source and the vocal tract". In: Stevens KN, Hirano M, eds. *Vocal Fold Physiology*. Tokyo:
Univ. of Tokyo Press, 305–328, 1980.
[4] I.R. Titze, "Nonlinear source–filter coupling in phonation: theory". *J. Acoust Soc. Am.* 123, 2733–2749, 2008.
[5] M. Rothenberg, "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing", *J Acoust Soc Am.* 53, 1632–1654, 1973.
[6] P. Alku, "Glottal inverse filtering analysis of human voice production — A review of estimation and parameterization methods of the glottal excitation and their applications", *Sadhana* 36, 623–650, 2011.
[7] P. Gramming, J. Sundberg, "Spectrum factors relevant to phonetogram measurement", *J Acoust Soc Amer* 83, 2352-2360, 1988.
[8] L. Maxfield, A. Palaparthi, I. Titze, "New evidence that nonlinear source-filter coupling affects harmonic intensity and fo stability during instances of harmonics crossing formants, *J Voice* 31, 149–156, 2016.
[9] S Imaizumi, H Saida, Y Shimura & H Hirose, "Harmonic analysis of the singing voice: - Acoustic characteristics of vibrato", in A. Friberg, J. Iwarsson, E. Jansson, J. Sundberg, eds *Proceeedings of the Stockholm Music Acoustic Conference 1993 (SMAC 93)*, Stockholm: Publication No 79 issued by the Royal Swedish Academy of Music, 197-200, 1994.
[10] I. Arroabarren, A. Carlosena, "Vibrato in singing voice: The link between source-filter and sinusoidal models", *EURASIP J. Appl. Signal Process*. 7, 1007–1020, 2004.

GLOTTIS IN VOCAL FRY ANALYZED BY HIGH SPEED DIGITAL PHONOSCOPY AND NYQUIST PLOTS

K. Izdebski[1,2], Y. Yan [2], M. Blanco [2]

[1] PVSF, Chairman, San Francisco, CA, USA
[2] Bioengineering, Santa Clara University, Santa Clara, CA, USA
kizdebski@pvsf.org, yyan1@scu.edu, mblanco@scu.edu

Abstract: **Glottal fry (GF) is the lowest range of human vocalization and can be produced voluntarily or as a part of dysphonia. Based on HSDP, spatio-temporal analysis of vibration patterns, multi-line kymograms of voluntary GF we conclude that supraglottic contraction assists in prolonged closed phase, and that this phase is elongated in VF. Further studies will contract these observations with various pathologic GF samples.**
Keywords : HSDP, LVS, female voice, vocal fry, mucosal wave, glottic cycle, GAW, PFFT, color analysis, Nyquist plots

I. INTRODUCTION

Vocal fry (VF)–also known as fry, pulse register, creaky voice, rattle, scrape, strohbass, or laryngealized voice (these terms used at random to describe either pathological or non-pathological voice quality)– recently became popular in use by young females in California, and it is also used to express vocal emotions [1-3]. VF although typically an expiratory sound can also be produced as an inspiratory sound. VF has been studied acoustically suggesting a very short opened glottis phase (pulse). Here we employed advances in biomedical optical to study VF specifically by using HSDP.

II. METHODS

All visual HSDP data were recorded using color HSV system (KayPENTAX Model 9710, NJ, USA). Standard phonoscopic signal acquisition without topical anesthesia was used. The rate of video frames was set to 2000 with maximum available image resolution of 512 x 512 pixels. Acquired signals were processed by KIPS (KayPENTAX) when appropriate and using custom programs such as Vocalizer® elsewhere [5-6].

HSDP images were compared to LVS images obtained from the same subject on the same date with the use of standard LVS equipment (Kay-Elemetrics RLS 9100, NJ, USA).

VF phonation was executed on exhalation by a female subject during production of a sustained /i/ sound. This mode was chosen to contrast these findings to pathological conditions in which such modes constitute primary symptoms. For example fry phonation presents in dystonia, traumatic brain injury, closed head injury, or various emotional voice qualities [1-2].

III. RESULTS

Muscular adjustments expressed in VF positioning were shown in detail using HSDP. Images were superior to those obtained from LVS.

A. Glottic wave

Fig. 1 (A & B) shows glottic view during VF from LVS (A) and HSDP (B) recordings respectively with a transoral rigid endoscope. Note that HSDP displays more vividly the moment of glottis separation.

Findings with KIPS Kymography showed extremely prolonged closed phase of the glottic cycle. This is illustrated in Fig. 2. Using Vocalizer ® technology [5], we showed that release of phonatory wave was very brief and consisted of a double pulse creating a single sound. This phase release predominated.

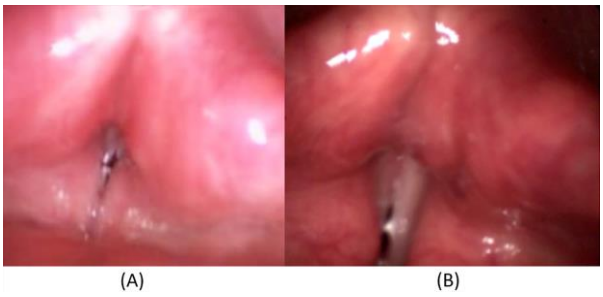


Fig. 1. Glottic separation observed by LVS (A) and by HSDP (B).

This double pulse was generated by two separate locations along the glottis (Fig. 2) and represented glottis separation of very short duration and a very long closed phase that was longitudinally limited to mid-glottic portion. These pulses are represented by the two

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

separate illuminated areas of the glottis (outlined in green). Also as shown in this illustration, a significant supraglottic constriction of the false vocal folds (FVF) is present during vocal fry production.

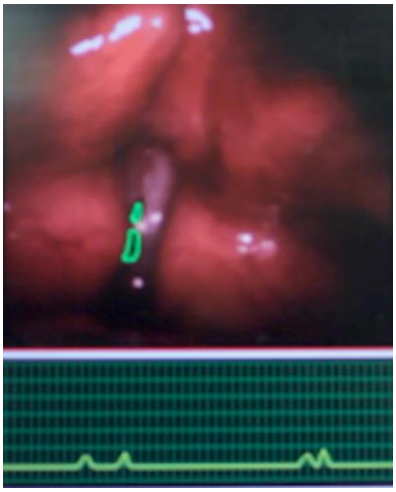


Fig. 2. Note extreme long closed phase of the glottis cycle and compression of the ventricular folds. Vocalizer® analysis demonstrating double glottic pulse in VF.

B. Nyquist plots

To provide more information on the acoustics of VF we also generated Nyquist plots form voice signals alone [6]. These findings (left plot) are contrasted to normal phonation (right plot) from the same subject shown in Fig. 3. Acoustic data corresponds well to the physiologic data, showing a double loop plot that in our opinion corresponds to the double pulse observed in the mucosal wave pattern as shown in Fig. 2.

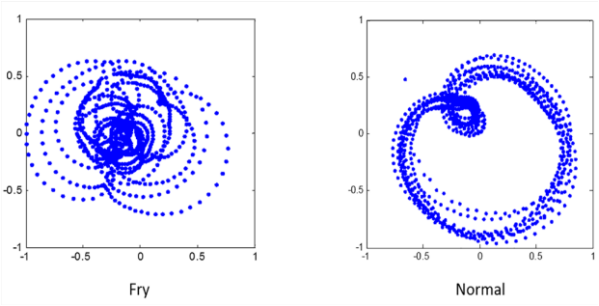


Fig. 3. Nyquist plot from acoustics for vocal fry contrasted against modal normal phonation for same female speaker. Note concentration of F0 in the center for fry.

IV. DISCUSSION

LVS and HSDP are informative about supraglottic topography, with LVS missing fine movements of these structures. HSDP is superior in demonstrating details of the mucosal wave.

II. CONCLUSIONS

Based on these findings we conclude that mucosal wave is suppressed in VF mode showing a very long closed portion of the glottal cycle and a very short pulse. Generated fry sound may be composed from double pulses generated at two locations within the glottis but proceeding very rapidly as to form a double pulse repetition pattern. Pulse is mixed with noise and medial compression of the entire glottis causes momentary cessation of the vibratory cycle.

VF mode also showed supraglottic contraction. This medial motion of the supraglottic structures is non-existing in normal (modal) phonation produced by the same subject. In VF, vocal fold oscillations are suppressed despite the true vocal folds and the FVF midline approximation. Acoustic Nyquist plots revealed fine structure of F0 of VF.

REFERENCES

[1] K. Izdebski, “Clinical voice assessment: The role & value of phonatory function studies (Chapter 29),” in *Current Diagnosis & Treatment Otolaryngology–Head and Neck Surgery*, third ed, A. Lalwani, Eds. New York: Lange, 2011.

[2] K. Izdebski, *Emotions in the Human Voice, Vol. I-III*. San Diego: Plural Publishing, 2008.

[3] I. Yuasa, “Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women,” *Am. Speech*, vol. 85, pp. 315-337, 2010.

[4] K. Izdebski, J. Ross, J. Klein, “Rigid transoral laryngovideostroboscopy (phonoscopy),” *Seminars in Speech and Hearing, Thieme*, 1990.

[5] Y. Yan, K. Izdebski, “Integrated spatio-temporal analysis of high-speed laryngeal imaging and acoustic signals: Their role and applications in the study of normal and abnormal vocal functions,” in *Speech and Language Technology, vol. 12/13*, G. Demenko, E. Wagner, Eds. Poznan: Polish Phonetic Association, 2010, pp. 15-38.

[6] Y. Yan, G. Demenko, K. Izdebski, A. Bhandari, “Using Nyquist plots to describe real and mimicking traumatic phonation,” *XVIII Annual Pacific Voice Conference PVSF/UCLA*, 2010.

WHITE LIGHT, NBI® & HSDP EXAM OF BAMBOO VOCAL FOLDS

K. Izdebski[1,2,] E. V. Osipenko[3], R. M. Cruz[1,4], M.Just,[1,5]

[1,2] PVSF, Chairman, San Francisco, CA, USA
[2]Bioengineering, Santa Clara University, Santa Clara, CA, USA
[3] Phoniatrics Department of Federal Research Clinical Otolaryngology Centre of the Russian Federation Healthcare Ministry, Moscow, Russia
[4] KP, Otolaryngology, Head & Neck Surg., Oakland, CA, USA
[5]DiagNova Technologies sp z o.o., Wrocław, Poland
kizdebski@pvsf.org, osipenko-lor@yandex.ru, Raul.Cruz@kp.org, marcin.just@diagnova.pl

Abstract: **The appearance and kinematics of the vocal folds (VF) in multiple cases of bilateral and extremely rare VF mucosal lesions referred to in the literature as "bamboo vocal folds" are presented. Using white light (WL), Narrow Band Imaging (NBI®) and HSDP visualization we found these structures to be avascular in nature. Our findings support immunologic, rather than traumatic causation.**

Keywords : B-nodes, bamboo VF, hoarseness, VF deposits, autoimmune diseases, NBI®

I. INTRODUCTION

Bamboo VF (B-nodes) have been given this name as reminiscent of the banding on a bamboo stalk [1-5] revealing transverse band-like appearing submucosaly with a somewhat elongated cystic or globular appearance in various mid-membranous portions of the VF. These lesions are bilateral, but not always opposing each other and in contrast to traditional nodes, B-nodes are oblong and transverse the entire dorsal (superior) surface of the VF on each side, displacing the VF mucosa upwards. This arrangement segments the mucosal wave, causing significant voice change. The most common voice characteristics in these are: instability of pitch and intermittent (aperiodic) dysphonia, diplophonia [1-5], and at times even momentary aphonia.

B-nodes occur in females only. Several etiologies have been postulated in sparse world literature [1-8]: systemic lupus erythematosus, variety of autoimmune disease processes including mixed connective tissue disorders, rheumatoid arthritis, relapsing polychondritis, and Hashimoto's and Sjögren's syndrome. One study advanced atraumatic etiology [9] describing a patient with no clinical autoimmune disease but high vocal demands, and suggested that histopathological analysis of her lesions revealed microscopic findings supportive of a traumatic etiology. In a review of the current literature, where vocational information regarding potential demand was noted, about 80% of patients reported demanding vocal usage, however demanding vocal usage need not be considered traumatic voice use [7].

II. METHODS

All our observations were made with WL and with NBI®, using Olympus NBI® system (Center Valley, PA, USA): Model OTV-S190 processor and CLV-S190ENT light source. Visualizations were performed with a distal chip flexible endoscope. The flexible scopes used were an ENF-VH scope, a 3.9mm OD 1080HD distal chip scope, or an ENF-V3 2.6mm OD high resolution distal chip scope. One case was studies with HSDP (Kay-PENTAX) and LVS (Kay-PENTAX, PENTAX Medical A Division of PENTAX of America, Inc. 3 Paragon Drive Montvale, New Jersey, 07645-1782 USA) and resultant data were processed using the DiagNova system (DigNova Technologies, Wroclaw Technology Park, Wroclaw, Poland). DiagNova provides analysis of VF amplitude for each VF (left or right), or both, as well as analysis of VF closure (opening and insufficiency), asymmetry, and phase differences, generating both kymograms and phonovibrograms.

III. RESULTS

A. Illumination and histology

Fig. 1 represents B-nodes illuminated with WL and with NBI®. NBI® illumination clearly reveals the location and the characteristics of the B-nodes in contrast to WL illumination. No evidence of vascular trauma typically present in phonotrauma is noted.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

Fig. 1. Vascular disruptions are minimal and do not display hemorrhagic events that we feel could be indicative of phonotrauma.

Fig. 2 represents H&E histopathology of the excised B-node at 10x and 40x magnification from a different case. The more intensely staining area is characteristic of these lesions. There are dense linear fibrinoid necrotic deposits. These appear as eosinophilic "rods" surrounded by fibroblasts, histiocytes, and occasional multinucleated giant cells. This represents a granulomatous reaction, often seen in autoimmune disorders. She had a dramatic improvement in voice quality and remains stable by examine and voice quality 12 months postoperatively.

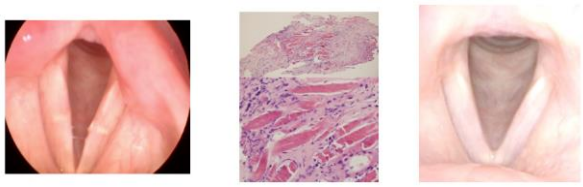


Fig 2. Shows B-nodes before and after removal and a H&E staining at 10x and 40x mag.

B. DiagNova Analysis

HSDP recordings were subjected to Diagnova analysis package comprising kymography and parametization to describe how the B-nodes disrupt or not disturb the symmetry of L vs. R VF vibratory patterns, hence based on only one case we are limited in concluding unequivocally on global voice characteristics of B-nodes. Findings included: 1) the B-node anomaly is not behaving as an independent "growth or membranous factor"; 2) the B-node lacks independently resonating frequency, but behaves rather like it was an integral part of the VF with only marginal global effects on the vibratory cycle; and 3) B-node introduces localized modification of the VF function, causing glottic gap and disturbing phase differences of the vibratory cycle.

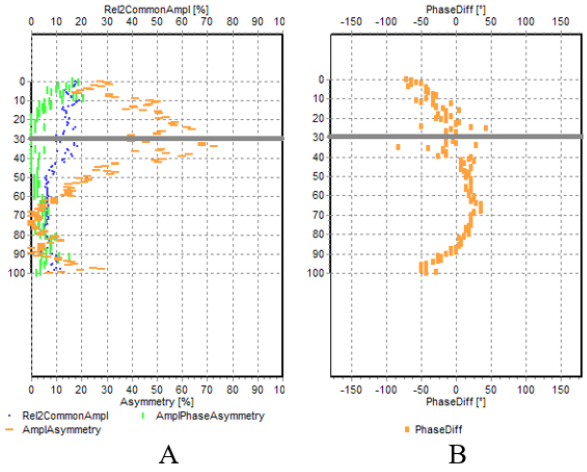


Fig. 3. A) shows parameterization results: Orange = ampl asymmetry; green = phase asymmetry; blue = glottal gap. B) show phase differences with respect to B-node deposit location. This difference is most visible at the actual B-node deposit and varies from negative to positive from the posterior to anterior positions. Indirectly, this phase difference tells us which VF (L or R) leads in phase.

IV. DISCUSSION

LVS and HSDP are informative about supraglottic topography, with LVS missing fine movements of these structures. HSDP is superior in demonstrating details of the mucosal wave.

V. CONCLUSIONS

NBI® give superior resolution of the B-nodes than does LVS. No vascular disruption was noted, what speaks against phonotrauma. HSDP DiagNova analysis showed vibratory characteristics, specifically pointing out the notions that the B-node works as the integral part of the VF, what speak against the traumatic etiology. Phase differences indicate different vibratory patterns of L and R VF, showing that one VF with the B-node located more towards the anterior commissure causes more havoc than the B-node located more posteriorly. However, to draw more universal conclusions, a larger data corpus is needed. Results from our five cases provide overall support that an autoimmune etiology rather than a traumatic one is responsible for B-nodes formation.

It also seems likely that if phonotrauma were the primary inciting event, documented cases in men would be reported. Additionally, four of five of our cases involved patients without high vocal demands. The asymmetrical location of the B-nodes in all five cases speaks against phonotrauma.

As previously mentioned, etiology is important as it may influence our treatment decisions. If these lesions are traumatic in nature, it seems that voice and anti-inflammatory therapy are more likely to be helpful. If they are not, more prompt surgical intervention may be warranted. Two of our five cases were offered voice therapy but did not respond to it. Four of five cases with documented autoimmune disease received medical therapy to control systemic disease but did not show a favorable voice response. The one case operated upon responded positively to surgery with a much-improved voice and no recurrence one year later. NBI® studies of more cases are anticipated and as these become available, we believe the atraumatic nature of these lesions and appropriate management will become better defined as more cases are studied.

ACKNOWLEDGMENT

This work was supported in part by PVSF, San Francisco, CA, USA. The authors express their gratitude to Mathew Blanco, MA (SCU & WVC) for his editorial skills.

REFERENCES

[1] C. Schwemmle, M. Ptok, "Bamboo nodes as a cause of juvenile dysphonia," *Klin. Padiatr.*, vol. 224, 2012, pp. 468-469.

[2] E. Hilgert, B. Toleti, K. Kruger, I. Nejedlo, "Hoarseness due to bamboo nodes in patients with autoimmune diseases: a review from the literature," *J. Voice*, vol. 22, 2008, pp. 343-350.

[3] E. Murano, et al., "Bamboo node: primary vocal fold lesion as evidence of autoimmune disease," *J. Voice*, vol. 15, 2001, pp. 441-450.

[4] R. Ylitalo, M. Heimbürger, P. Lindestad, "Vocal fold deposits in autoimmune disease. An unusual cause of hoarseness," *Clin. Otolaryngol. Allied Sci.*, vol. 28, 2003, pp. 446-450.

[5] H. Yamashita, Y. Takahashi, T. Kano, A. Mimori, "A case of systemic lupus erythematosus with bamboo joint-like corditis as an antecedent symptom," *Rheumatology*, vol. 52, 2013, pp. 759-761.

[6] K. Izdebski, et al., *Normal and Abnormal Vocal Folds Kinematics. High-Speed Digital Phonoscopy (HSDP), Optical Coherence Tomography (OCT) & Narrow Band Imaging (NBI®). Volume I: Technology*. San Francisco: PVSF e-Q&A-b, 2015.

[7] K. Izdebski, "The role & value of the phonatory function studies (Chapter 29)," in *Current Diagnosis & Treatment in Otolaryngology: Head & Neck Surgery, 3rd ed.*, A. Lalwani, Ed. New York: Lange, 2012, pp. 416-429.

[8] E. Osipenko, K. Izdebski, R. Cruz, "NBI® value in evaluating non-cancerous laryngeal lesions. Russian and US experience," *3rd Congress of European ORL-HNS*, Prague, Czech Republic, 2015.

[9] L. Li et al., "A pathological study of bamboo nodule of the vocal fold," *J. Voice*, vol. 24, pp. 738-741, 2010.

SESSION VI:

VOCAL FOLDS DYNAMICS II

A METHOD FOR ANALYSIS OF THE VOCAL FOLD VIBRATIONS IN CONNECTED SPEECH USING LARYNGEAL IMAGING

M. Naghibolhosseini [1], D. D. Deliyski [1], S. R. C. Zacharias [2-4], A. de Alarcon [3,4], R. F. Orlikoff [5]

[1] Department of Communicative Sciences and Disorders, Michigan State University, East Lansing, Michigan, USA
[2] Department of Otorhinolaryngology, Mayo Clinic Hospital, Phoenix, Arizona, USA
[3] Division of Pediatric Otolaryngology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA
[4] Department of Otolaryngology-Head and Neck Surgery, University of Cincinnati, Cincinnati, Ohio, USA
[5] College of Allied Health Sciences, East Carolina University, Greenville, North Carolina, USA
naghib@msu.edu, ddd@msu.edu, stefcotton@hotmail.com, alessandro.dealarcon@cchmc.org, orlikoffr16@ecu.edu

Abstract: **This study proposes an algorithm for segmentation of the vibrating vocal folds during connected speech. The data were obtained via laryngeal high-speed videoendoscopy (HSV) during reading of the "Rainbow Passage" using a custom-designed color HSV system. To address the complexity of the HSV image in connected speech, the segmentation consists of three stages: temporal segmentation, motion compensation, and spatial segmentation. The temporal segmentation determines the time locations of the vibrating vocal folds across the HSV frames. The motion compensation allows for removing unwanted motion of the glottis. The spatial segmentation employs an active contour analysis, which is performed on the vocal-segments' HSV kymograms. The active-contour model is based on energy optimization and describes analytically the edges of the vocal folds in the kymograms during phonation. The results suggest motion compensation was successful in detecting the vibrating vocal folds location and extracting the kymograms in the presence of tissue maneuvers. The active-contour algorithm made it possible to describe the vocal fold edges in the kymograms. HSV-based voice assessment of connected speech can lead to significant improvements in clinical voice practice. Developing automatic algorithms for HSV analysis is a necessary step allowing the extraction of clinically relevant information from big data.**

Keywords: **High-speed videoendoscopy, laryngeal imaging, connected speech, motion compensation, spatial segmentation**

I. INTRODUCTION

Voice disorders usually reveal during communication using connected speech. Therefore, it would be essential to perform instrumental functional voice assessment in the context of connected speech. Although, videostroboscopy is often used in running speech, its utility is limited to assess only the gross laryngeal movements, because its principles do not allow for the visualization and analysis of intra-cycle vibratory characteristics outside the context of prolonged sustained phonation [1,2]. Laryngeal high-speed videoendoscopy (HSV) enables the recording of vocal fold vibrations with high temporal resolution [2,3]; but to use HSV in connected speech, the challenge has been to couple it with flexible fiberoptic endoscopes. Recently, this challenge has been overcome [4,5]. HSV captured during connected speech results in huge datasets, which require developing automated algorithms and methodologies for big-data analysis. Such automated methods can help extract and emphasize clinically relevant information from the HSV data. To this end, we have recently developed an automatic temporal segmentation algorithm to extract timestamps of the vibratory onsets and offsets, and the epiglottic obstructions of the glottis [6]. The findings of the temporal segmentation method were applied in developing an automated spatial segmentation technique, which provides analytic representation of the edges of the vocal folds. Prior to performing the spatial segmentation, motion compensation needs to be performed. This study describes an automatic algorithm for segmentation of HSV in connected speech, which consists of three stages: temporal segmentation, motion compensation and spatial segmentation. The emphasis in this article is on the motion compensation, which aligns the vocal folds across frames to overcome the problem of laryngeal maneuvers in connected speech. Motion compensation is necessary for performing kymography-based spatial segmentation on the results from the temporal segmentation. An active-contour modeling approach was applied to the HSV-derived kymograms toward detection and description of the vocal fold edges for spatial segmentation [7].

II. METHODOS

HSV data collection: A vocally normal 38-year-old female participated in this study. The examination was performed at the Center for Pediatric Voice Disorders, Cincinnati Children's Hospital Medical Center. The

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

participant did not have history of voice disorder. A custom-built flexible fiberoptic HSV system was used to record a "Rainbow Passage" production from the participant. The HSV system was set at 4,000 frames per second and integration time of 249 μs. The spatial resolution of the HSV images was set to 256x256 pixels. The length of the recording was 29.14 s (total of 116,543 frames). The HSV system included a FASTCAM SA-Z color high-speed camera (Photron Inc., San Diego, CA) equipped with a 12-bit color image sensor, 64 GB of cache memory, a 300-W xenon light source, model 7152A (PENTAX Medical Company, Montvale, NJ), and a 3.6-mm Olympus ENF-GP Fiber Rhinolaryngoscope (Olympus Corporation, Tokyo, Japan). After recording, the HSV sequence was saved as an uncompressed 24-bit RGB AVI file.

Data analysis: A motion compensation method was developed and applied to each vocalized segment extracted by the temporal segmentation method described in [6]. The motion compensation was done using a gradient-based algorithm. The gradient-based algorithm was developed to suppress the motions unrelated to the vocal fold vibrations and the tissue maneuvers. The gradient was computed as the time differential of the red channel data with step size of 1.5 ms (6^{th} frames). The red channel was used since it contained the main information regarding the vocal fold vibrations and less noise.

Each frame of the gradient was spatially filtered using a Gaussian filter to remove the noise. The result was then temporally filtered using a Hamming bandpass filter (cutoff frequencies of 70 and 1000 Hz). Although the aforementioned steps removed noise significantly, the presence of moving edges unrelated to the vocal fold vibrations were observed. To remove the anterior-to-posterior edge movements, the analyzed gradient was added to its absolute value (termed positive gradient). In addition, the absolute value was subtracted from the analyzed gradient to remove the posterior-to-anterior edge movements (termed negative gradient). The positive and negative gradients were then bandpass filtered and multiplied with the analyzed gradient. The resulting denoised gradient was used to determine the location of the vibrating vocal folds across the frames.

The location of the vibrating vocal folds in each frame was determined based on the first moment of inertia of the denoised gradient multiplied by a motion window. The motion window was in a shape of an ellipse and was computed using a moving-average gradient-based algorithm that is explained in [6]. The motion window was the smallest window that enclosed the location of the vibrating vocal folds across all frames. The first horizontal and vertical moments, denoted by $M_1(x, t_i)$ and $M_1(y, t_i)$, were computed as follows:

$$M_1(x, t_i) = \frac{\sum_{x=1}^{256}\sum_{y=1}^{256} f(x, y, t_i)x}{\sum_{x=1}^{256}\sum_{y=1}^{256} f(x, y, t_i)}, \quad (1)$$

$$M_1(y, t_i) = \frac{\sum_{x=1}^{256}\sum_{y=1}^{256} f(x, y, t_i)y}{\sum_{x=1}^{256}\sum_{y=1}^{256} f(x, y, t_i)}. \quad (2)$$

Based on the estimated location of the vocal folds in each frame, the vocal folds were aligned across the frames. Next, the kymograms of the HSV data were extracted inside a rectangular window (with a specific location) that enclosed the vibrating vocal folds in each frame. The size of the window was so that the motion window was inscribed in it. The kymograms were then extracted by passing a line in the medial section of the frames to capture the vocal fold vibrations over time.

The active contour modeling approach was applied to the HSV kymograms of each vocalized segment to provide an analytic description of the vocal fold edges across the frames. During the active contour modeling, a pair of active contours (open-curve snakes) were used that could deform toward the edges of the vocal folds [7]. The snakes were attracted to pixels with large spatial gradient values that were associated with the edges of the vocal folds (glottis boundaries). These deformable models work through an energy-minimization procedure. The goal was to minimize an energy function composed of the internal forces of the snakes and the external forces derived from the spatial gradient of each kymogram. The internal force acting on each snake was a function of first and second spatial derivatives of pixel intensities to adjust the snake's rigidity and elasticity. The initialization of the snakes was done using the scaled vertical second moment of inertia. The energy function was optimized using time-delayed dynamic programming.

The performance of the motion compensation algorithm was investigated by visually checking the analyzed HSV data to ensure that the location of the vibrating vocal folds was captured across the frames.

III. RESULTS

The result of the gradient-based algorithm for denoising of the HSV data for two frames are shown in Fig. 1. The denoised data in Fig. 1-C was computed based on the gradients of the red channels for frame #4469 (Fig. 1-A) and frame #4475 (Fig. 1-B). As seen in Fig. 1-C, the noise is effectively removed and only the glottis area, associated with the vocal fold vibrations, remains.

The digital kymogram (red channel) for the first vocalization of the Rainbow Passage is shown in Fig. 2. Fig. 2-A provides an example of a kymogram before motion compensation. The bright area in Fig. 2-A between frame #4460 and 4700 results from the

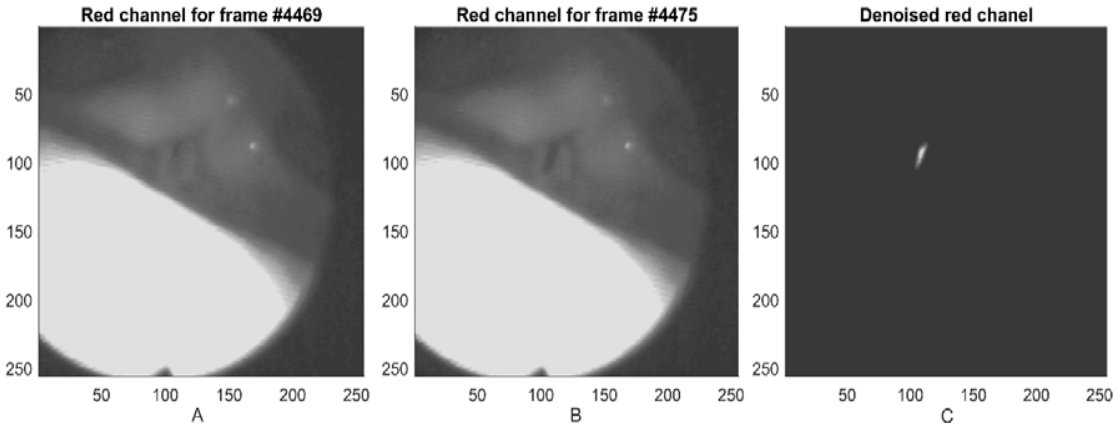


Figure 1: (C) Denoised HSV (red channel data) computed based on the gradients of frame #4469 (A) and frame #4475 (B).

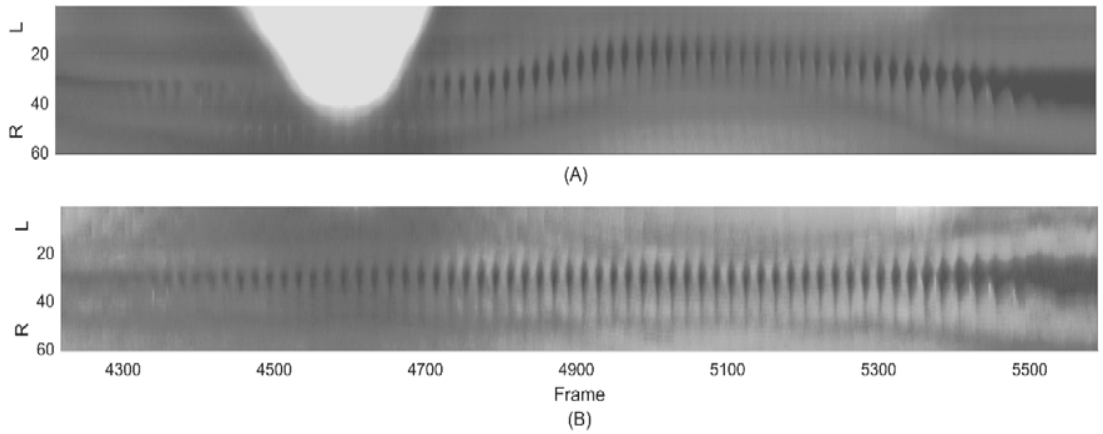


Figure 2: (A) Digital kymogram at medial section of the vocal folds for the first vocalized segment of the Rainbow Passage (frame #4217–5588). (B) Digital kymogram after performing motion compensation for similar frames as in (A). The L and R on the y-axis denote the left and right side of the image in HSV frames, respectively.

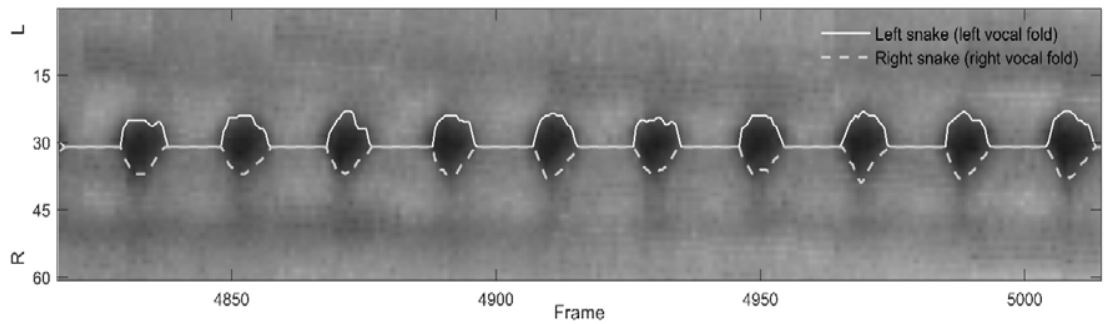


Figure 3: Digital kymogram of 200 frames (frame #4817-5017, selected from the kymogram in Fig. 2-B) after motion compensation. The edges of the left and right vocal folds, extracted using active contour modeling, are shown by the solid and dashed lines, respectively. The L and R on the y-axis denote the left and right side of the image in HSV frames, respectively.

reflection of the epiglottis. In this case, the vocal folds are moving longitudinally relative to the camera, from the anterior toward the posterior and backward, resulting into kymographic scans of various portions of the glottis and the epiglottis in the same kymogram. While the left-right motions of the vocal folds relative to the camera are seen as up-down fluctuations of the glottal opening in the kymogram. Fig. 2-B shows the kymogram after applying the motion compensation to the HSV frames. The motion in the left-right and anterior-posterior planes are compensated for. The epiglottis reflection is no longer seen in Fig. 2-B because the laryngeal structures' motion in the anterior-posterior direction is compensated for. As also seen in Fig. 2-B, the glottis opening and closing occurs on a straight line after the motion compensation was performed, which prepares the data for further analysis using active contour modeling.

The result of active contour modeling applied to 200 frames (frame #4817-5017) of the kymogram in Fig. 2-B is shown in Fig. 3. The left and right vocal fold edges extracted using the active contour modeling are depicted by the solid and dashed lines, respectively. As seen in this figure, the active contour approach was successful in detecting the edges of the vocal folds and providing an analytical description of the edges.

IV. Discussion

The gradient-based denoising algorithm was able to remove the information irrelevant to the vocal fold vibrations. Hence, the motion compensation was done successfully and the location of the vibrating vocal folds was determined correctly based on the comparison performed with the visual rating of the HSV data. Due to the satisfactory performance of the motion compensation algorithm, the alignment of the vocal folds and HSV-based kymogram extraction were done successfully. Hence, the use of active contour approach enabled the analytic description of the edges of the vocal folds at the medial section of the vocal folds.

V. Conclusion

The motion compensation algorithm, proposed in this study, was successful in compensating for the relative motion of the vocal folds and the endoscope for all segments resulting from the temporal segmentation. The active contour modeling was shown to be successful in detecting the vocal fold edges in the medial section of the vocal folds across the frames. This algorithm will be further tested and modified in order to achieve spatial segmentation on the full length of the vocal folds. Further, the proposed algorithm will be tested on a larger HSV dataset to address the inter- and intra-subject reliability.

ACKNOWLEGMENTS

Funding for this study was provided by the National Institutes of Health, NIDCD, R01 DC007640 "Efficacy of Laryngeal High-Speed Videoendoscopy" and by the Michigan State University Foundation.

REFERENCES

[1] A.E. Aronson and D. Bless, *Clinical voice disorders,* 4th ed, Thieme, 2009.

[2] *Laryngeal evaluation: Indirect laryngoscopy to high-speed digital imaging*. K.A. Kendall and R.J. Leonard, Eds. New York, NY: Thieme, 2010.

[3] D.D. Deliyski and R.E. Hillman, "State of the art laryngeal imaging: research and clinical implications," Curr. Opin. Otolaryngol. Head Neck Surg., vol. 18, n. 3, pp. 147-152, 2010.

[4] M. Zañartu, D. D. Mehta, J. C. Ho, G. R. Wodicka, and R. E. Hillman, "Observation and analysis of in vivo vocal fold tissue instabilities produced by nonlinear source-filter coupling: A case study," J. Acoust. Soc. of Am., vol. 129, n. 1, pp. 326-339, 2011.

[5] D. D. Mehta, D. D. Deliyski, S. M. Zeitels, M. Zañartu, and R. E. Hillman, "Integration of transnasal fiberoptic high-speed videoendoscopy with time-synchronized recordings of vocal function," in *Normal & Abnormal Vocal Folds Kinematics: High Speed Digital Phonoscopy (HSDP), Optical Coherence Tomography (OCT) & Narrow Band Imaging (NBI®), Volume I: Technology,* vol. 12, K. Izdebski, Y. Yan, R.R. Ward, B.J.F. Wong, and R.M. Cruz, Eds. San Fransisco, CA: Pacific Voice & Speech Foundation, 2015, pp. 105-114.

[6] M. Naghibolhosseini, D. D. Deliyski, S. R. C. Zacharias, A. de Alarcon, and R. F. Orlikoff, "Temporal segmentation for laryngeal high-speed videoendoscopy in connected speech," J. Voice, S0892-1997(17)30137-6 [Epub ahead of print], 2017.

[7] H. Moukalled, D. Deliyski, R. Schwarz, and S. Wang, "Segmentation of Laryngeal High Speed Videoendoscopy in Temporal Domain using Paired Active Contours," in *Proceedings of the 6th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications MAVEBA,* vol. 6, C. Manfredi Ed. Firenze, Italy: Firenze University Press, 2009, pp. 137-140.

QUANTIFICATION OF INTRAGLOTTAL PRESSURE DURING THE MODAL VIBRATION CYCLE

P. H. DeJonckere [1] J. Lebacq[2],

[1]Federal Agency for Occupational Risks, Brussels & Department of Neurosciences KULeuven, University of Leuven, Leuven, Belgium
[2]Neurosciences Institute, University of Louvain, Brussels, Belgium
philippe.dejonckere@kuleuven.be

Abstract:

Intraglottal pressure is the driving force of vocal fold vibration. Its time course during the open phase of the vibratory cycle is essential for understanding the mechanics of phonation, but measuring it directly is difficult and may hinder spontaneous voicing. However, the intraglottal pressure can be computed from the *in vivo* measured transglottal flow and glottal area (hence the air particle velocity) on the basis of the Bernoulli energy law. Calculations are presented for three shapes of glottal duct : uniform, convergent and divergent. When the airflow curve is skewed to the right relative to the glottal area curve, whatever the glottal duct configuration, the intraglottal pressure during the opening phase systematically exceeds that during the closing phase, which is the basic condition for sustaining vocal fold oscillation. The skewing results from air compressibility and vocal tract inertance. The intraglottal pressure becomes negative during the closing phase.

Keywords : **intraglottal pressure, glottal shape, Bernoulli's equation.**

I. INTRODUCTION

Intraglottal pressure is the driving force of vocal fold vibration. Its time course during the open phase of the vibratory cycle is essential for understanding the mechanics of phonation. However measuring it directly (by tracheal puncture or using a transglottal tranducer) is difficult and may hinder spontaneous voicing [1]. Some previous studies have found a sharp peak of subglottal pressure synchronous with the vocal fold impact, but it is probably the result rather than the cause of the movement of the vocal folds.

A positive flow of energy from the airstream to the tissue can be realized only if the net aerodynamic driving force has a component in phase with the tissue velocity (i.e. the first derivative of displacement, thus with a phase lead of 90° over displacement) [1;2]. By using a model in which the intraglottal pressure P is computed from the transglottal flow and the air particle velocity on the basis of the Bernoulli energy law,

$$P + \tfrac{1}{2} \rho v^2 = constant \quad (1)$$

where ρ is fluid density and *v* is particle velocity, it can be shown that, when the airflow curve is skewed to the right with respect to the glottal area curve (Fig.1), the intraglottal pressure during the opening phase exceeds that during the closing phase. The skewing results from air compressibility and vocal tract inertance.

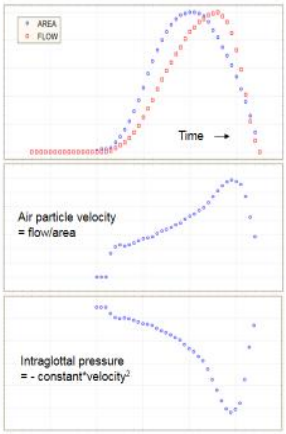


Fig. 1 (after Titze). Graphic simulation of a single vibration cycle of the vocal folds in a typical normal phonation of a male subject (modal register). The upper panel shows the glottal area and the airflow as a function of time. Both signals increase upwards. The central panel represents the shape of the air particle velocity waveform, obtained by dividing the flow waveform by the displacement waveform. Only the open part of the vibration cycle is shown. The bottom panel is the waveform of the intraglottal pressure, computed on the basis of Bernoulli's energy law.

However this is only a first approximation of the driving force on the tissue since it assumes a laminar, incompressible fluid flow that remains attached to the

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

glottal wall and has negligible viscous loss. Bernoulli's law is not valid for compressible flows (variable air density), but the glottal air flow may be considered incompressible for Mach numbers < 0.3. Actually, the compression rate of air at the glottal level is limited: for speaking voices, the volume changes due to air compression are in the range 1–2 % (subglottal pressures of 10 to 20 hPa). More importantly, flow separation from the wall and vortex formation in a divergent glottal duct also causes a departure from Bernoulli's law. The pressure does not recover completely in an expanding (diverging) duct (expansion angle >5°) [3].

In adult larynges during modal exhalatory phonation, the glottis takes on - during each open phase - three successive shapes: convergent, uniform and divergent (Fig.2), the uniform shape being only the brief transition from the convergent one to the divergent one.

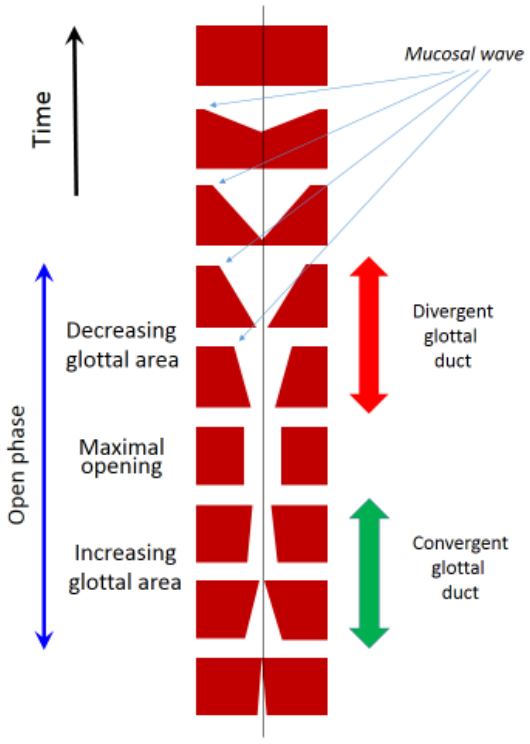


Fig. 2. Schematic frontal section at midpoint of the glottis showing the two parts of the open phase.

The estimate of intraglottal pressure according to Bernoulli's law may be considered valid during the opening phase. If flow separation from the glottal wall as well as flow vorticity occur during the closing phase, Bernoulli's law becomes less valid and the intraglottal pressure is to some extent affected by the supraglottal acoustic pressure, which modifies the overall pressure distribution in the glottis. If flow separation and supraglottal acoustic pressure are included, the computation of intraglottal pressure can theoretically be divided into two parts, one upstream and the other one downstream of the flow separation:

upstream: $P = P_s - k\,(1/2)\,\rho v^2$ (2) (derived from Eq. 1)

downstream: $P = I\,dU/dt$ (3)

where P_s is subglottal pressure (the lung pressure created by contraction of expiratory muscles and/or recoil of thoracic elastic elements), k is a pressure loss coefficient for glottal entry and viscous drag, I is supraglottal acoustic inertance, and U is airflow. The value of k was set at 1.37, according to van den Berg & al. [4] and Fulcher & al.[5], for soft to medium voicing. Actually, eq. (3) intervenes as soon as the glottis opens: dU/dt is mainly positive during glottal opening (when flow is increasing) and mainly negative during glottal closing (when flow is decreasing). The inertance of an air column is defined as the air density multiplied by the length of the column (along the direction of acceleration or deceleration) and divided by its cross-sectional area (perpendicular to the acceleration or deceleration). The inertance I can be estimated as

$$I = \rho L/S \qquad (4)$$

where S is the cross-sectional area of the supraglottal air column and L its effective length [6]. Inertance can be thought of as density of an air column per unit length. Units are g/cm^4 or kg/m^4 (1 g/cm^4 = 10^5 kg/m^4). L and S may be considered as constant during emission of a sustained vowel, as is the case in our experiments. Inertance also depends on frequency, but this is not relevant here. When written out in terms of Newton's second law of motion, which states that force = mass x acceleration, Eq. (3) means that
Vocal tract input pressure = (inertance) x (acceleration of the air column).
Force is analogous to the vocal tract input pressure, mass is analogous to inertance, and acceleration remains the same. Values of L and S were chosen according to Titze [6].

Exactly defining when and to what extent equations (2) and (3) are applicable is impossible, but one may expect that equation (2) carries greater weight.

Since the work of van den Berg & al. [4], much attention has been paid to possible negative values of the intraglottal pressure during the closing phase. However, mechanically, the only requirement for maintaining the vocal folds in oscillatory motion is that the driving force be less positive during closing (including tissue recoil) than during opening, and that the net driving force over the whole cycle be sufficient to overcome frictional forces. The important point is the asymmetry of the

pressure curve between the opening portion and the closing portion of the cycle.

In a previous paper [7], the focus was mainly put on the ratio [P during opening phase/P during closing phase], in order to investigate the effect of voice intensity on the ratio. The present work is an attempt to quantify the intraglottal pressure values during the open phase in different intensity conditions by using *in vivo* calibrated flow and area measurements, and applying Eq. (2) (mainly suited for upstream of flow separation) and Eq. (3) (mainly suited for downstream of flow separation).

The subject was a healthy trained vocalist, whose average subglottic pressure values as function of intensity have been investigated previously [8].

II. METHODS

Glottal area

The glottal area was derived from a photometric record, obtained by transilluminating the trachea. The light flux was detected by a photovoltaic transducer in the pharynx. The transducer, a BP104 Silicon Photodiode (Vishay Precision Group, Malvern, PA), was glued onto a small laryngoscopic mirror (Nr. 3), the handle of which was introduced - together with the sensor lead - through the hermetically sealed hole normally intended for the handpiece of a Rothenberg mask [7]. The current produced by the photodiode was preamplified by a current-to-voltage converter with a linear response up to 2 kHz. Calibration was described earlier [7].

Transglottal flow

The glottal flow waveform (flowglottogram) was recorded using a Rothenberg mask and the MSIF2 inverse filtering system of Glottal Enterprises (Syracuse, NY). The mask is equipped with a compressible seal and is firmly pressed against the face of the subject to avoid any air leakage. Again, the calibration procedure was described earlier [7].

Acoustic signal

A small condenser microphone (Ø 5.6 mm) was fixed laterally inside the Rothenberg mask, exactly fitting an opening of the mask opposite the pressure transducer. SPL of the voice samples were evaluated using the *Praat* software (www.praat.org). The microphone sound levels were calibrated with a Wärtsilä 7178 sound level meter in a position corresponding to a direct measurement at 10 cm from the lips.

All signals were recorded using a 4-channels Pico Scope 3403D module (Pico Technology Ltd, St Neots, England, UK) and stored in a PC computer.

Three typical voicing conditions were selected for detailed analysis: 62.35, 68.60 and 74.70 dB (10cm from the lips), at an average speaking frequency of around 110 Hz.

III. RESULTS & DISCUSSION

The area, flow and intraglottal pressure curves for the three typical conditions are illustrated in Figs. 3, 4 & 5. The area curves define the separation between the opening phase and the closing phase. Intraglottal pressure was calculated according to Eq. 2 (upstream) and 3 (downstream). It can be seen that whatever the voicing condition and the equation used, the average intraglottal pressure is systematically larger during the opening phase (convergent duct) than during the closing phase (divergent duct).

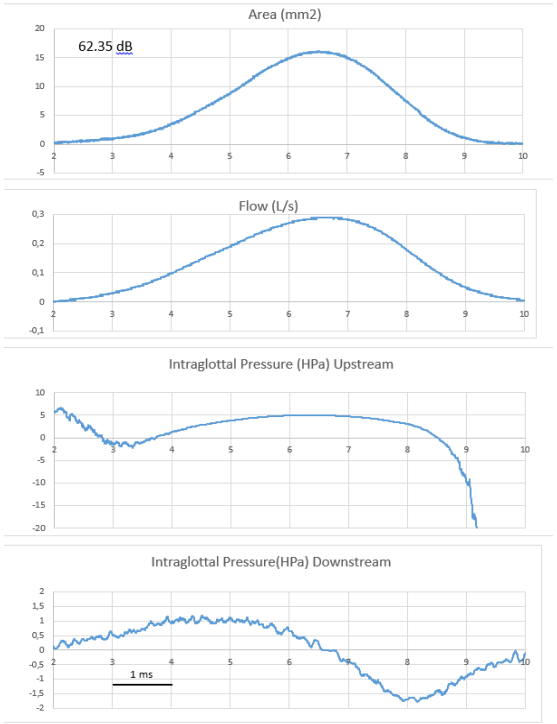


Fig.3. Top to bottom: Glottal area, flow and intraglottal pressure calculated by equations 2 and 3. The closed phase is very short and limited skewing of the flow trace is visible. With Eq. 2 as well as with Eq. 3, the area under the pressure curve is obviously larger during the opening than during the closing phase, even becoming negative during the closing phase. Maximum intraglottal pressure is about 5 hPa.

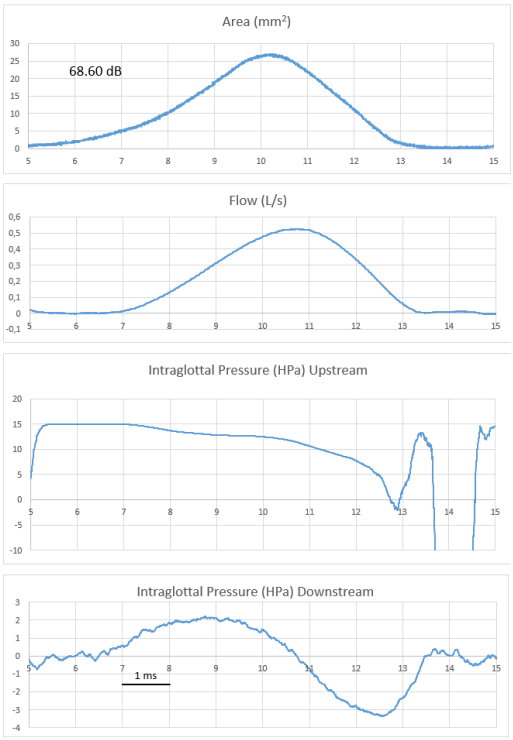


Fig.4. Top to bottom: Glottal area, flow and intraglottal pressure computed by equations 2 and 3. Same comment as in Fig. 3. Maximum intraglottal pressure is about 15 hPa.

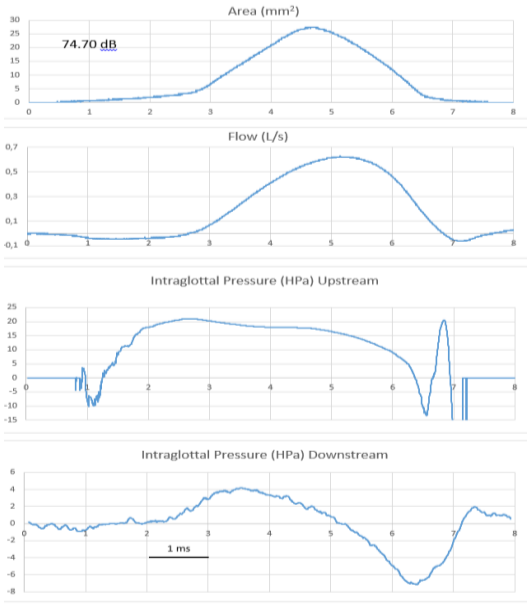


Fig.5. Top to bottom: glottal area, flow and intraglottal pressure computed by equations 2 and 3. Same comment as in Fig. 3. Maximum intraglottal pressure is about 21 hPa.

IV. CONCLUSION

The results confirm *in vivo* the data obtained by modelling: over one whole cycle, the driving force performs net positive work, accounting for sustained vocal fold motion. When the airflow curve is skewed to the right with respect to the glottal area curve, whatever the glottal duct configuration, the intraglottal pressure during the opening phase systematically exceeds that during the closing phase, the basic condition for sustaining oscillation. The skewing results from air compressibility and vocal tract inertance. Quantitatively, the intraglottal pressure becomes negative during the closing phase. Importantly, the general downward trend of the mean intraglottal pressure during the open phase of the cycle matches the downward trend of the tissue velocity; in other words the tissue displacement shows a phase delay of $\pi/2$ radians (in ideal conditions, without friction) with respect to the driving force.

REFERENCES

[1] P.H. DeJonckere , J. Lebacq, "Phase relationship between dynamics of the subglottic pressure and oscillatory movement of the vocal folds. I. Sustained phonation." *Arch. Internat. Physiol. Bioch.* Vol. 88: pp. 31 – 32, 1980.

[2] I.R. Titze , *Principles of voice production*. 2nd Printing. National Center for Voice and Speech. Iowa City IA USA, 2000.

[3] E.M. Sparrow, J.P. Abraham, W.J. Minkowycz, "Flow separation in a diverging conical duct: Effect of Reynolds number and divergence angle." *International Journal of Heat and Mass Transfer* vol. 52: pp. 3079 - 3083, 2009.

[4] Jw. Van den Berg , J. Zantema , P. Doornenbal Jr, "On the air resistance and the Bernoulli effect of the human larynx." *J Acoust Soc Am.* vol. 29 : pp. 626 – 631, 1957.

[5] L. Fulcher, R.C. Scherer, N. Anderson. "Entrance loss coefficients and exit coefficients for a physical model of the glottis with convergent angles." *J Acoust Soc Am.* vol. 136/3: pp. 1312 – 1319, 2014.

[6] I.R. Titze, Acoustic interpretation of resonant voice. *JVoice* vol. 15: pp. 519-528, 2001.

[7] P.H. DeJonckere , J. Lebacq, I. Titze, Dynamics of the driving force during the normal vocal fold vibration cycle. *JVoice* in press, 2017.

[8] P.H. DeJonckere, J. Lebacq, L. Bocchi, S. Orlandi C. Manfredi, "Automated tracking of quantitative parameters from single line scanning of vocal folds: a case study of the 'messa di voce' exercise", *Logop. Phoniatr. Vocol.* vol. 40: pp. 44–54, 2015.

KINEMATIC MODEL FOR SIMULATING MUCOSAL WAVE PHENOMENA ON VOCAL FOLDS

P. K. Subbaraj[1], J. G. Svec [1]

[1]Voice Research Lab, Department of Biophysics, Faculty of Science, Palacký University, Olomouc, Czech Republic.
pravin.subbaraj@upol.cz, jan.svec@upol.cz

Abstract: **Mucosal waves have been found of crucial importance for evaluating vocal fold vibrations in laryngological practice. While they are routinely evaluated visually, the knowledge on the physical phenomena related to mucosal wave propagation is limited. Kymographic imaging in particular reveals various mucosal wave features that deserve more understanding in order to advance diagnostics of voice disorders. Here, a kinematic model is presented which is intended for simulating mucosal waves on human vocal folds. The vibration characteristics including the mucosal wave movements are then visualized using a synthetic kymogram generated by local illumination method.**

Keywords : **Mucosal wave, vocal fold vibrations, kinematic model, synthetic kymogram**

I. INTRODUCTION

Mucosal wave is an important parameter in the diagnosis of voice disorders as it reveals about the pliability and healthiness of the vocal folds. It originates from the lower margins of the vocal fold during phonation and travels to the upper margins, creating a wave like motion on the vocal fold surfaces [1].

The presence and extent of mucosal wave reveals certain information about the vocal fold characteristics which helps clinicians in the diagnosis of voice pathology [2]. For instance, the reduction in the mucosal wave amplitude may indicate the presence of vocal fold lesions and scarred tissues [3].

Mucosal wave is effectively assessed through visual inspection of the kymographic images [2, 4]. Though these images are often successful in exhibiting various mucosal wave features, physical phenomena of mucosal wave propagation are not yet completely understood.

This study attempts to simulate mucosal waves on kinematic model of the human vocal folds. A synthetic kymogram generated by local illumination method is then used to visualize the vibratory characteristics including the mucosal wave movements.

II. METHODS

The vocal fold geometry is constructed based on the specifications of the so called M5 model [5, 6]. Mucosal wave velocity is then related to the vocal fold vibration amplitude using the formula

$$V_m = 2\pi f_o \tau A_u \tag{1}$$

where τ is a factor relating the mucosal wave velocity to the maximum opening speed of the vocal fold upper margin, f_0 is the fundamental frequency, and A_u is the amplitude of upper margin of the vocal fold.

Mucosal wave motion is modelled as

$$x_i = x_{0i} + A_i \sin(2\pi f_0)\left(t - \frac{D}{V_m}\right),$$

$$y_i = y_{0i} + A_i \cos(2\pi f_0)\left(t - \frac{D}{V_m}\right), i = (1,2,..N) \tag{2}$$

where t is the time instant, $D = 0.01\,\text{cm}$ is the equidistance between the samples defining the vocal fold surface, (x_0, y_0) and (x_i, y_i) are the initial and final vocal fold coordinates during mucosal wave propagation, and A_i is the vibration amplitude varying from lower to upper margin according to the equations

$$A_i = 0 \quad \forall i \in (K, Z),$$

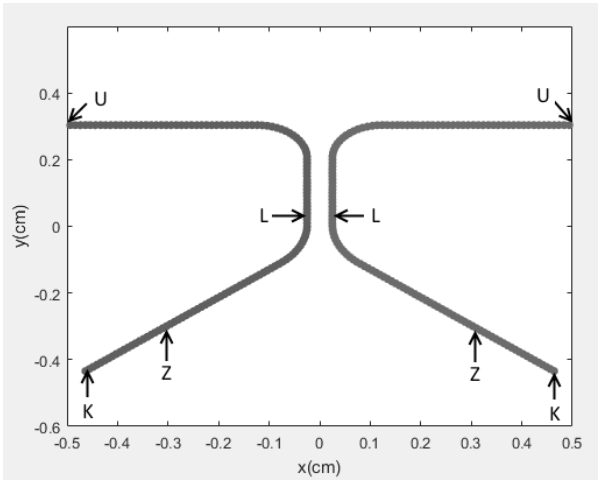
$$A_i = \left(\frac{A_u}{L}\right) i \quad \forall i \in [Z, L],$$

$$A_i = A_u \quad \forall i \in (L, U) \tag{3}$$

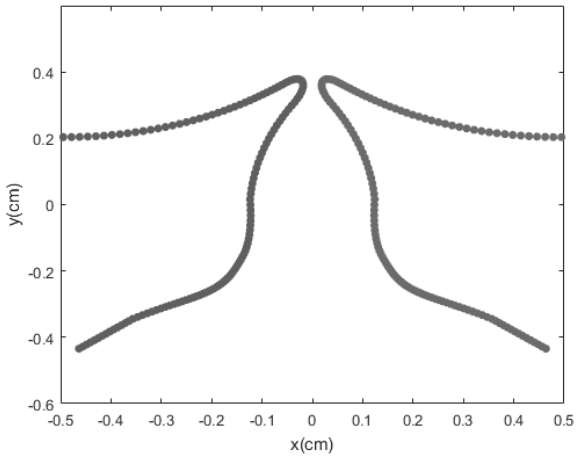
where K is the index of the low end of the model surface, Z is the index of the point at which the subglottal mucosa starts vibrating, L is the index of the vocal fold lower margin at which $y = 0$, and U is the index of the vocal fold upper margin as defined by Li et al. [5], The indexes used in (3) are as shown in Fig.1A.

Laryngoscopic image of the vibrating model is then simulated using the Radiance and Phong specular reflectance term known in computer graphics for rendering realistic specular highlights, for calculating

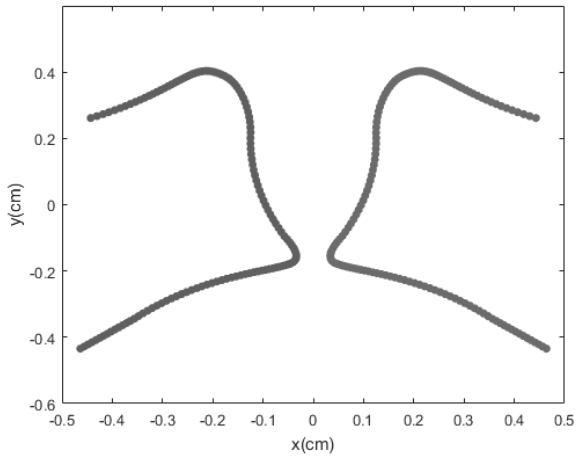
Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)



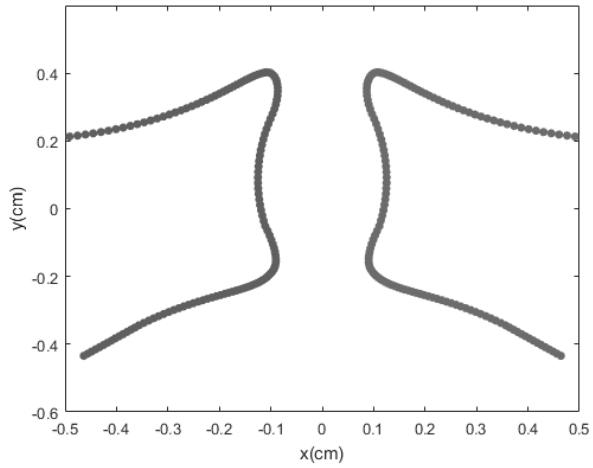
A



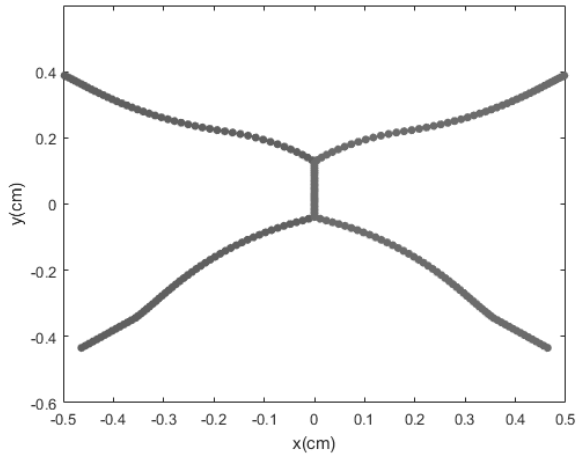
D



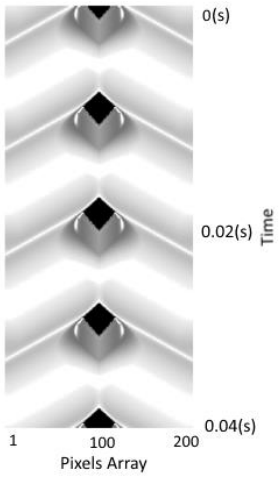
B



E



C



F

Fig.1. Kinematic (M5-shaped [5]) model of human vocal fold in neutral position (A), exemplary motions of the model under the influence of mucosal wave propagation through its surface (B-E) and kymographic image (F) computationally generated by capturing surface motions from the top (viewing direction orthogonal to the model). [7]. Exemplary images are chosen to illustrate closing (B), complete closure (C), opening (D), and maximally open (E) phases of the glottal cycle.

surface illumination, and for determining pixel values. From there, kymographic image is computationally generated by dynamically mapping the mucosal wave motion to pixels.

III. RESULTS

Fig.1 provides an example of mucosal wave motion on the model by specifying $f_0 = 100\,\text{Hz}$, $A_u = 0.1\,\text{cm}$ and $\tau = 1.6$, and an example of the resulting kymogram. Fig. 1A shows the vocal fold model in neutral position prior to vibration. Figs. 1B-E show examples of model kinematics representing different phases of a glottal cycle.

Fig. 1F shows the kymographic image algorithmically computed from the point of view of a virtual line camera orthogonally placed above the glottal midline of the vocal fold model.

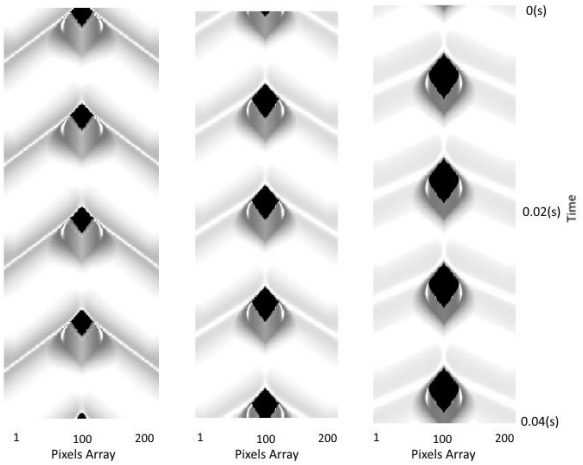
Time is incremented in steps of 1/sampling rate till it reaches sampling rate/number of frames per second (default values of sampling rate and number of frames per second are 7200 Hz and 25 fps respectively [8].

In a given time instance, pixels array positions 1 to 200 with the step size of 1 are mapped to the vocal fold points in the range -0.5 cm to 0.5 cm with the step size of 0.005 cm. Pixel value in a given vocal fold points range is calculated from the slopes of the points in that range and normalized differences of their positions in the x – coordinate, followed by local illumination calculations. In case no vocal fold points are present in a range then the respective array is assigned with the pixel value of the immediate adjacent non-empty range.

Pixel array values computed are stored in a single row of the synthetic kymographic image. Similarly, the next rows are updated with the pixel array values computed for the remaining time instances. Thus, a total of 288 rows (sampling rate / number of frames per second = 7200/25 = 288) are created covering all time instances. Each of these rows is duplicated in order to imitate the behavior of the 2nd generation videokymograpic cameras as they store each scan lines twice [9]. This makes the final dimension of the kymographic image as 200 columns X 576 rows.

Fig. 2 demonstrates the changes in the kymogram when increasing the speed of the mucosal wave by increasing the parameter τ as defined in (1). The increasing speed of the mucosal wave is visible in the kymogram through the varying slope of the contour of the mucosal wave travelling over the upper surface of the vocal folds laterally – the slope of the contour is changing towards the horizontal.

Furthermore, it can also be observed that the increasing mucosal wave speed causes the width of the mucosal wave contour to increase. This reflects an increasing width of the mucosal wave tissue bulge on the upper vocal fold surface.



(a) (b) (c)

Fig. 2 Examples of vocal fold kymograms with $\tau = 1.4$ (a), $\tau = 1.8$ (b) and $\tau = 2.6$ (c). Notice the differences in mucosal wave speed (i.e. the changes in diagonality of the mucosal wave contour) and in the breadth of the mucosal wave contour.

IV. DISCUSSION AND CONCLUSION

The developed model allows relating mathematically defined mucosal wave features to kymographic images.

This approach is useful for deeper understanding of the appearance of mucosal waves on the vocal folds and their variability when changing the driving parameters. The model can also be used to improve and verify the algorithms for automatic image analysis of vocal fold vibrations and for detecting the mucosal wave phenomena in clinically obtained kymographic images. This is desirable for advancing the diagnostic possibilities of kymographic imaging in laryngology.

V. ACKNOWLEDGEMENT

The work has been supported by the Czech Science Foundation (GA CR) project no. GA16-01246S. The authors greatly acknowledge Dr. Jan Nauš, Department of Biophysics, Faculty of Science, Palacký University, Olomouc, and Mr. Lukáš Vrajík, ConfigAir, Olomouc, for their helpful suggestions on the calculation of illumination of the synthetic kymogram.

REFERENCES

[1] I. R. Titze, J. J. Jiang, and T. Y. Hsiao, "Measurement of Mucosal Wave-Propagation and Vertical Phase Difference in Vocal Fold

Vibration," *Annals of Otology Rhinology and Laryngology*, vol. 102, pp. 58-63, Jan 1993.

[2] C. R. Krausert, A. E. Olszewski, L. N. Taylor, J. S. McMurray, S. H. Dailey, and J. J. Jiang, "Mucosal Wave Measurement and Visualization Techniques," *Journal of Voice*, vol. 25, pp. 395-405, Jul 2011.

[3] D. M. Bless, M. Hirano, and R. J. Feder, "Videostroboscopic evaluation of the larynx," *Ear Nose Throat J*, vol. 66, pp. 289-96, Jul 1987.

[4] J. G. Svec, F. Sram, and H. K. Schutte, "Videokymography in voice disorders: What to look for?," *Annals of Otology Rhinology and Laryngology*, vol. 116, pp. 172-180, Mar 2007.

[5] S. Li, R. C. Scherer, M. X. Wan, and S. P. Wang, "The effect of entrance radii on intraglottal pressure distributions in the divergent glottis," *Journal of the Acoustical Society of America*, vol. 131, pp. 1371-1377, Feb 2012.

[6] R. C. Scherer, D. Shinwari, K. J. De Witt, C. Zhang, B. R. Kucinschi, and A. A. Afjeh, "Intraglottal pressure profiles for a symmetric and oblique glottis with a divergence angle of 10 degrees," *Journal of the Acoustical Society of America*, vol. 109, pp. 1616-1630, Apr 2001.

[7] M. Pharr and G. Humphreys, *Physically Based Rendering (Second Edition)* Morgan Kaufmann, 2010.

[8] Q. Qiu and H. K. Schutte, "Real-time kymographic imaging for visualizing human vocal-fold vibratory function," *Rev Sci Instrum*, vol. 78(2): 024302, Feb 2007.

[9] J. Sedlar, "Image Analysis in Microscopy and Videokymography," PhD Dissertation, Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Charles University, Prague, 2012.

FUNCTIONAL MODELS OF THE NEURAL CAUSES AND INTRA-FOLD MODULATION OF VOCAL FREQUENCY JITTER

J. Schoentgen[1]
[1] Bio-, Electro- and Mechanical Systems, Faculty of Applied Sciences, Université Libre de Bruxelles, Brussels, Belgium

jschoent@ulb.ac.be

Abstract: **The presentation is devoted to functional models of the generation and modulation of vocal jitter. A TA muscle twitch model simulates vocal frequency jitter in terms of muscle tension jitter, which is the outcome of the concurrent activity of several motor units. The control parameters of the model are the dead time and firing rate of the motor neurons, the number of active motor units as well as the duration and shape of the muscle fiber twitch. The presentation also includes a discussion of a possible fold-internal modulation of vocal jitter, which is unrelated to motor unit activity per se.**
Keywords: **Vocal jitter, TA-muscle jitter, body-cover fold model.**

I. INTRODUCTION

The objective of the presentation is to discuss models of the neural causes as well as the intra-fold modulation of vocal frequency jitter. The models are numerically compact so that they may be used for speech synthesis.

Lists of possible physiological causes of vocal jitter have been published that refer to a wide range of vocal irregularities most of which are better known under names other than vocal jitter.

Kreiman et al. include apart from muscle tension jitter the following as physiological causes of (vocal frequency) jitter: asymmetries in muscle tension, randomness in sub-glottal pressure and trans-glottal airflow, mucous on the folds, regional blood flow as well as tongue pull [1]. They also mention "perturbations in muscular innervation", but it is not clear from the context what conditions they refer to.

Left-right fold asymmetries cause left-right phase shifts when small and diplophonia or biphonation when large. Random fluctuations of air pressure and flow rate are expected to cause additive noise rather than modulation noise given their small size. Reported perturbative influences of mucous on the folds appear to be a misunderstanding [2]. Regional blood flow owing to the heart cycle is one of the causes of physiological tremor and the effect of tongue pull is known as the intrinsic vocal frequency, which is a special case of micro-prosody.

Generally speaking, jitter designates deviations from true periodicity of a presumably periodic signal. The perturbations must be rapid (> 10 Hz) and jitter may refer to perturbations of any quantity (amplitude, cycle length, instantaneous frequency, etc.) [3][4].

Vocal frequency jitter may be measured reliably and is expected to be a salient feature of human vocal timbre if (i) only the true vocal folds vibrate; (ii) the vibrations stay within the same vocal mechanism (aka vocal register); (iii) the vibrations are pseudo-periodic and monophonic (i.e. "chaos" or biphonation or diplophonia are absent); (iv) vocal tremor and breathiness are "negligible" and (v) the intonation contour is "flat".

Acoustic cues reporting vocal frequency jitter are popular as well as criticized. They are popular because they are expected to be relevant descriptors of the voice quality of a majority of speakers. They are criticized because users may "overlook" some of the conditions of applicability cited above, but also because these conditions suggest that vocal frequency jitter may contribute negligibly to the more extreme or spectacular voice qualities.

One expects a skeletal muscle force to be jittered because it is the outcome of the superposition of many individual muscle twitches. Muscle tension jitter of the thyro-arytenoid (TA) muscle therefore is the most likely source of vocal frequency jitter when the true vocal folds vibrate exclusively, pseudo-periodically and monophonically. Force jitter of muscles other than the TA muscle, e.g. the crico-thyroid (CT) muscle, may be disregarded in a first approximation because of the inertia of the thyroid cartilage that is expected to smooth CT jitter.

A (skeletal) muscle force is the effect of the concurrent activity of many motor units. A motor unit is a motor neuron that innervates a group of muscle fibers that contract synchronously when the motor neuron emits an electrical spike. The synchronous contraction is called a muscle twitch. The overall muscle force is the outcome of the co-activity of many muscle twitches that overlap in space (owing to the

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

concurrent activity of several motor units) and time (owing to the rapid firing of a single motor neuron).

Titze has proposed in 1991 a model of TA muscle tension jitter in vocal fold vibration [5]. Components of the model are (i) a three-parameter model of the time course of a muscle twitch, (ii) the linear summation over time of several twitches that follow each other at variable time intervals (simulating the activity of one motor unit) and (iii) the linear summation of several twitch sequences that occur asynchronously (simulating the concurrent activity of several motor units). The emission of spikes by the motor neurons is not modeled explicitly.

The muscle twitch model involves the multiplication of three elementary curves and three control parameters. The parameters fix the amplitude as well as the rise time and time origin of the twitch. Observations of canine and human muscle twitches suggest fixing the rise time between 20 and 30 *ms*. The decay time has a default value.

The simulation of the activity of one motor unit involves (i) obtaining random inter-twitch time intervals followed by (ii) the generation and (iii) summation of the individual twitches. The result is a sequence of twitches that overlap when the inter-twitch time intervals are short. The inter-twitch time intervals, which correspond to putative inter-spike time intervals of the firing motor neuron, are obtained via a Gaussian distribution that has been fitted to the histogram of the observed inter-spike intervals from a single motor unit of a normal subject's TA muscle. The typical inter-spike intervals are between 30 *ms* and 100 *ms*. Titze considers coefficients of variation of the inter-spike intervals between 10 % and 100 % with a preference for small percentages.

The concurrent activity of several motor units is simulated via a second sum. A second Gaussian is considered that shifts single motor unit twitch sequences one with regard to the other to take into account their asynchronous activity.

To sum up, possible control parameters are the twitch amplitude and rise time, the inter-twitch duration as well its coefficient of variation and the number of concurrently active motor units. The twitch amplitude is believed to depend on the number of muscle fibers innervated by one motor neuron.

Practical and theoretical issues with Titze's model are the following. First, the model is numerically cumbersome and not well suited for speech synthesis because it requests a large overhead as well as the generation of thousands of distinct twitch models for one second of speech.

Second, inter-twitch intervals are drawn randomly from a Gaussian distribution even though time intervals must be positive by definition. The number of unacceptable interval lengths increases with the coefficient of variation. For instance, 16 % of the inter-twitch intervals are negative when the Gaussian distribution is centered on a typical interval and the coefficient of variation is 100 %. This means that the interval lengths must be tracked and those that are not acceptable must either be discarded or shifted to positive values. This intervention changes the statistical properties of the twitch sequences uncontrollably.

Third, Titze does discuss the dead time of the muscle fibers, but not the dead time of the motor neurons. The former can be neglected, but not the latter. That is, the Gaussian distribution is problematic not only because it predicts negative inter-twitch intervals, but also inter-twitch intervals that are so short that they are physiologically unlikely because of the refractive period of the motor neurons. Possible dead times can be observed in the histogram reported by Titze that displays inter-spike intervals from a single motor unit of a human TA muscle. Extra-short intervals (< 30 *ms*) are missing. This observation must, however, be taken with a grain of salt, because the observed number of intervals was small. Expected durations of dead times are circa 5 *ms* [6].

Fourth, Titze's model involves muscle twitches only. The activity of the motor neurons indirectly enters the picture only via the number of motor units, the typical inter-twitch length and its coefficient of variation. The percentage of vocal frequency jitter that is so generated is in the interval of observed jitter, i.e. 0.1 – 1 % and so indirectly confirms the relevance of the model. However, laryngeal conditions are known to exist that influence perceived hoarseness and measured jitter, but which are unlikely to influence directly the activity of motor units. A possible mechanism that would enable controlling the gain of jitter fold-internally and that is not directly related to neural activity is therefore discussed in section III.

Ignoring dead times and having recourse to Gaussian distributions to simulate inter-spike intervals was less likely to cause raised eyebrows in 1991 than today. The reason is that till recently the Poisson process, the inter-event intervals of which can be approximated by Gaussian distributions under favorable conditions, was believed to be a universal point process. That is, a superposition of point processes was thought to give rise to a Poisson process. However, this appears not to be the case. The point process resulting from the superposition of several point processes is not known in general [7].

II. A NUMERICAL MODEL OF THE NEURAL CAUSES OF VOCAL JITTER

A. Preliminaries

The model of muscle tension jitter that is discussed hereafter is inspired by [7]. The model takes into account that motor neurons have a dead time and it is free of assumptions that would generate negative inter-spike time intervals. Also, a single point process replaces the simulation of the collective activity of many individual motor units.

General properties that are made use of hereafter are the following.

(1) The (spatial) superposition of the muscle twitches can be replaced by the superposition of the spikes of distinct motor neurons when the spike-to-twitch models are linear and do not differ too much from each other.

(2) A superposition of several Poisson processes with dead time is another Poisson process with dead time [7].

(3) A superposition of N Poisson processes with a firing rate λ (i.e. the inverse of the average inter-spike interval) is equivalent to another Poisson process with a firing rate equal to $\lambda \times N$ [7].

B. Muscle twitch model

The observed shape of a muscle twitch (i.e. a rapid rise from zero followed by a slow decay towards zero without undershoot) suggests simulating a muscle twitch via the unit pulse response of a linear filter. The filter is a second-order Butterworth low-pass the control parameter of which is the cut-off frequency. The unit pulse response is a large positive impulse the rise time of which is shorter than its decay time. The large positive pulse is followed by a small negative undershoot that does not exceed in absolute value 5 % of the positive pulse amplitude. The negative undershoot is not observed in natural muscle twitches and is considered to have a negligible influence on the simulation of muscle tension jitter. Typical muscle twitch lengths are between 80 and 100 *ms*, which correspond to cut-off frequencies of the filter between 6 *Hz* and 8 *Hz* when the length is considered over which the pulse is positive.

The auto-regressive coefficients a_i and gain g_0 of the filter are the following when the cut-off frequency f_c and sampling step Δt are given [8].

C. Single motor neuron function

The function of a single motor neuron is simulated by means of a Poisson process with dead time d (~5 *ms*).

$$u = 1/\tan(\pi f_c \Delta t)$$
$$w = 2\cos(\pi/4)$$
$$a_0 = 1/(1 + uw + u^2)$$
$$a_1 = 2a_0(1 - u^2)$$
$$a_2 = a_0(1 - uw + u^2)$$
$$g_0 = (1 + a_1 + a_2)$$

The dead time takes into account that once a motor neuron has fired it enters into a refractory period during which it cannot fire again. In practice, the Poisson processes is used to generate at each time step a number 0, 1, 2, etc. equal to the number of spikes emitted within time interval Δt. No spike is emitted when this number is 0, a unit pulse is emitted when this number is 1, a pulse of twice the amplitude is emitted when this number is 2, etc.

The emitted spikes are inputted to the linear second-order filter to simulate the muscle twitches that are due to the spikes emitted by a single motor neuron.

D. Collective motor neuron function

When more than one motor unit is active, their muscle twitches are superimposed. The superposition of the twitches can be replaced by a superposition of the spikes because of the linearity of the twitch model. The relevance of the switch of filter and sum is that the superposition of twitch sequences is replaced by a superposition of spike sequences that can be simulated by a single process [7]. That is, a collective of motor units may be simulated by means of a single Poisson process with dead time emitting spikes that are filtered to generate the jittered muscle tension. Muscle tension jitter is used to simulate vocal frequency jitter after normalization and gain control.

The following is an implementation of the algorithm given by Deger et al. [7] that generates the number Q of spikes emitted at each time step when n Poisson processes with dead time $D\Delta t$ are superimposed. This number is broken up into $n = n_E + n_1 + n_2 + \ldots n_D$. Symbol n_E designates the number of neurons that are able to emit and n_1 designates the number of neurons that are in their refractive state since one time step, n_2 since two time steps and n_D since D time steps after which they become able to emit in the next time step. Symbol λ_c designates the collective firing rate.

$$Q \leftarrow Binomial(n_E, \lambda_c \Delta t)$$
$$n_E \leftarrow n_E - Q + n_D$$
$$n_{i+1} \leftarrow n_i, \text{shift: } i = D - 1 \ldots 1$$
$$n_1 \leftarrow Q$$

III. A FUNCTIONAL MODEL OF THE INTRA-FOLD MODULATION OF VOCAL JITTER

The following is speculative. It is an attempt to discuss a possible intra-fold mechanism that would modulate vocal frequency jitter. It is indeed the case that laryngeal conditions are known to exist that are expected to influence perceived hoarseness and measured jitter, but which are unlikely to influence directly the activity of motor units. For instance, one may observe that light laryngitis, menstruation, vocal loading in dry air or while consuming dehydrating beverages or injection of atropine in the folds may increase jitter and perceived hoarseness. An increase of the *passive* tension of the folds may decrease jitter, however.

The model assumes that the body and the cover of a vocal fold vibrate sinusoidally at the same frequency, but at different amplitudes A_B and A_C. The instantaneous phases φ of the body and cover are assumed to be the same up to a slight perturbation of the phase of the fold body owing to vocal jitter. The instantaneous phase of the fold cover is assumed to be unjittered.

The movement of the fold edge is then the sum of the sinusoidal motions of the body and cover when the folds do not touch. A model of the glottal entrance width $w_{g,entr}$ may then be rewritten according to [9] as follows, assuming left-right symmetry.

$$w_{g,entr} = 2 \times \max(0, A_{abd} + A_{cov} \sin \varphi_{cov} + A_{bod} \sin \varphi_{bod})$$

$$\theta_{jit} = \varphi_{cov} - \varphi_{bod}$$

A rule of elementary trigonometry enables transforming the sum of the two sines as follows [10].

$$A_{cov} \sin \varphi_{cov} + A_{bod} \sin \varphi_{bod} = A_{cov} \sqrt{1 + 2 \times \frac{A_{bod}}{A_{cov}} \cos \theta_{jit} + \frac{A_{bod}^2}{A_{cov}^2}} \times \sin(\varphi_{cov} - \tan^{-1} \frac{\frac{A_{bod}}{A_{cov}} \sin \theta_{jit}}{1 + \frac{A_{bod}}{A_{cov}} \cos \theta_{jit}})$$

Taking into account that θ_{jit} is small enables inserting the following approximations.

$$\cos \theta_{jit} \approx 1 \quad \sin \theta_{jit} \approx \theta_{jit} \quad tan^{-1} \theta_{jit} \approx \theta_{jit}$$

One so obtains an expression that suggests that the instantaneous frequency jitter of the glottal width is the instantaneous frequency jitter of the fold body weighted by the amplitudes of vibration of the cover and body. The instantaneous frequency jitter is obtained by taking the temporal derivative of the instantaneous phase.

$$w_{g,entr} = 2 \times \max[0, A_{abd} + (A_{cov} + A_{bod}) \times \sin(\varphi_{cov} - \frac{A_{bod}}{A_{bod} + A_{cov}} \theta_{jit})]$$

$$f_{fold} = f_{cov} - \frac{1}{1 + \frac{A_{cov}}{A_{bod}}} \Delta f_{jit}$$

The previous expression suggests that the frequency jitter of the body of the folds is weighted so that a decrease of the amplitude of the cover or an increase of the amplitude of the body increase observed jitter. This suggests that an increase of the viscosity of the cover may increase observed jitter and that a (passive) increase of the stiffness of the body may decrease observed jitter.

REFERENCES

[1] J. Kreiman, D. Sidtis, "*Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception*," Wiley-Blackwell, 2011, p. 55.

[2] G. A. Bryant, M. G. Haselton, "Vocal cues of ovulation in human females," *Biol. Lett.* vol. 5, 2009, 12–15.

[3] F. Hargrave, *Hargrave's Communication Dictionary*, Wiley-IEEE Press, 2001, p. 283

[4] "Jitter," in *Wikipedia: The Free Encyclopedia*. Wikimedia Foundation, Inc. Retrieved 11 September 2017, en.wikipedia.org/wiki/Jitter.

[5] I. R. Titze, "A Model for Neurologic Sources of Aperiodicity in Vocal Fold Vibration," *J. Speech and Hearing Res.,* vol. 34, 1991, 460-72.

[6] R.M. Roark, C.L. J. Li, S. Schaefer, A. Adam, C. De Luca, "Multiple Motor Unit Recordings of Laryngeal Muscles: The Technique of Vector Laryngeal Electromyography," *The Laryngoscope*, vol. 112, 2002, 2196-2202.

[7] M. Deger, M. Helias, C. Boucsein, S. Rotter, "Statistical properties of superimposed stationary spike trains," J. Comput. Neurosci., vol. 32, 2012, 443-463.

[8] U. Zölzer, "*DAFX - Digital audio effects*," Wiley, 2002, p. 34.

[9] I. R. Titze, "*The Myoelastic Aerodynamic Theory of Phonation*," The National Center for Speech and Voice, 2006, p. 259.

[10]"List of trigonometric identities," in *Wikipedia: The Free Encyclopedia*. Wikimedia Foundation, Inc. Retrieved 11 September 2017, en.wikipedia.org/wiki/List_of_trigonometric_identities

SESSION VII:

VOCAL FOLDS DYNAMICS III

CALIBRATION OF EXTERNAL LIGHTING AND SENSING PHOTOGLOTTOGRAPH

A. Bouvet[1], A. Van Hirtum[1], X. Pelorson[1], S. Maeda[2], K. Honda[2], A. Amelot[2]

[1] CNRS UMR216, Gipsa-lab, Université Grenoble-Alpes, Grenoble, France
[2] Laboratoire de Phonétique et Phonologie (LPP), Université Paris 3, Paris, France
anne.bouvet@grenoble-inp.fr

Abstract: **Observation and measurement of vocal folds vibration and glottal opening during speech requires techniques as little invasive as possible for the subject. The LPP has developed the External Photoglottograph (ePGG) system. It consists of illuminating the glottis through the neck skin with an infrared light and recording light variation intensity modulated by glottal movement with a photodiode placed across the larynx. The system is tested on two mechanical larynx replicas. The first one consists of two rigid half cylinders in forced oscillation controlled by a step motor. The second one is flow driven and uses latex tubes filled with water in order to reproduce vocal folds self-oscillation. Time-varying glottal area is measured accurately for both replicas. Experimental results are compared to ePGG recordings in order to assess the correlation between area measurements and ePGG signal. This characterization is used to propose a calibration of the glottal opening as a function of parameters affecting the ePGG signal (distance, angle, skin, tissue, system setting, etc.).**

Keywords: **ePGG, glottal area, vocal folds auto-oscillation, non-invasive measurement, mechanical replica**

I. INTRODUCTION

Observation and measurement of the glottal opening and vocal folds vibration during speech requires the measurement of small displacements (of order of a millimeter) at a high sampling rate. Further, the technique must be as little invasive as possible in order not to prevent from normal speech and/or articulation.

External lighting and sensing ElectroPhotoGlottoGraph (ePGG), developed at LPP [4], relies on transillumination technique (Fig. 1), which consists of illuminating the glottis through the neck skin with infrared light and recording the variation of light intensity modulated by the vibration of the vocal folds with a sample rate of 20 kHz. In contrast to common laryngeal illumination techniques no visible light is used to light the glottal area. Instead light in the near infrared (IR) spectral range is used since wavelengths in this range (700-1000 nm) are reported to transilluminate large sections (many centimeters) of human tissue [1-3]. Compared with other photoglottographic techniques, this device has the advantage that both the light source (IR) and the light sensor (S) are positioned on the speaker's neck as illustrated in Fig. 1. Non-invasive, it allows to perform measurements without coercion for the speaker and can be used without medical assistance.

The objective of this paper is therefore to assess ePGG measurements during auto-oscillation of a mechanical glottal replica for which the varying glottal area can be accurately quantified during oscillation. The influence of all parameters potentially affecting the outcome of an ePGG measurement is studied in a controlled and repeatable way in order to propose a calibration procedure or/and provide objective guidelines for ePGG usage.

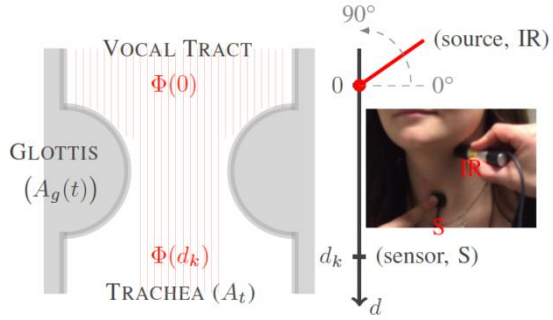


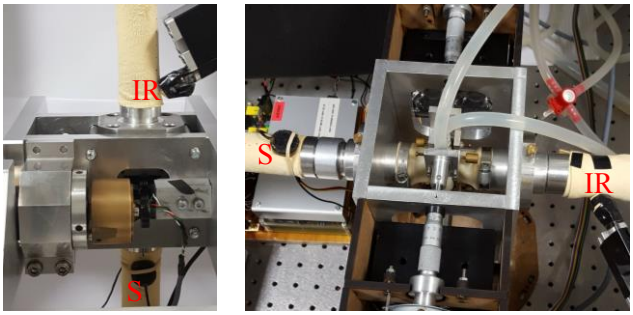
Fig. 1: Glottal transillumination principle.

II. METHODS

In order to realize this objective, experiments are performed using the following experimental setups.

Firstly, to characterize the ePGG outcome, the system (source and sensor) is mounted on an optic bench, so that the emitter and receiver system can be characterized (positioning, settings) and environmental measurement disturbances can be identified. Next, a uniform Plexiglas tube (diameter 25 mm) is used to represent the human trachea and pharynx. The tube is covered with layers of lamb leather to simulate the absorption of light by the human skin. The impact of leather thickness and IR source positioning (angle, distance to sensor) are assessed.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)



a) **b)**

Fig. 2: ePGG system source/emitter IR and optic sensor/receiver (S) placed at a) motor-driven rigid replica, b) flow-driven auto-oscillating replica.

Secondly, the ePGG system is placed on a mechanical glottal replica [6] to which two tubes are added representing the trachea and the pharynx as illustrated in Fig. 2.a. The "vocal folds" consist of two rigid half-cylinders, one of which is forced into motion by an eccentric motor. A second glottal replica is used (Fig. 2.b), the "vocal folds" consist then of latex tubes filled with water so that the replica is able to self-oscillate in interaction with an airflow [10,11]. In both cases, the initial glottal area prior to oscillation can be imposed, so that steady geometrical configurations can be systematically quantified. Usage of both replicas ensures that slow (rigid motor-driven, frequency range < 15 Hz, aperture < 2 mm) as well as fast (auto-oscillation, frequency range ~100-300 Hz, aperture < 2 mm) vocal folds displacements can be studied accurately whereas the glottal area can be measured either using an optical sensor (type OPB700) for the first replica or using a high-speed camera (Motion BLITZ Eosens Cube 7) at 525 frames per second and standard image processing techniques validated on the auto-oscillating replica [4]. These replicas allow to study parameters in the range relevant to orders of magnitudes observed on human speakers [5-9].

III. EPGG SIGNAL MODEL

The ePGG system is assessed on mechanical replicas. Since experimental setups are equipped to measure the glottal area, the relationship between ePGG signal and glottal area can be studied on these replicas as a function of parameters potentially affecting the ePGG signal (Fig. 1). In the following, the experimental ePGG signal characterization is presented firstly for static geometrical configurations with constant glottal area and secondly for dynamic geometrical configurations with a time-varying glottal area. During all experiments, the room temperature was 21 ± 1 °C.

A. Static glottal area

Firstly, the effect of source-sensor distance d (Fig. 1) on the ePGG signal is sought. The ePGG system is positioned on the mechanical airway replica with constant area (A_u = 491 mm^2). The source-sensor distance d, is systematically varied in the range 2 mm ≤ d ≤ 200 mm and the orientation angle is 27°.

In addition, in order to mimic the influence of wall tissue thickness, measurements are performed adding two (thickness 1.4 mm) or three leather layers (thickness 2.1 mm). Measured mean ePGG signals are plotted in Fig. 3. The ePGG signal decreases with d regardless of wall thickness. Linear fitting of measured ePGG signals in the range d ≤ 100 mm (appropriate for human subjects) and in the range d ≥ 100 mm (appropriate for mechanical replicas), results in R^2 ≥ 98.9%. Consequently, a first order linear approximation can be used to characterize the evolution of ePGG signal with source-sensor distance d, while the negative slope depends on wall absorption (thickness) and distance d. All remaining experiments are done with 2 layers (thickness 1.4 mm).

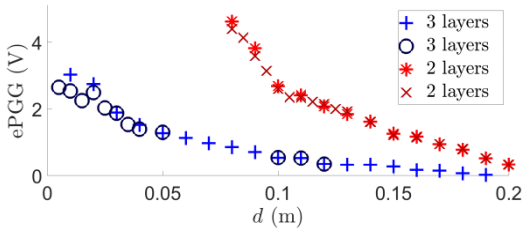


Fig. 3: Mean ePGG signal as a function of source-sensor distance d for the airway replica with 2 and 3 leather layers.

Secondly, static geometrical configurations are done to determine the effect of the source orientation angle in the mid-coronal plane (Fig. 1) on the ePGG signal. The ePGG system is again positioned on the uniform mechanical airway replica, i.e. in absence of a glottal constriction (no glottal replica). The source orientation angle is systematically varied from 0° up to 40° and the source-sensor distance is held constant to d = 100 mm. Measured mean ePGG signals are plotted in Fig. 4. For orientation angles up to about 15°, the ePGG signal is minimum and only marginally (< 0.3 V) affected by the orientation angle due to the source (IR) half beam angle of 22.5 ± 2.5°. Further increasing the orientation angle above 15° results in a linear (R^2 = 98.1%) increase of the mean ePGG signal. All remaining experiments are done for orientation angle 27°.

Thirdly, static geometrical configurations are performed to determine the effect of glottal area on the ePGG signal (Fig. 1). The rigid and deformable vocal

folds mechanical replicas are used to which a uniform mechanical airway wall replica is attached at each end.

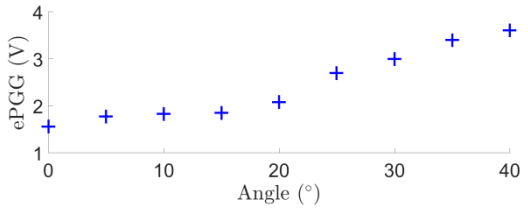


Fig. 4: Mean ePGG signal as a function of source orientation.

Source and sensor are positioned on each airway replica (trachea end and vocal tract end) so that the glottal area of the mechanical replica corresponds to the minimum area of the channel portion between source and sensor. The (minimum) source sensor distance is d = 150 mm for the rigid and d = 257 mm for the deformable glottal replica. Glottal area A_g is varied in the range 0-55 mm^2 (rigid) and 20-100 mm^2 (deformable). Measured mean ePGG signals are plotted in Fig. 5. The ePGG signal increases linearly with A_g for both the rigid (R^2 = 99.2%) and the deformable (R^2 = 98.2%) replica. So that ePGG signal and glottal area relate well using a linear approximation. Note that, in general, differences in slope and offset can occur due to 1) positioning of the ePGG system (source-sensor distance d, orientation angle) and 2) channel wall properties affecting absorption (thickness, material, etc.). In the next section, time-varying glottal areas are considered.

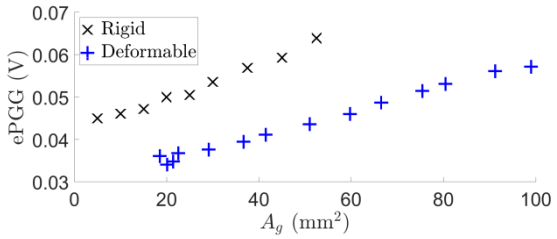
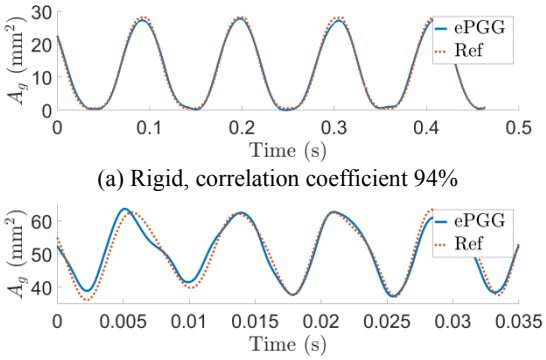


Fig. 5: Mean ePGG signal as a function of static glottal area A_g for a rigid and a deformable mechanical glottal replica.

B. Time-varying glottal area

The correlation between the time-varying ePGG signal and the time-varying glottal area is quantified for the motor driven rigid (oscillation frequency $f_0 \in \{2, 5, 10, 12\}$ Hz, $0 \leq Ag \leq 40$ mm^2, d = 150 mm) and the flow-driven deformable (fundamental frequency $f_0 \in \{113; 125; 129; 131\}$ Hz for mean upstream glottal pressures $Pu \in \{500, 570, 720, 840\}$ Pa, $20 \leq Ag \leq 100$ mm^2, d = 257 mm) mechanical replica. Typical examples of correlated time signals for slow (rigid) and fast (deformable) vocal folds displacement are plotted in Fig. 6. Correlation coefficients between ePGG signals and glottal area $A_g(t)$ yield > 90% for the rigid and > 85% for the deformable glottal folds replica.



(a) Rigid, correlation coefficient 94%

(b) Deformable, correlation coefficient 90%

Fig. 6: Illustration of correlated time signals of scaled ePGG (full line, ePGG) and glottal area Ag (dashed line, Ref): a) rigid replica at f_0 = 10 Hz, b) deformable replica at f_0 = 129 Hz.

Consequently, the ePGG signal and glottal area are correlated at all times during the oscillation. In the following section, it is aimed to model the relationship between ePGG signal and glottal area $A_g(t)$ accounting for the different variables affecting the ePGG signal.

IV. EPGG SIGNAL MODELING

In Section III, it was shown that the ePGG signal is mainly determined by 1) the source-sensor distance, 2) the minimum area of the channel portion between the source and the sensor and 3) the measurement condition determined by the combination of wall properties (e.g. absorption), environment (*e.g.* light) and ePGG system settings and positioning (*e.g.* orientation angle). In the following, an ePGG signal model is proposed accounting for each of these factors. Next, its parameters estimation and initialization is outlined. Finally, the sought relationship between ePGG signal and glottal area $A_g(t)$ is discussed.

Following the transillumination principle shown in Fig. 1, ePGG sensor voltage U is proportional to light intensity I at distance d_k from the light source,

$$U(d_k) \propto I(d_k) \tag{1}$$

Transmitted light intensity $I(d_k)$ at sensor position d_k is then expressed using light flux ϕ as

$$I(d_k) = \iint_{A_{min}} \phi(d_k) dA, \tag{2}$$

where

$$A_{min}(d_k) = \min_{d \in [0,d_k]}(A(d)) \quad (3)$$

is the minimum area encountered by the transmitted light flux between the source position and the sensor position. Furthermore, in Section III was shown that the dependence on d and A_{min} can be described using a first order linear approximation. Consequently, light flux $\phi(d) > 0$ can be approximated by model $\phi_m(d)$ defined as

$$\phi_m(d) = \alpha_d d + \beta_d, \quad (4)$$

with slope $\alpha_d < 0$ and offset $\beta_d > 0$ (see Section IV). From (2), $I(d_k)$ is now modeled as

$$\begin{aligned} I_m(d_k) &= A_{min}(d_k) \cdot \phi(d_k), \\ &= A_{min}(d_k) \cdot (\alpha_d d_k + \beta_d). \end{aligned} \quad (5)$$

Inserting (5) in (1) results in modeling the ePGG voltage $U(d_k)$ as $U_m(d_k)$ given by

$$\begin{aligned} U_m(d_k) &= \gamma\big((\alpha_d d_k + \beta_d) \cdot A_{min}(d_k)\big) + \eta \\ &= (\alpha_v d_k + \beta_d) \cdot A_{min}(d_k) + \eta. \end{aligned} \quad (6)$$

where $\eta > 0$ is the remaining signal measured for $A_{min}(d_k) = 0$ and $\gamma > 0$ is the scaling factor of (1). For sake of simplicity, let us denote $\alpha_v = \gamma\alpha_d < 0$ and $\beta_v = \gamma\beta_d > 0$. It is worth noting that $A_{min} = 0$ corresponds to glottal closure so that no direct light is transmitted. Therefore, η is independent of d and A_{min} so that η reflects solely the measurement condition. Considering now a time-variation of the glottal opening, model (6) can be directly extended as

$$U_m(d_k, t) = (\alpha_v d_k + \beta_v) \cdot A_{min}(d_k, t) + \eta. \quad (7)$$

Consequently using model (7), extracting area $A_{min}(d_k,t)$ from measured ePGG signals $U(d_k,,t)$ reduces to a parameter estimation problem which is solved using the iterative gradient descent method with appropriate initialization [12].

V. CONCLUSION

An experimental systematic characterization of the ePGG system has been assessed. Main parameters affecting the ePGG outcome are identified. Based on this result, glottal area modulation is studied experimentally on a motor-driven and on a flow driven mechanical larynx replica. It is seen that the correlation between ePGG signal and area measurement yields more than 85%. Next, a calibration method for the ePGG system is proposed. Further research focuses on a systematic validation of the proposed method.

ACKNOWLEDGMENT

Partly funded by ArtSpeech (ANR-15-CE23-0024).

REFERENCES

[1] D. Delpy and M. Cope, “Quantification in tissue near-infrared spectroscopy,” *Phil. Trans. R. Soc. Lond. B*, vol. 352, pp. 649–659, 1997.

[2] T. Lister, P. Wright, and P. Chappell, “Optical properties of human skin,” *J. Biomed. Optics*, vol. 17, pp. 1–15, 2012.

[3] S. Jacques, “Optical properties of biological tissues: a review,” *Phys. Med. Biol.*, vol. 58, pp. R37–R61, 2013.

[4] H. Kim, K. Honda, and S. Maeda, “ePGG, airflow and acoustic data on glottal opening in Korean plosives,” *Proc. 18th Int. Conf. of Phonetics Sciences (ICPhS)*, Glasgow, UK, 2015, p. 4.

[5] D. O’Shaughnessy, Speech Communication Human and Machine. *Addison-Wesley Publishing Company*, 1987.

[6] J. Cisonni, A. Van Hirtum, X. Pelorson, and J. Willems, “Theoretical simulation and experimental validation of inverse quasi one-dimensional steady and unsteady glottal flow models,” *J. Acous. Soc. Am.*, vol. 124, pp. 535–545, 2008.

[7] J. Burnard West, Bio-engineering aspects of the lung, Volume 3, *M. Dekker, Ed. Marcel Dekker, INC.*, 1977.

[8] T. Brancatisano, P. Collett, and L. Engel, “Respiratory movements of the vocal cords,” *J. Appl. Physiol.*, vol. 54, pp. 1269–1276, 1983.

[9] A. Scheinherr, L. Bailly, O. Boiron, A. Lagier, T. Legou, M. Pichelin, G. Caillebotte, and A. Giovanni, “Realistic glottal motion and airflow rate during human breathing,” *Med. Eng. Physics*, vol. 37, pp. 829–839, 2015.

[10] A. Van Hirtum and X. Pelorson, “High-speed imaging to study an auto-oscillating vocal fold replica for different initial conditions,” *Int. J. Applied Mech.*, Accepted.

[11] J. Haas, P. Luizard, X. Pelorson, and J. Willems, “Study of the effect of a moderate asymmetry on a replica of the vocal folds,” *Acta Acustica*, vol. 102, pp. 230–239, 2016.

[12] J. Nocedal and J. Wright, Numerical Optimization. Springer, 2006.

MODELLING OF RANDOM EXTRA PULSES DURING QUASI-CLOSED GLOTTAL CYCLE PHASES

P. Aichinger[1], I. Roesner[1], J.Schoentgen[2], F. Pernkopf[3]

[1] Department of Otorhinolaryngology, Division of Phoniatrics-Logopedics, Medical University of Vienna, Austria
[2] F.N.R.S. & Université Libre de Bruxelles, Laboratories of Image, Signal Processing and Acoustics, Faculty of Applied Sciences, Brussels, Belgium
[3] Signal Processing and Speech Communication Lab, Graz University of Technology, Austria
philipp.aichinger@meduniwien.ac.at, imme.roesner@meduniwien.ac.at, jschoent@ulb.ac.be, pernkopf@tugraz.at

Abstract: **The presence of random extra pulses during quasi-closed glottal cycle phases may constitute a distinct voice quality type relevant to the clinical care of disordered voices. In this paper, we propose for this voice type a glottal area waveform model that includes automatic parameter estimation. The model involves (1) extraction of the fundamental frequency, (2) estimation of the cyclic pulse times, heights and shapes, (3) Fourier synthesis of a cyclic pulse train model, (4) closed phases estimation via fitting an inverted parabola to the averaged pulse shape, (5) estimation of the random extra pulses' positions and shapes, and (6) pulse shape filtering based synthesis of the random extra pulses. For a typical voice sample, the root mean square error energy level of the purely cyclic model E_1 = -13.2 dB, which improves by 1.5 dB when extra pulses are added to the model.**

Keywords: **Glottal area waveform, voice quality, random extra pulses, waveform model, detection**

I. INTRODUCTION

The assessment of voice quality is pivotal to the clinical care of disordered voice. Voice quality is a functional condition or outcome that guides the indication, selection, evaluation, and optimization of treatment. On the level of auditory perception, breathiness and roughness are routinely assessed. Breathiness is the "auditory impression of turbulent air leakage through an insufficient glottic closure", and roughness is the "audible impression of irregular glottic pulses, abnormal fluctuations in Fo, and separately perceived acoustic impulses (as in vocal fry), including diplophonia and register breaks." [1].

Irregular voices may contain extra pulses that are added to the cyclic pulse trains typical for normal phonation. The occurrence of extra pulses is related to glottal mechanics and aerodynamics, which are subject to clinical treatment, e.g., phonosurgery, or voice therapy. Despite their potential relevance to clinical treatment of voice disorders, extra pulses are not explicitly described in clinical practice most of the time, partly because their detection is tedious and difficult.

In this paper, we propose a numerical model for glottal area waveforms (GAW) including random extra pulses. Applications of the model are the synthesis of phonatory signals, and the model-based detection of extra pulses in phonatory signals. Synthesis is needed for auditory experimentation and listener training, as well as for creating test signals for voice analyzers. Detection is needed for the automatic distinction of extra pulses from other types of irregularity in clinical practice.

The proposed GAW model involves: (1) extraction of the fundamental frequency, (2) estimation of the cyclic pulse times, heights and shapes, (3) Fourier synthesis of a cyclic pulse train model, (4) closed phases estimation via fitting an inverted parabola to the averaged pulse shape, (5) estimation of the random extra pulses' positions and shapes, and (6) pulse shape filtering based synthesis of the random extra pulses.

II. METHODS

A voice sample that is simultaneously tonal and raspy has been selected from a database of laryngeal high-speed videos and audio recordings of pathological and non-pathological voices [2]. We define "tonal raspiness" as the auditory percept of a vocal pitch constituting one distinct auditory stream mixed with a second distinct auditory stream evocative of a rasping or grating noise. The voice sample is 125 ms long, i.e., 500 video frames at a frame rate $f_r' = 4\ kHz$. The GAW is obtained with the seeded region growing algorithm [3], [4].

Fig. 1 shows the block diagram of the proposed GAW model for random extra pulses during quasi-closed glottal cycle phases. The fundamental frequency of a cyclic quasi-unit pulse train u_1 is f_o. The pulse train u_1 is jittered and shimmered. It is pulse shape filtered with shape r_1 to obtain the cyclic pulse train d_1, which contains pulses with shape r_1 centered at positions of the pulses in u_1. The pulse shape filter is realized as a Fourier synthesizer. Random extra pulses d_2 are modelled by multiplying a phase shifted version

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

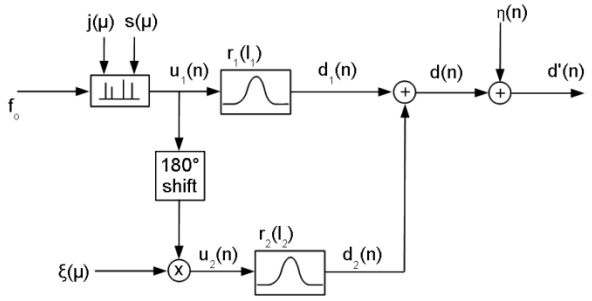


Figure 1: Block diagram of the proposed GAW model containing random extra pulses during quasi-closed glottal cycle phases.

of u_1 by random numbers $\xi \in \{0,1\}$, and subsequent pulse shape filtering of u_2 with shape r_2. The cyclic pulse train d_1 and the randomly generated extra pulses d_2 are added to obtain the noiseless GAW d. Furthermore, white Gaussian noise η is added, which results in the noisy GAW d'.

In more detail, the cyclic pulse train $u_1(n) = \sum_\mu s(\mu) \cdot \delta[n - \mu \cdot N_1 - j(\mu)]$, where n is the discrete time index, $\mu \in \mathbb{Z}$ is the pulse index, $\delta(x)$ is the unit pulse function, i.e., $\delta(x) = 1$ at $x = 1$, and 0 elsewhere, the cycle length in samples $N_1 = f_r / f_o$, $s(\mu)$ is the height of the μ^{th} pulse, which represents shimmer, and $j(\mu)$ is the time shift of the μ^{th} pulse, which represents jitter. In particular, s and j are random numbers drawn from uniform distributions in the intervals $0.5 \le s \le 2$, and $-\frac{N_1}{4} \le j \le \frac{N_1}{4}$. Thus, u_1 decodes the positions and heights of the pulses of d_1. The cyclic pulse train

$$d_1(n) = \mathrm{A}(n) \cdot d'_1(n), \text{ with} \quad (1)$$

$$d'_1(n) = a_0 + \sum_{p=1}^{P} \big[a_p \cdot \cos\big(p \cdot \Theta(n)\big) + b_p \cdot \sin\big(p \cdot \Theta(n)\big)\big]. \quad (2)$$

The Fourier coefficients a_p and b_p are the real and imaginary parts of the discrete Fourier transform of $r_1(l_1)$, with $l_1 \to p$, where l_1 is the pulse shape index that goes from $-\frac{N_1}{2} + 1$ to $\frac{N_1}{2} - 1$, and $p \in \{1, \dots, P\}$ is the partial index. The instantaneous phase $\Theta(n) = \pi \cdot \sum_{\mu \in \mathbb{Z}} [2 \cdot \mu + 1]$ at pulse locations of $u_1(n)$, i.e., at $n = \mu \cdot N_1 - j(\mu)$, and spline interpolated in between. The amplitude modulation function $\mathrm{A}(n) = u_1(n)$ at pulse locations of u_1, and obtained by shape-preserving cubic interpolation in between.

The random pulse train $u_2(n) = \sum_\mu \xi(\mu) \cdot \delta\left[n - N_1 \cdot \mu - \frac{N_1}{2}\right]$ decodes the positions of the pulses of d_2. The random numbers $\xi(\mu)$ are drawn from a Bernoulli distribution, which models the random occurrence of extra pulses. The term $\frac{N_1}{2}$ shifts u_2 by 180°, such that the random pulses are centered temporally between the adjacent cyclic pulses. The random pulse train d_2 is obtained by pulse shape filtering, i.e., convolution, as $d_2(n) = \sum_{l_2} u_2(n) \cdot r_2(n - l_2)$.

Fig. 2 illustrates the parameter estimation procedure. The parameters are estimated for signal blocks of 32 ms with 50 % overlap, using a Hann window. First, $\hat{f}_o$ is estimated from the GAW d' as described elsewhere [5], with the maximal number of f_os set to 1. All signals are processed at a resampled frame rate $f_r = 50\,kHz$. Second, the pulse shape estimate $\hat{r}_1$ is obtained by normalized cross-correlation, i.e., $\hat{r}_1(l_1) = \frac{1}{\sum_n \hat{u}_1(n)} \cdot \sum_n \hat{u}_1(n) \cdot d'(n - l_1)$, where $\hat{u}_1$ is a unit pulse train driven by $\hat{f}_o$, and d' is the observed GAW. The cyclic pulse train $\hat{d}_1$ is estimated via Fourier synthesis (1), (2), with $P = 10$. The Fourier coefficients are spline interpolated with respect to time, which enables smooth slow pulse shape modulation. Third, the model error E_1 is obtained as the root mean square (RMS) error level of the error waveform $e_1(n) = d'(n) - \hat{d}_1(n)$ given in dB, with reference to the observed GAW d', i.e.,

$$E_1 = 20 \cdot \log_{10}\left[\sqrt{\overline{e_1(n)^2}} \Big/ \sqrt{\overline{d'(n)^2}}\right]. \quad (3)$$

Fourth, the modulation noise is estimated as $[\hat{\jmath}(\mu),\ \hat{s}(\mu)] = \mathrm{argmin}_{j(\mu),s(\mu)}\{E_1(j(\mu), s(\mu))\}$. The interior-point algorithm is executed for each pulse one by one. It optimizes parameters $j(\mu)$ and $s(\mu)$, in order to minimize objective E_1 [6], [7]. Each variation of $j(\mu)$ and $s(\mu)$ requires an update of the pulse train $\hat{d}_1$, which includes updates of the pulse train $\hat{u}_1$, the pulse shape $\hat{r}_1$, its Fourier coefficients $\hat{a}_p$ and $\hat{b}_p$, the instantaneous phase $\widehat{\Theta}$, and the amplitude modulation function $\widehat{\mathrm{A}}$. After the last pulse of the sequence has been optimized, the procedure is started again from the first pulse on, until convergence, i.e., until the model error improvement cumulated from the first to the last pulse decreases below 0.01 dB.

Fifth, the random pulse train estimate $\hat{d}_2$ is obtained. The steepness of a clipped inverted parabola is estimated, i.e., $q_{opt} = argmin_q \left\{\frac{1}{N_1} \cdot \sum_{l_1} [max(0, 1 - q \cdot |l_1|^2) - \hat{r}_1'(l_1)]\right\}$, via golden section search and parabolic interpolation. The pulse shape $\hat{r}_1'$ is normalized to the range [0,1], and the parabola is fit to $\hat{r}_1'$ optimally in a least squares sense. The muting cycle $m'(l_1) = 1$, where $q_{opt} \cdot |l_1|^2 < 1$, and 0 elsewhere. The cyclic muting function is obtained by convolution, i.e., $m(n) = \sum_{l_1} \hat{u}_1(n) \cdot m'(n - l_1)$. The positions of

the extra pulses are obtained by picking peaks in the cyclically muted model error $e_1'(n) = e_1(n) \cdot m(n)$. The unit pulse train estimate $\hat{u}_2$ is generated with pulses at the peak positions. The random extra pulses' shape estimate $\hat{r}_2$ is obtained by normalized cross-correlation, i.e., $\hat{r}_2(l_2) = \frac{1}{\sum_n \hat{u}_2(n)} \cdot \sum_n \hat{u}_2(n) \cdot e_1(n - l_2)$, where l_2 is the pulse shape index that goes from $-\frac{N_1}{2} + 1$ to $\frac{N_1}{2} - 1$. The random pulse train estimate $\hat{d}_2$ is obtained by convolution, i.e., $\hat{d}_2(n) = \sum_{l_2} \hat{u}_2(n) \cdot \hat{r}_2(n - l_2)$.

Finally, the model error E is obtained as the RMS error level of the error waveform $e(n) = e_1(n) - \hat{d}_2(n)$ in dB, with reference to the observed GAW d', i.e.,

$$E = 20 \cdot \log\left[\sqrt{\overline{e(n)^2}} \Big/ \sqrt{\overline{d'(n)^2}}\right]. \quad (4)$$

III. RESULTS

Fig. 3 shows several signals involved in the modelling of random extra pulses during quasi-closed glottal cycle phases. In (a), the observed GAW d' is shown. Twenty-one cyclic pulses are marked by arrows and six random extra pulses are marked by double arrows. Two of the extra pulses interfere with adjacent cyclic pulses at approximately 1.405 and 1.415 s. Inspection of the video confirms the existence of these extra pulses. The Bernoulli parameter is approximated as $p(\xi = 1) = \frac{6}{21} \approx \frac{1}{3}$. Fig. 3 (b) shows the amplitude modulation function A, the model of the cyclic pulses $\hat{d}_1$, and the quasi-unit pulse train $\hat{u}_1$. A and $\hat{u}_1$ are rescaled for comparability. The cyclic pulses of $\hat{d}_1$ agree with the cyclic pulses of d' in terms of timing, height and shape. Some minor pulses exist in $\hat{d}_1$, which are bias artefacts in the estimation of the cyclic pulse shapes $\hat{r}_1$, evoked by the random extra pulses. The amplitude modulation function A seems to be favorable smooth. The instantaneous frequency is the instantaneous phase Θ derived with respect to time and shown in Fig. 3 (c). It varies slightly and smooth in the vicinity of 0.02-0.025 rad/sample, which corresponds to approximately 160 – 200 Hz. Fig. 3 (d) shows the error waveform e_1, which contains the extra pulses. Also other pulses exist in e_1, which are errors due to cycle-to-cycle shape modulations in d' that are not considered in the model. These additional pulses in e_1 require the muting during quasi-closed phases prior to peak picking.

The root mean square (RMS) energy level of the error E_1 is 13.2 dB, which reflects the agreement of $\hat{d}_1$ with d'. The model of the extra pulses $\hat{d}_2$ and the unit pulse train $\hat{u}_2$ are shown in Fig. 3 (e). $\hat{u}_2$ is rescaled for comparability. All extra pulses are correctly detected, and only one false alarm occurs (arrow). Fig. 3 (f) shows the error waveform e. E is 14.7 dB, which is an improvement of 1.5 dB as compared to the model $\hat{d}_1$.

IV. DISCUSSION

A GAW model including random extra glottal pulses during the quasi-closed phases is proposed. The model is tested on a typical voice sample including automatic parameter estimation. Two versions of the waveform model are obtained and compared with the observed GAW. The first contains cyclic pulses only, whereas the second contains extra pulses. Considering extra pulses in the model improves the model error by 1.5 dB in the tested sample.

Limitations are listed with reference to additional suggestions for future work. First, this is a prototypical proposal, i.e., only one voice sample was used for testing. Thus, the sample size needs to be increased. However, recruiting patients with extra pulses as based on auditory screening is difficult as one must first learn how extra pulses sound. Fortunately, computational screening may be used to find extra pulses in existing data [2], and synthesized data can also be used. Second, our proposed peak picking involves two adjustable parameters, i.e., the minimum peak prominence, and the minimum peak height, which is a weakness. Other solutions for detecting randomly occurring repetitive events may involve less degrees of freedom [8], [9]. Third, it was assumed that extra pulses are unjittered, which may not be true. Thus, this assumption was relaxed by picking peaks to estimate the times of extra pulses. However, jittering of extra pulses may be added to the model explicitly in the future, to enable more sophisticated detection. Fourth, cycle-to-cycle shape modulations are not considered in the model, which causes additional pulses in the error waveform e_1. Consideration of cycle-to-cycle shape modulations may help in the future to get rid of these additional pulses that may disturb the detection of the extra pulses. Finally, our estimation of the cyclic pulse shapes is biased such that residuals of extra pulses occur in the model of the cyclic pulses. This artefact may increase with frequency of occurrence of extra pulses. A strategy to suppress the small pulses in double pulsed pulse shapes may thus help to improve the parameter estimation and signal segregation.

V. CONCLUSION

The presence of random extra pulses during quasi-closed glottal cycle phases may be a clinically relevant vocal phenomenon. It may constitute a type of dysphonic voice quality that is distinct on the levels of

glottal kinematics, radiated sound, and auditory perception. It may relate to mechanical and aerodynamical parameters of phonation that may be controllable in clinical treatment. Moreover, it may relate to higher order vocal features like, e.g., vocal fatigue and endurance, vocal timbre, and the identity of a speaker. However, not much is known about this voice type.

The proposed model establishes an analogy of the signal domain and the perceptual domain. In particular, human listeners segregate auditory streams that may also be segregated and analyzed computationally. Cyclic pulses provoke the tonal auditory stream, and random extra pulses provoke the raspy stream.

Detailed insights into a promising option for the modelling of the encountered voice type are presented. The functionality of the model and the automatic parameter estimation framework is demonstrated on a typical voice sample. Also, the estimation of modulation noise from the glottal area waveform (GAW) is brought into focus. The most urgent next step will be to reproduce our preliminary success with new natural and synthetic data, in order to establish normative ranges of the frequency of occurrence of random extra pulses during quasi-closed glottal cycle phases.

VI. ACKNOWLEDGEMENTS

The authors would like Richard Wolf GmbH for providing the HRES ENDOCAM 5562 and the Division of Phoniatrics and Pediatric Audiology, University Hospital Erlangen, Germany, for providing the Glottis Analysis Tools (GAT).

REFERENCES

[1] P. Dejonckere, P. Bradley, P. Clemente, G. Cornut, L. Crevier-Buchman, G. Friedrich, P. Van De Heyning, M. Remacle, and V. Woisard, “A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques,” *Eur. Arch. Otorhinolaryngol.*, vol. 258, no. 2, pp. 77–82, 2001.

[2] P. Aichinger, I. Roesner, M. Leonhard, D. Denk-Linnert, W. Bigenzahn, and B. Schneider-Stickler, “A database of laryngeal high-speed videos with simultaneous high-quality audio recordings of pathological and non-pathological voices,” in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2016, vol. 10, pp. 767–770.

[3] R. Adams and L. Bischof, “Seeded region growing,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 6, pp. 641–647, Jun. 1994.

[4] J. Lohscheller, U. Eysholdt, H. Toy, and M. Dollinger, “Phonovibrography: Mapping High-Speed Movies of Vocal Fold Vibrations Into 2-D Diagrams for Visualizing and Analyzing the Underlying Laryngeal Dynamics,” *IEEE Trans. Med. Imaging*, vol. 27, no. 3, pp. 300–309, Mar. 2008.

[5] P. Aichinger, M. Hagmüller, I. Roesner, B. Schneider-Stickler, J. Schoentgen, and F. Pernkopf, “Fundamental frequency tracking in diplophonic voices,” *Biomed. Signal Proces.*, vol. 37, pp. 69–81, 2017.

[6] R. H. Byrd, J. C. Gilbert, and J. Nocedal, “A trust region method based on interior point techniques for nonlinear programming,” *Math. Program. Ser. B*, vol. 89, no. 1, pp. 149–185, 2000.

[7] R. A. Waltz, J. L. Morales, J. Nocedal, and D. Orban, “An interior algorithm for nonlinear optimization that combines line search and trust region steps,” *Math. Program.*, vol. 107, no. 3, pp. 391–408, 2006.

[8] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection theory*. Prentice-Hall, 1993.

[9] S. M. Kay, *Fundamentals of Statistical Signal Processing: Practical algorithm development*. Prentice Hall, 2013.

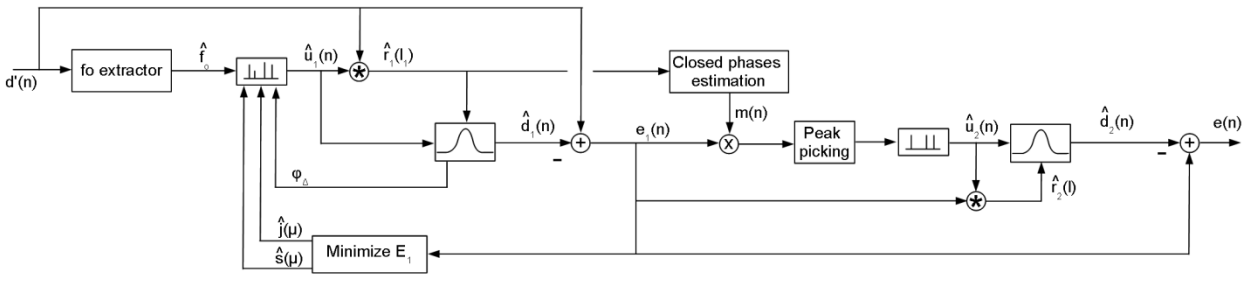


Figure 2: Block diagram of the proposed framework for parameter estimation.

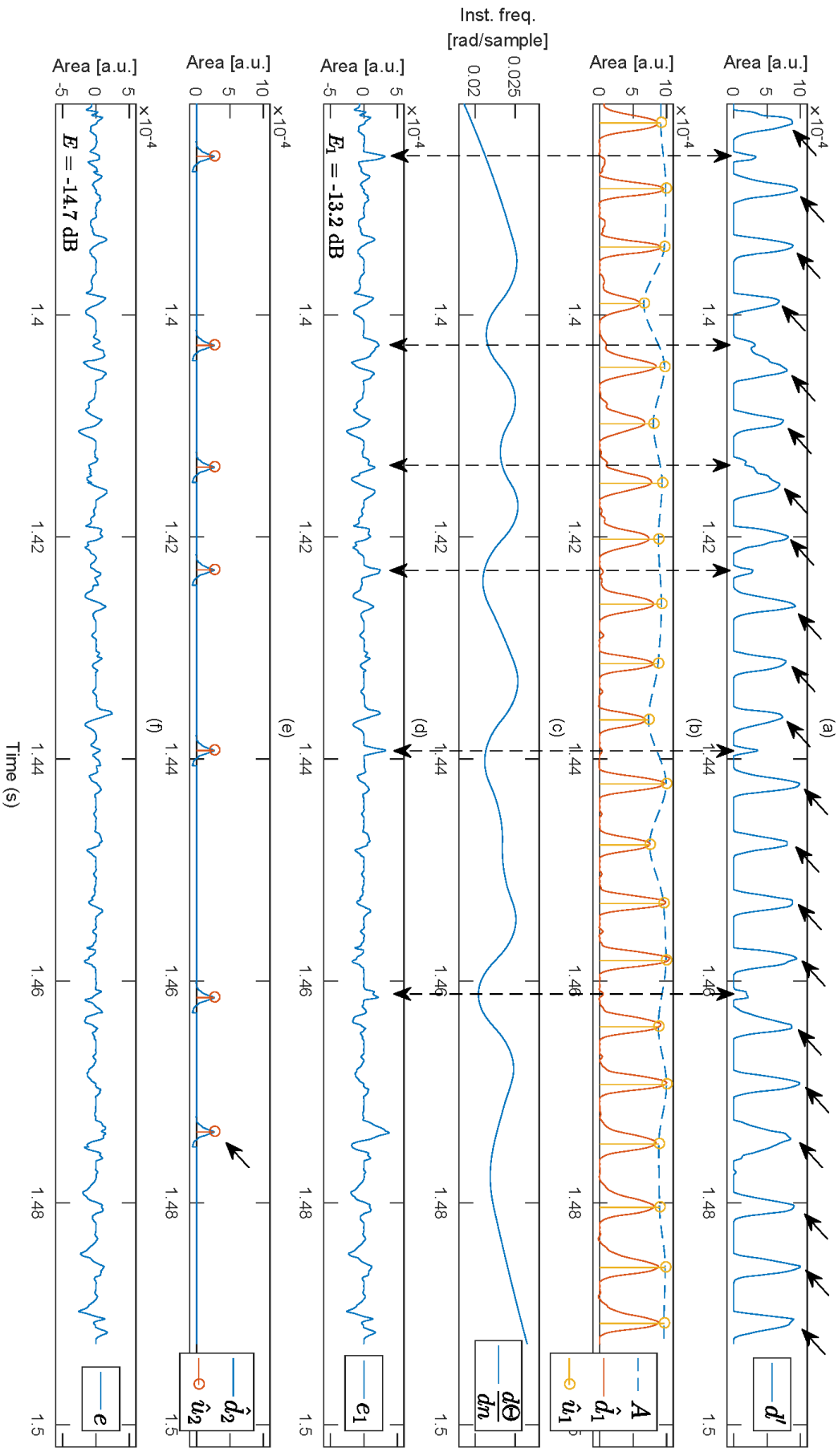


Figure 3: Signals involved in the modelling of random extra pulses during quasi-closed glottal cycle phases.

IDENTIFICATION OF THE GLOTTAL WAVE STRUCTURE WITH THE USE OF THE VOICE SOURCE SIGNAL RECORDING METHOD

A. E. Barabanov[1], K. V. Evgrafova[2], V.V. Evdokimova[2], P.A.Skrelin[2], T.V. Chukaeva[2]

[1] Saint Petersburg State University, Theoretical Cybernetics Department, Saint-Petersburg, Russia
[2] Saint Petersburg State University, Department of Phonetics, Saint-Petersburg, Russia
andrey.barabanov@gmail.com, karinaevgr@mail.ru, postmaster@phonetics.pu.ru, skrelin@phonetics.pu.ru, chukaeva68@mail.ru

Abstract: **The given research is aimed at analyzing the structure of a voice signal recorded with the microphone placed in the proximity of the vocal folds. In recording two microphones were used. The one was located in the larynx near the vocal folds and the other one was near the lips. The speech signal containing isolated Russian vowels was registered synchronously through both microphones. It was discovered that the signal in the inner microphone contains the first formant of the corresponding vowel and a very weak echo of other formants. The formants were suppressed with simple FIR filters for all the signals from the inner microphone. After that all the signals recorded by the inner microphone were transformed into the signals of a triangle form. The resulting triangular signals can be considered as a general model of the vocal fold output. The transfer functions of the vocal tract for the vowels were calculated using the signals recorded by the outer microphone. A cross validation of the sets of the transfer functions and the decorrelated triangular signals is provided by comparing the results of filtering with signals recorded by the outer microphone.**

Keywords: **voice generation, glottal wave structure, numerical methods, speech synthesis**

I. INTRODUCTION

The traditional approach to phonetic research of the vocal tract assumes that there are several successive stages of speech production which are initialization, phonation, articulation and radiance of speech signal. Almost all the internal organs are parts of the bio-mechanical oscillating system that generates the voice signal. This signal is individual and optimized by nature [1], [2], [3], [5], [9]. The periodic sequence of lung pressure differences in larynx is called the glottal wave [6], [7]. The frequency of these pulses corresponds to the fundamental frequency in speech signal. The fact of the interaction between the two parts of the vocal tract does not make the traditional linear source–filter theory completely consistent. Obtaining the vocal fold signal detached from the influence of the articulation system and analysing its nature is an important up-to-date problem for different fields of speech science and speech technology. There exist different voice source models that are applied to the majority of linguistic research and speech technology applications.

The LF-model (Lilencrants and Fant) of the voice source was one the first models of the vocal tract. It was developed in the 80-s by G. Fant [6], [7], [8]. It described the glottal wave as a sequence of pulses of the given shape. The voice source constituents were obtained from the signal using the inverse filtering. However, it is more complicated to use it for real time analysis of voice.

Apart from LF-model, there exist biomechanical models of the voice source and the vocal folds. Single-mass models could be called single degree-of-freedom vibration models for the vocal folds because each vocal fold is modeled with a single mass spring system [12], [14]. These models are of particular interest theoretically because they must explain the net work done on the vocal folds by air flow in 1 cycle in terms of asymmetries between opening and closing phases in air flow conditions. The two-mass vocal fold model introduced by Stevens [9] consists of two pairs of masses. Larger ones represent the inferior part of the vocal folds, and smaller ones represent the superior part of the vocal folds.

The source–filter interactions that involve changes in vocal fold vibration have been demonstrated by investigators [4], [11], [13], [14], [15]. However, the data presented are sometimes fragmentary and inconsistent. The main Titze's [11] goal was to determine the proportion of irregularities that are due to nonlinear source–tract interactions and to provide a theoretical framework for the bifurcation phenomena in vocal fold vibration with a nonlinear source–filter construct.

Our research is aimed at analyzing the signal of the voice source and the output speech signal to consider the non-linearity of the vocal tract system and the glottal source signal structure. A glottal wave carries

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)

important information about the vocal folds. The inverse filtering theory does not provide a reliable separation of the speech signal into a generating glottal wave and a transform in the vocal tract. In this paper we investigate the features of the signal recorded in the proximity of the vocal folds analysing the results of our recent recording experiments.

II. MATERIAL AND METHODS

A miniature microphone QueAudio (with d=2.3 mm, waterproof) was inserted through the nasal cavity and located in the proximity of vocal folds in the output air stream. An "inner" signal recorded by this microphone was expected to represent the output of the vocal folds. The second "outer" microphone was located near the lips. The two recordings (in the beginning and in the end of the vocal tract) are sufficient to estimate its transfer function. A speaker was asked to pronounce a vowel with pitch changing slowly within approximately an octave, increasing and then decreasing. The recordings were made for the vowels /a/, /o/, /æ/ /e/, /i/ by three male and two female speakers.

The main difference from the previous experiments [16], [17], [18], [19], [20] is the change of the fundamental frequency in a wide band. This gave us an opportunity to estimate precisely a spectral envelope of each vowel near the vocal folds. A full vocal tract transfer function is successfully estimated as well. Each recording is divided into frames of the length of a pitch period. A precise mathematical technique [21] was used to estimate the fundamental frequency, amplitudes and phases of all subharmonics in each frame.

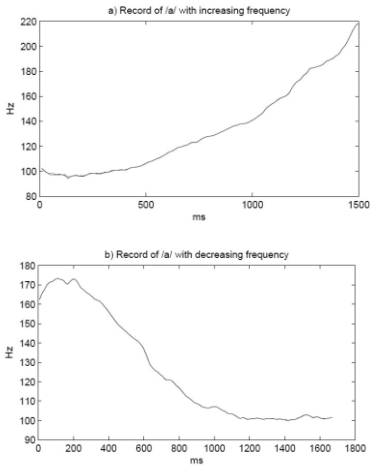


Fig 1. Estimated frequency curves in the two recordings of the vowel [a], with an increasing Fundamental frequency (a) and with a decreasing Fundamental frequency (b). Pitch periods were estimated for the inner signal and for the outer signal.

Frequencies of subharmonics are multiple to the fundamental frequency that changes in each frame. Therefore mapping “frequency to amplitude” forms a nearly solid spectral envelope of the inner signal for each vowel on a wide frequency band, see Fig.2, 3. Shades of gray mark the multiple number of a harmonic frequency with respect to the fundamental one.

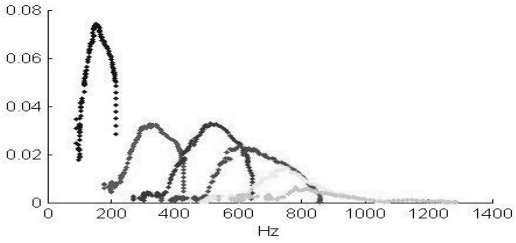


Fig.2. Spectral envelopes of the inner signal. The vowel [e], a male voice.

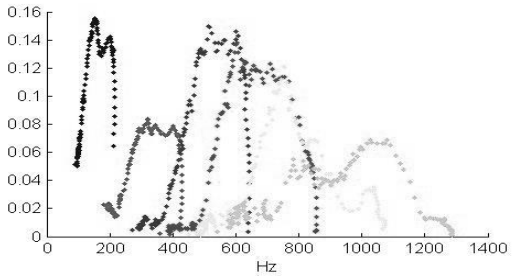


Fig.3. Spectral envelopes of the corresponding outer signal (right).They are nearly proportional in the low frequency band. The vowel [e], a male voice.

Thus, the inner signal contains the first formant and sometimes a very weak echo from other formants. This conclusion conforms to the physical conception of the direct dependence between the first formant and a width of the glottis opening.

The vocal tract transfer function can be estimated by the measured input and output. Its input is the output of the vocal chords. Its output is measured by the outer microphone located near the lips. The estimated transfer functions are shown in Fig. 4.

The sound with the vowel [e] contained also the vowel [ɛ]. Therefore the spectrum contains a mixture of the corresponding formants (1500 and 2200 Hz).

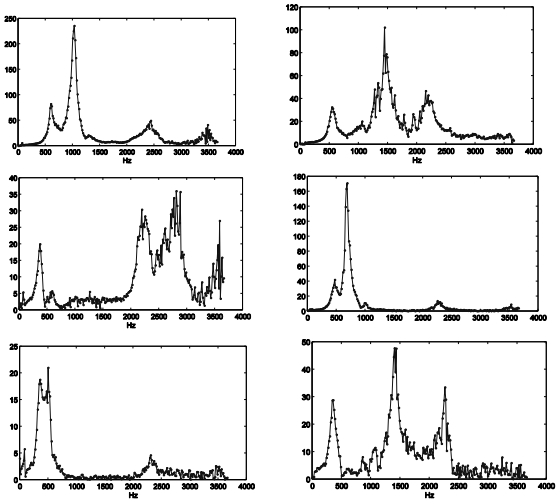


Fig 4. The transfer functions of the vocal tract for the vowels [a], [o], [e], [i], [u], [ɨ].

It is possible to separate these transfer functions. Or, in general, it is possible to interpolate the estimated transfer functions of the speaker under investigation to the transfer functions of other vowels. In the next Section a method of interpolation based on the Linear Spectral Pairs is presented.

It is possible to synthesize cross sounds. A vocal fold output can be represented by the triangle signals obtained from the one recording. The vocal tract can be represented by the impulse response obtained from the other recording. Filtering of the vocal fold output by the vocal tract filter model produces a new sound. As it was expected, the sound was completely determined by the vocal tract parameters.

III. RESULTS

1. The transfer functions of the vocal tract were calculated for each vowel and for each speaker by the superfast Schur algorithm of the Toeplitz matrix inversion [22]. All of the functions show the correct location of the corresponding formants. The transfer functions have also shown which spectral parts almost do not change in the vocal tract.

2. The inner signal is a low frequency, not exceeding 1100 Hz for each speaker and each vowel.

3. For each speaker a shape of the inner signal is similar for all vowels. A slight difference can be explained by the first or the second formant.

4. The shape of the low-frequency envelope of the output signal up to 1100 Hz is the same as that of the inner signal (Fig. 2, 3). It can be seen from spectra directly and from the absolute value of the transfer function of the vocal tract in the band [600, 1000] Hz which is nearly constant for each vowel. Thus, the low frequency spectrum part of the speech signal is formed mainly near the vocal folds.

5. The slopes in periods of the inner and outer signals are connected for all vowels and for all speakers. Each period of the inner signal contains a vertical fall that means a volume speed drop. The output speech signal contains the main increasing slope exactly in the same time interval, see Fig. 4. A beginning sample of a period is commonly chosen on this slope, in the maximum or at the zero intersection, for instance, in PSOLA. This correspondence is observed for all speakers and for all vowels.

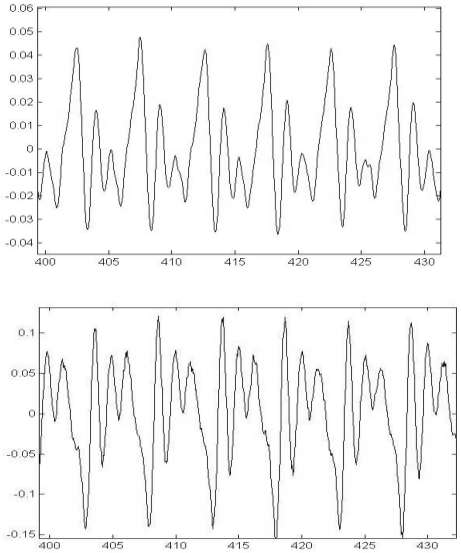


Fig.4. A drop in the inner signal (upper) is synchronized with the increasing slope in the outer signal (lower)

IV. CONCLUSION

For each speaker the recorded inner signals appeared to be similar for all vowels and for each fundamental frequency constituents. However, the signals are not the same due to the tracks of the low frequency formants. We can suppose that the common shape of the recorded inner signals of a speaker is close to that of his/her glottal wave. This shape essentially differs from that of the theoretical models that are either parametric or biomechanical. Another important result is an opportunity to extract the inner signal as a low frequency part of the speech signal with an appropriate phase correction.

V. Acknowledgements

The work was supported by Saint Petersburg State University projects 6.37.349.2015 , 31.37.353.2015, 31.40.94.2017

REFERENCES

[1] L. V. Bondarko, *Phonetics of Russian modern language*. Saint-Petersburg State University, 1998 (in Russian).

[2] S. V. Kodzasov, and O. F. Krivnova, *General Phonetics*. Moscow, 2001.

[3] G. / Fant, *Acoustic Theory of Speech Production*. Netherlands: Mouton, 1960.

[4] J. L. Flanagan, "Source-system interaction in the vocal tract,"*Ann. N.Y. Acad.Sci.* vol. 155, pp. 9-17, 1968.

[5] J. L. Flanagan, *Speech Analysis, Synthesis, and Perception*. New York: Springer, 1972.

[6] G. Fant., "The voice source in connected speech," *Speech Communication*, vol. 22, 1997.

[7] G. Fant, J. Liljencrants, and Q. Lin, *A four-parameter model of glottal flow*. STL-QPSR, Tech. Rep., 1985.

[8] G. A. Alzamendi, G. Schlotthauer, and M. E. Torres, "Formulation of a stochastic glottal source model inpired on deterministic Lilencrants-Fant model", *International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications MAVEBA 2015*, Florence, pp. 15-18, 2015.

[9] K. Stevens, *Acoustic Phonetics*. Cambridge, MA 02141: The MIT Press,1998.

[10] M. S. Howe, and R. S. McGowan, On the single-mass model of the vocal folds. Fluid Dyn. Res. 42, 015001.doi: 10.1088/0169-5983/42/1/015001.2010 (2010)

[11] I. R. Titze, "Non-linear source-filter coupling in phonation: Theory," *J. Acoust.Soc.Am.* vol. 123, 27332749.doi:10.1121/1.2832337. 2008 (2008)

[12] M. Zanartu, L. Mongeau, and G. R. Wodicka, "Influence of acoustic loading on an effective single mass model of the vocal folds," *J. Acoust. Soc. Am.*, vol. 121, pp. 1119-1129, doi:10.1121/1.2409491. 2007, 2007.

[13] H. Hatzikirou, W. T. S. Fitch, and H. Herzel, "Voice instabilities due to source-tract interactions," *Acta.Acust.Acust.*, vol. 92, pp. 468-475, 2006.

[14] D. G. Miller, and H. K. Schutte, "Mixing the registers: Glottal source or vocal tract," *Folia Phoniatr. Logop.*, vol. 57, pp. 278-291, 2005.

[15] P. Mergell, H. Herzel, "Modeling biphonation - The role of the vocal tract," *Speech Commun.*, vol. 22, pp. 141-154. doi: 10.1016/S0167-6393(97)00016-2. 1997, 1997.

[16] K. Evgrafova, V. Evdokimova, P. Skrelin, T. Chukaeva, and N. Shvalev, "A new technique to record a voice source signal," *International Workshop onModels and Analysis of Vocal Emissions for Biomedical Applications - MAVEBA 2013*, Florence, pp. 181-182, 2013.

[17] V. Evdokimova, K. Evgrafova, P. Skrelin, T. Chukaeva, and N. Shvalev, "Detection of the frequency characteristics of the articulation system with the use of voice source signal recording method," *Speech and Computer, Lecture Notes in Computer Science*, Springer, vol. 8113, pp. 108-115, 2013.

[18] A. Barabanov, V. Evdokimova, and P. Skrelin, "Estimation of Vowel Spectra Near Vocal Chords with Restoration of a Clipped Speech Signal," *Speech and Computer, Lecture Notes in Computer Science*, Springer, vol. 9319, pp. 209-216, 2015.

[19] V. Evdokimova, P. Skrelin, K. Evgrafova, T. Chukaeva, and N. Shvalev, "Investigating voice source signal filtering be articulation component," *Teoreticheskaya i prikladnaya lingvistika*, vol. 1, no. 3, pp. 37-49, 2015 (in Russian).

[20] V. Evdokimova, K. Evgrafova, and P. Skrelin, "Investigating source-filter interaction to specify classic speech production theory," in *Proceedings of the 18th International Congress of Phonetic Sciences*,The Scottish Consortium for ICPhS 2015 (Ed.), Glasgow, UK: the University of Glasgow. ISBN 978-0-85261-941-4. Paper number 04.621.1-5 retrieved from http://www.icphs2015.info/pdfs/Papers /ICPHS0462.pdf, 2015.

[21] A. Barabanov, V. Magerkin, and E. Vikulov, "Precise estimation of harmonic parameter trend and modification of a speech signal. *Speech and Computer, Lecture Notes in Artificial Intelligence*, Springer, vol. 9811, pp. 547–554, 2016.

[22] G. S. Ammar, and W. B. Gragg, "Superfast solution of real positive definite Toeplitz Systems," *SIAM J. Matrix Analysis and Appl.*, vol. 9, no. 1, pp. 61–76, 1988

AUTHOR INDEX

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 10 th international workshop : December 13-15, 2017*, ISBN 978-88-6453-606-4 (print) ISBN 978-88-6453-607-1 (online)