



UNIVERSITÀ
DEGLI STUDI
FIRENZE

DINFO
DIPARTIMENTO DI
INGEGNERIA
DELL'INFORMAZIONE

11th
INTERNATIONAL
WORKSHOP

MODELS AND
ANALYSIS
OF VOCAL
EMISSIONS
FOR
BIOMEDICAL
APPLICATIONS
December 17-19, 2019
Firenze, Italy



PROCEEDINGS



PROCEEDINGS E REPORT

ISSN 2704-601X (PRINT) | ISSN 2704-5846 (ONLINE)

**MODELS AND ANALYSIS OF VOCAL
EMISSIONS FOR BIOMEDICAL
APPLICATIONS**

11th INTERNATIONAL WORKSHOP

**December 17-19, 2019
Firenze, Italy**

**Edited by
Claudia Manfredi**

Firenze University Press
2019

Models and Analysis of Vocal Emissions for Biomedical Applications : 11th International Workshop, December, 17-19, 2019 / edited by Claudia Manfredi. – Firenze : Firenze University Press, 2019.
(Proceedings e report ; 122)

<https://www.fupress.com/isbn/9788864539454>

ISSN 2704-601X (print)

ISSN 2704-5846 (online)

ISBN 978-88-6453-951-5 (print)

ISBN 978-88-6453-961-4 (online)

Cover: designed by CdC, Firenze, Italy.

Peer Review Process

All publications are submitted to an external refereeing process under the responsibility of the FUP Editorial Board and the Scientific Committees of the individual series. The works published in the FUP catalogue are evaluated and approved by the Editorial Board of the publishing house. For a more detailed description of the refereeing process we refer to the official documents published on the website and in the online catalogue (www.fupress.com).

Firenze University Press Editorial Board

M. Garzaniti (Editor-in-Chief), M. Boddi, A. Bucelli, R. Casalbuoni, A. Dolfi, R. Ferrise, M.C. Grisolia, P. Guarnieri, R. Lanfredini, P. Lo Nostro, G. Mari, A. Mariani, P.M. Mariano, S. Marinai, R. Minuti, P. Nanni, G. Nigro, A. Perulli.

🌀 The on-line digital edition is published in Open Access on www.fupress.com.

The present work is released under Creative Commons Attribution 4.0 International license (CC BY 4.0: <http://creativecommons.org/licenses/by/4.0/legalcode>). This license allows you to share any part of the work by any means and format, modify it for any purpose, including commercial, as long as appropriate credit is given to the author, any changes made to the work are indicated and a URL link is provided to the license.

© 2019 Firenze University Press
Published by Firenze University Press
Firenze University Press
Università degli Studi di Firenze
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

*This book is printed on acid-free paper
Printed in Italy*



MAVEBA 2019

Firenze, Italy

The MAVeBA 2019 Workshop is sponsored by:

Università degli Studi di Firenze
Department of Information Engineering - DINFO



Int. Journal Biomedical Signal Processing and Control, Elsevier I.t.d

and is supported by:



Department of Information Engineering - DINFO



Fondazione Cassa di Risparmio di Firenze - FCRF

FONDAZIONE
CR FIRENZE

REGIONE TOSCANA



Regione Toscana - Consiglio Regionale

Consiglio Regionale



CoMeT Collegium Medicorum Theatri

CONTENTS

Foreword	XIII
----------------	------

SESSION I – PARKINSON’S DISEASE AND SPEECH

FUNDAMENTAL FREQUENCY AND DURATION CUES FOR SENTENCE FOCUS IN MANDARIN: PARKINSON’S AND HEALTHY SPEAKERS	17
Xi Chen, Diana Sidtis	

TEMPORAL REVERSION OF PHONATION INSTABILITY IN PARKINSON’S DISEASE BY NEUROACOUSTICAL STIMULATION	21
Pedro Gómez-Vilda, Gerardo Gálvez-García, Andrés Gómez-Rodellar, Daniel Palacios-Alonso, Guillermo de Arcas-Castro	

A NEUROMECHANICAL MODEL OF JAW-TONGUE ARTICULATION IN PARKINSON’S DISEASE SPEECH	25
A.Gómez, A. Tsanas, P. Gómez, D. Palacios, A. Álvarez, R. Martínez	

PARKINSON’S DISEASE CLASSIFICATION BASED ON VOWEL SOUND	29
Daria Hemmerling, David Sztaho	

ANALYSIS OF PHONATORY FEATURES FOR THE AUTOMATIC DETECTION OF PARKINSON’S DISEASE IN TWO DIFFERENT CORPORA.....	33
Laureano Moro-Velázquez, Jorge Andrés Gómez-García, Najim Dehak, Juan Ignacio Godino-Llorente	

JOINT ANALYSIS OF VOCAL JITTER, FLUTTER AND TREMOR IN VOWELS SUSTAINED BY NORMOPHONIC AND PARKINSON SPEAKERS	37
J. Schoentgen, A. Kacha, F. Grenez	

THE EFFECTS OF DEEP BRAIN STIMULATION ON SPEECH ARTICULATION AND VOCALIZATION IN PARKINSON’S DISEASE	41
John J. Sidtis, Diana Van Lancker Sidtis	

BIOMEDICAL SPEECH SIGNAL INSIGHTS FROM A LARGE SCALE COHORT ACROSS SEVEN COUNTRIES: THE PARKINSON’S VOICE INITIATIVE STUDY	45
Athanasios Tsanas, Siddharth Arora	

SESSION II – SINGING VOICE

TOWARDS A SOMATOSENSORY TRAINING DIGITAL ENVIRONMENT FOR LYRIC SINGING PEDAGOGY	51
Angelakis Evangelos, Georgaki Anastasia	

ASSESSING ARTICULATION IN SINGING DURING A VOWEL MATCHING EXERCISE USING ARTICULOGRAPHY: A PILOT STUDY	55
Helena Daffern, Amelia Gully	

VIII

TUNING TENDENCIES IN A SINGING QUINTET: EVOLUTION ACROSS REHEARSALS	59
Sara D'Amario, David M. Howard	
FACE VIBRATIONS MEASUREMENT IN SINGING — PILOT STUDY	63
Marek Frič, Pavel Dlask, Viktor Hruška	
COMPARISON OF SOUND RADIATION BETWEEN POP AND CLASSICAL SINGERS	67
Iva Podzimková, Marek Frič	
DISSOCIATION OF SPOKEN AND SUNG VOCAL PRODUCTION	71
Diana Van Lancker Sidtis, Y-J. Kim, John J. Sidtis	
SHOULD I OPEN MY MOUTH MORE TO SING LOUDER?	75
Allan Vurma	

SESSION III - SPECIAL SESSION - EDUCATION AND REHABILITATION OF THE ARTISTIC VOICE

SOVTE: BETWEEN MYTH AND REALITY	81
F. Fussi	
THE SOVTE PROTOCOLS: INDICATIONS AND CRITICISM IN VARIOUS SINGING STYLES	81
E. Biavati, E. Bruni	
G.E.M.M.A. TRAINING: A LINK BETWEEN PSYCHOLOGY AND LOGOPAEDICS IN THE TREATMENT OF SINGING VOICE DISORDERS	81
E. Rosa	

SESSION IV - VOICE AND EMOTIONS

THE CORRELATION BETWEEN POETIC RHYTHM AND HEART RATE VARIABILITY IN SUBJECTS READING AND PERCEIVING RUSSIAN POETRY	85
Karina Evgrafova, Pavel Skrelin, Vera Evdokimova, Tatjana Chukaeva	
EMOTIONALLY EXPRESSED VOICES ARE RETAINED IN MEMORY FOLLOWING A SINGLE EXPOSURE	89
Y-J. Kim, J. J. Sidtis, D. Van Lancker Sidtis	
ELECTRODERMAL ACTIVITY AND SPEECH FEATURES AS PREDICTORS FOR AROUSAL LEVEL CHANGES AFTER AFFECTIVE WORD PRONUNCIATION"	93
Nicola Vanello, Alberto Greco, Claudia Marzi, Enzo Pasquale Scilingo	
A COMMONLY SHARED CODE FOR SADNESS IN HUMAN VOCALIZATIONS AND MUSIC REVEALED BY HUMAN INFANT CRIES	97
G. Zeloni, F. Pavani	

SESSION V - VOICE QUALITY

TOWARDS ROBUST FEATURES IN VOICE DISEASE DIAGNOSIS: MFCCs VS. PNCCs.....	103
M. Madruga, Y. Campos-Roca, C. J. Pérez	

ELECTROGLOTTOGRAPHIC VOICE MAPS OF UNTRAINED VOCALLY HEALTHY ADULTS, WITH GENDER DIFFERENCES AND GRADIENTS	107
Sten Ternström, Rita R. Patel	

ACOUSTIC AND ELECTROGLOTTOGRAPHIC PARAMETRISATION OF PHONATORY QUALITY PROVIDE VOICE PROFILES OF PATHOLOGICAL SPEAKERS	111
Manfred Pützer, Wolfgang Wokurek	

VOCAL QUALITIES OF SARCASTIC UTTERANCES: CROSS-LINGUISTIC STUDY OF ENGLISH AND KOREAN	115
Seung-yun Yang	

SESSION VI - VOCAL FOLDS DYNAMICS

PERTURBATION OF CYCLE LENGTHS AND CYCLE PEAK AMPLITUDES IN DIPLOPHONIC VOICES	121
Philipp Aichinger	

TRACKING OF MULTIPLE FUNDAMENTAL FREQUENCIES IN STANDARD TEXT READINGS OF DIPLOPHONIC SPEAKERS.....	125
Philipp Aichinger	

AERODYNAMICS OF GLOTTAL VIBRATION ONSET	129
P. H. DeJonckere, J. Lebacqz	

A GLOTTAL AREA WAVEFORM MODEL FOR MULTI-PULSED VOCAL FRY.....	133
Vinod Devaraj, P. Aichinger	

MODELLING LONGITUDINAL PHASE DIFFERENCES IN A LUMPED AND DISTRIBUTED ELEMENTS VOCAL FOLD MODEL	137
Carlo Drioli, Philipp Aichinger	

EXTRACTING VOCAL FOLD KINEMATIC PARAMETERS FROM VIDEOKYMOGRAMS VIA SIMULATION OF CLINICALLY OBSERVED DATA.....	141
Sridhar Bulusu, S. Pravin Kumar, Jan G. Svec, Philipp Aichinger	

VOCAL FOLD OSCILLATORS AT LARGE ASYMMETRIES	145
Jorge C. Lucero, Xavier Pelorson, Annemie Van Hirtum	

PHYSICAL STUDY OF THE INFLUENCE OF LEFT-RIGHT VOCAL FOLDS ANGLE ASYMMETRY ON PHONATION.....	149
Annemie Van Hirtum, Anne Bouvet, Xavier Pelorson, Isao Tokuda	

THEORETICAL AND EXPERIMENTAL MODELING OF LESIONS OF THE VOCAL FOLDS.....	153
Xavier Pelorson, Anne Bouvet, Annemie Van Hirtum	

VOICE SOURCE EFFECTS OF FUNDAMENTAL FREQUENCY VARIATION.....	157
Johan Sundberg	

SESSION VII - COMET SESSION - ACTOR'S AND ACTRESS' VOICES

SPECTRAL SPECIFICITIES OF ACTING VOICE IN PROFESSIONAL ACTRESSES	163
P.H. DeJonckere, H. Stoffels	
SHORT TERM EFFECT OF 'SEMI-OCCLUDED VOCAL TRACT EXERCISES' ON HEALTHY ACTORS' VOICES	167
Valentina Di Natale, Giovanna Cantarella, Claudia Manfredi, Annaclara Ciabatta, Cosimo Becherini, Philippe DeJonckere	
SINGING WHILE ACTING AND VICE-VERSA	171
Orietta Calcinoni	
THE SHOUTING VOICE AND THE IMPORTANCE OF SHOUTING ABILITY AS A 'FITNESS' PARAMETER FOR ALL VOICE USERS.....	175
Josef Schlömicher - Thier, Hannes Tropper, Ingolf Franke	
THE ACTORS VOICE: LAUGHING, CRYING AND SHOUTING IN THE MRI	179
Bernhard Richter, Louisa Traser, Michael Burdumy, Matthias Echternach, Claudia Spahn	
DISTORTED VOCALITY: SUPRAGLOTTIC MANAGEMENT IN ACTING AND DUBBING	181
Franco Fussi, Eleonora Bruni	

SESSION VIII – VOCAL FOLDS PARALYSIS/ABNORMALITIES

PHONOSURGICAL TREATMENT OF BILATERAL LARYNGEAL PARALYSES.....	185
Giovanna Cantarella	
SELECTIVE SURFACE STIMULATION IN UNILATERAL VOCAL FOLD PARALYSIS (UVFP)...	189
Berit Schneider-Stickler, Matthias Leonhard, Matthias Krenn, Winfried Mayr	
ULTRAHIGH RESOLUTION OPTICAL COHERENCE TOMOGRAPHY FOR DETECTING TISSUE ABNORMALITIES OF THE ORAL AND LARYNGEAL MUCOSA: A PRELIMINARY STUDY	195
Niels Møller Israelsen, Anders Overgård Jønsson, Mette Pedersen	

SESSION IX - KEYNOTE LECTURE

DEVELOPING NEW SPEECH SIGNAL PROCESSING ALGORITHMS FOR BIOMEDICAL AND LIFE SCIENCES APPLICATIONS: PRINCIPLES, FINDINGS, CHALLENGES, AND A VIEW TO THE FUTURE.....	201
Athanasios Tsanas	

SESSION X - BIOMECHANICS/DEVICES

EXPERIMENTAL MODELLING OF GLOTTAL AREA DECLINATION RATE IN VOWEL AND RESONANCE TUBE PHONATION.....	205
Jaromír Horáček, Radolf Vojtěch, Bula Viteslav, Anne-Maria Laukkanen	

DEVELOPMENT AND USE OF AN ANECHOIC SUBGLOTTAL TRACT FOR EXCISED LARYNX EXPERIMENTS.....	209
Hugo Lehoux, Vít Hampala, Jan G. Švec	

SHAKER: PRELIMINARY OBSERVATIONS OF A POTENTIAL DEVICE FOR VOICE TRAINING AND THERAPY.....	213
Anne-Maria Laukkanen , Jaromír Horáček , Vojtěch Radolf	

SESSION XI - SPEECH

BEATBOX SOUNDS RECOGNITION USING A SPEECH-DEDICATED HMM-GMM BASED SYSTEM.....	219
Solene Evain, Adrien Contesse, Antoine Pinchaud, Didier Schwab, Benjamin Lecouteux, Nathalie Henrich Bernardoni	

AM-FM DECOMPOSITION OF SPEECH SIGNAL: APPLICATIONS FOR SPEECH PRIVACY AND DIAGNOSIS.....	223
Petr Motlicek, Hynek Hermansky, Srikanth Madikeri, Amrutha Prasad, Sriram Ganapathy	

THE TOMATIS ELECTRONIC EAR EFFECTS ON SIMPLE VOCALIZATIONS	227
Marco Prenassi, Walter Coppola, Giovanni Ramponi, Tiziano Agostini, Sara Marceglia	

SESSION XII - VOICE VS OTHER PHYSIOLOGICAL SIGNALS/DISEASES

THE INTELLIGIBILITY OF SPEECH IN POLISH SPEAKERS BORN WITH CLEFT LIP AND/OR PALATE.....	233
Wiktor Gonet, Maria Hortis-Dzierzbicka, Edyta Zomkowska	

THE EFFECT OF PHYSICAL ACTIVITY ON THE PHONETIC CHARACTERISTICS OF SPEECH.....	237
V. V. Evdokimova, E. A. Zakharchenko	

CHANGES TO VOICE PRODUCTION CAUSED BY LONG-TERM HEARING LOSS (HL).....	241
I.V. Kastyro, A.N. Kovalenko, Torshin V.I., Draginskaya E.S.	

COMPARISON OF IMMEDIATE EFFECTS OF VOCAL BREATHING EXERCISES AND PHYSICAL EXERCISES ON HEART RATE VARIABILITY (HRV) IN HEALTHY STUDENTS	245
A.N. Kovalenko, I.V. Kastyro, V.I. Torshin, Y.S. Guhschina, E.S. Doroginskaya, N.A. Kamanina	

AN ALGORITHM FOR DETECTING THE ONSET OF LINGUISTIC SEGMENTS IN CONTINUOUS ELECTROENCEPHALOGRAM SIGNALS	249
C. Tonatiah Hernández-del-Toro, Carlos A. Reyes-García	

DISCRIMINATION BETWEEN CHILDREN AND ADULT FACES USING BODY AND HEAD RATIO AND GEOMETRIC FEATURES	253
C.A. Reyes-García, E. Morales-Vargas, H. Peregrina-Barreto, C. Manfredi	

SONIFICATION TECHNIQUES APPLIED TO EEG SIGNALS OF NONMOTOR GENERALIZED ONSET EPILEPTIC SEIZURES	257
L. Frassinetti, R. Guerrini, C. Barba, F. Melani, F. Piras, C. Manfredi	

BIOVOICE: A MULTIPURPOSE TOOL FOR VOICE ANALYSIS.....	261
Maria Sole Morelli, Silvia Orlandi, Claudia Manfredi	
INDEX OF AUTHORS	265



MAVEBA
2019
Firenze, Italy

FOREWORD

This book of Proceedings includes the contributions presented at the 11th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications – MAVeBA 2019, held in Firenze from 17 to 19 December, 2019. That is, 20 years since the very first MAVeBA in 1999!

Looking back to those days, I remember well the spirit of adventure that inspired this initiative, both on my side and on that of my colleague Piero Bruscaaglioni, with whom I also shared many subsequent MAVeBA editions.

MAVeBA started because of our curiosity and continued thanks to the enthusiasm of the participants. And today? Curiosity and enthusiasm are still there, with the awareness of a fascinating and increasingly interdisciplinary world. The large number of contributions collected in this Proceedings is the clear demonstration of this.

The main subjects concern methods for analyzing hoarseness and retrieving features of the human voice related to particular physiological or neurological conditions, with the aim of assessing reliable procedures for objective, quantitative definition of levels of voice disorders, singing voice parameters, newborn cry features, vocal fold and vocal tract modelling. The interdisciplinarity, that has always characterized the MAVeBA workshops, is well highlighted by the themes addressed, listed below.

I wish to give special thanks and greetings to the CoMeT Association, that is present at MAVeBA with a large number of its members. This year is a special one for CoMeT, celebrating the 50th anniversary from its foundation, and I am happy and proud to celebrate it together with the twenty-year anniversary of MAVeBA!

The papers presented at MAVeBA and collected in this volume are divided into nine Sessions, two Special Sessions, professionally coordinated by Dr. Franco Fussi and Dr. Philippe Dejonckere, and a Keynote lecture given by Thanasis Tsanas.

SESSION I – PARKINSON’S DISEASE AND SPEECH

SESSION II – SINGING VOICE

SESSION III - SPECIAL SESSION – EDUCATION AND REHABILITATION OF THE ARTISTIC VOICE

Coordinator: Franco Fussi

SESSION IV - VOICE AND EMOTIONS

SESSION V - VOICE QUALITY

SESSION VI - VOCAL FOLDS DYNAMICS

SESSION VII - COMET SESSION – ACTOR’S AND ACTRESS’ VOICES

Coordinator: Philippe Dejonckere

SESSION VIII – VOCAL FOLDS PARALYSIS/ABNORMALITIES

SESSION IX - KEYNOTE LECTURE

DEVELOPING NEW SPEECH SIGNAL PROCESSING ALGORITHMS FOR BIOMEDICAL AND LIFE SCIENCES APPLICATIONS: PRINCIPLES, FINDINGS, CHALLENGES, AND A VIEW TO THE FUTURE

Thanasis Tsanas

SESSION X - BIOMECHANICS/DEVICES

SESSION XI – SPEECH

SESSION XII - VOICE VS OTHER PHYSIOLOGICAL SIGNALS/DISEASES

I am very grateful to the authors for their contribution and to all participants that stimulated the discussion and helped to propose new research themes and methodologies of analysis in a field that will always be evolving, even and hopefully in the next twenty years.

Claudia Manfredi

ACKNOWLEDGEMENTS

I greatly acknowledge the ScaramuzziTeam Congress Agency for its great professionalism, Dr.Eng. Alice Cavaliere, who manages and constantly updates the website, and Dr. Eng. Lorenzo Frassinetti, PhD student, who collaborated in reviewing the Proceedings and solving the daily difficulties with patience and professionalism.

SESSION I
PARKINSON'S DISEASE AND SPEECH

FUNDAMENTAL FREQUENCY AND DURATION CUES FOR SENTENCE FOCUS IN MANDARIN: PARKINSON'S AND HEALTHY SPEAKERS

Xi Chen^{1,2}, Diana Van Lancker Sidtis^{1,2}

¹ Department of Communicative Sciences and Disorders, New York University, New York, US

² Brain & Behavior Laboratory, Nathan Kline Institute, Orangeburg, NY, US
x.chen@nyu.edu, diana.sidtis@nyu.edu

Abstract: Speech prosody is impaired in persons with Parkinson's Disease (PWP) due to the disfunction of motor control. PWP demonstrate deficits in pitch, temporal cues, and loudness in vocal production. However, there is a lack of evidence regarding vocal disorders in PWP who speak tone languages. In a tone language such as Mandarin, pitch carries a heavier functional load. Pitch carries not only affective and linguistic information as it does in non-tone languages, but also serves to distinguish lexical meanings. Being an important aspect of speech prosody, sentence focus, used to convey intonation contrasts for discourse purposes, is assumed to have a complex prosodic pattern in Mandarin speaking PWP. The current study is designed to investigate the acoustic characteristics of sentence focus in this population. Mandarin speaking PWP and healthy controls were recruited to produce sentences with and without contrastive focus in elicited speech. Acoustic analysis was conducted on the vocal characteristics, pitch and duration.

Keywords: Sentence focus, tone language, Parkinson's disease speech, acoustic analysis, F0 and duration cues

I. INTRODUCTION

Sentence focus is the use of vocal emphasis on a syllable or word to create contrast with the rest of the utterance for discourse or pragmatic purposes [1]. In many languages, such as English and German [2], sentence focus is realized through raising the fundamental frequency (F0) and increasing the duration of the focused words. Studies of Mandarin reported greater pitch range [3] and increased duration [4] on focused syllables. These studies were based on healthy speakers. To our knowledge, sentence focus in persons with Parkinson's Disease (PWP) has not been investigated in any of the languages.

Parkinson's disease (PD) is a progressive neurodegenerative disorder characterized by dopaminergic deficiency in the brain [5]. It not only damages patients' physiological movement, but also impairs voice and speech functions. The loss of dopamine in basal ganglia results in reduced physiological support of breath and control of the

phonatory and articulatory organs [6]. Consequently, PWP demonstrate impaired voice quality and speech prosody characterized by breathiness, monoloudness, and aberrant pitch and temporal features. Previous studies on speech and voice of PWP were based on non-tone language speakers. In tone languages, pitch is not only used to mark the intonational curve as it does in non-tone language, but it is also used to differentiate lexical meanings. Therefore, speech prosody in PWP who speak a tone language may have more complicated patterns.

The present study endeavors to answer the following questions: 1. Do PD and healthy speakers make a significant difference in F0 and/or duration measures on words targeted for focus as compared with neutral exemplars; 2. In producing sentence focus, do PD speakers exhibit different patterns of performance from healthy controls in use of F0 and/or duration? The findings of this study will demonstrate the acoustic properties of sentence focus produced by PWP and healthy speakers of Mandarin. This research will also contribute to the understanding of cerebral processing of voice, pitch, and temporal cues.

II. METHODS

Procedures

Sixteen individuals with PD (8 males, 8 females) and 16 age-matched healthy control (HC) speakers (8 males, 8 females) were recruited. All participants were Mandarin monolinguals who are the residents of Baoji, China.

The study utilized an elicitation task, whereby drawings depicting activities of animal figures were shown to the speakers and the participants were instructed to describe what the animal is doing on each picture with a given sentence structure (Subject + Verb + Object). The descriptions were spoken with neutral focus (no contrastive sentence focus). The experimenter then asked three questions addressing the three different components of the original sentence. Each question provided a contrastive content in order to elicit the correct, focused lexical item from the participant. The participants' task was to answer each question. The

correct answers were the original sentence type, but each spoken with a different contrastive sentence focus.

For example, with a line drawing depicting an animated turtle sitting on an airplane with its hands on the rudder, the experimenter asked (in Mandarin): who is doing what? The participants were trained to answer: wū guī kāi fēi jī (A turtle is flying a plane). The experimenter then asked a question regarding a component of this picture, such as: is a dog flying the plane? The participants were supposed to answer: bù, shì wū guī kāi fēi jī (no, the turtle is flying the plane). The correct answer should have the original sentence the participants used to describe the picture, but with a contrastive focus on a particular word to highlight the information they want to emphasize. In the example answer, the word “wū guī” (turtle) would be emphasized by the participants.

Acoustic analysis

The acoustic analysis was performed using the acoustic analysis software Praat. F0 maximum, F0 minimum and duration were measured for words carrying contrastive focus and the corresponding neutral exemplar. F0 range were obtained by subtracting F0 minimum from F0 maximum. Considering gender difference, ratios were used in the study expressing the degree to which F0/duration measures in the focused and neutral words differ. The analysis is still on-going. The current preliminary analysis involved data of 9 female participants (5 PD and 4 HC).

III. RESULTS

For within-group analyses, one-sample t-tests on ratios of F0 range and ratios of duration change (from the contrastive focus tokens to corresponding neutral exemplars) were performed for the two groups respectively. This is to examine the amount of change in F0 range and duration made by the speakers in signaling sentence focus. The results showed that the ratio of F0 range was not significantly different from zero for either the PD group ($t = -1.76, p = 0.10$) or the HC group ($t = 1.03, p = 0.16$). However, the ratio of duration was significantly different from 0 (i.e., greater than 0) within each group: PD ($t = 3.27, p = 0.002$) and HC ($t = 5.60, p < 0.001$). Regarding group differences, independent t-tests were conducted on F0 range ratios and duration ratios between the two groups in order to examine whether the PD group utilized F0 and duration cues differently from the control group. Fig. 1 shows F0 range ratios of the PD and the HC groups. It was found that the F0 range ratios were not significantly different between the two groups ($t = 0.87, p = 0.39$). Fig. 2 displays duration ratios between the two groups. Unlike F0 range ratios, the results showed that HC group had

significantly greater duration ratios than the PD group ($t = -2.59, p = 0.02$).

Fig. 1, F0 range ratios between PD and HC groups

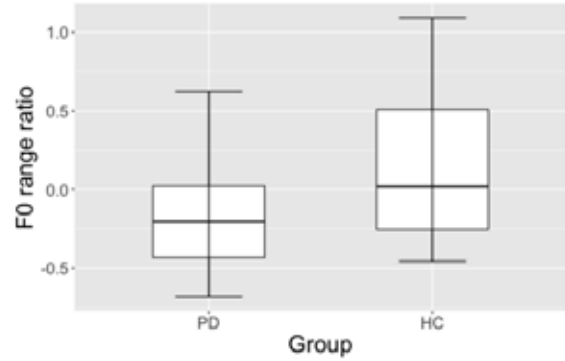
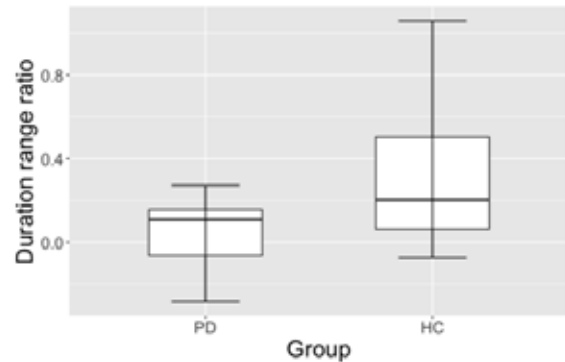


Fig. 2, Duration ratios between PD and HC groups



IV. DISCUSSION

The current study investigated the characteristics of F0 and duration cues of sentence focus produced by PWP and healthy speakers who are speakers of a tone language (i.e., Mandarin). The preliminary results suggested that duration plays an important role for each group in producing contrastive sentence focus. The PD group is less efficient in using duration to signal sentence focus compared to the healthy controls, but both groups produced longer duration in contrastive focus compared to the neutral exemplar. Since the PD group in our data was able to expand duration range for the salience of contrastive sentence focus, it is postulated that the ability to use duration cues is intact in the PWP, even though this ability is compromised to some degree. In contrast, F0 range was a less crucial cue for sentence focus in either the PD group or the HC group. Nor was there any significant difference in the use of F0 range between the two groups.

Previous research has found that PWP have abnormal temporal patterns of speech. It was reported that PWP had accelerated speech rate and reduced

syllable duration [7][8]. On the other hand, PWP were also found to have decreased speech rate [9]. Despite these reported global temporal deficits of speech rate, extending over phrases and sentences, not all temporal cues are compromised in PWP. A research study showed that the relational timing in word production is preserved in PWP. In that study, PWP were found to be able to produce a shorter syllable in a long word just as the healthy speakers do [10]. The findings of the current study support the notion that temporal cues in speech, in unit-level planning, might be preserved in PWP. It can be used for pragmatic or discourse purposes to compensate the impairment of F0 cues, which may be damaged more severely. As for our findings of F0 range, the discrepancy with previously reported results remains to be explained. The current analysis is preliminary with 9 participants (5 PD and 4 HC). More data probing additional F0 measures and extended across 32 participants may provide a fuller profile of voice characteristics in communicating focus.

REFERENCES

- [1] Bolinger, D. L. (1958). A theory of pitch accent in English. *Word*, 14(2-3), 109-149.
- [2] Eady, S. J., Cooper, W. E., Klouda, G. V., Mueller, P. R., & Lotts, D. W. (1986). Acoustical characteristics of sentential focus: Narrow vs. broad and single vs. dual focus environments. *Language and speech*, 29(3), 233-251.
- [3] Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f0 contours. *Journal of Phonetics*, 27(1), 55-105. doi: <https://doi.org/10.1006/jpho.1999.0086>
- [4] Liu, F., & Xu, Y. (2005). Parallel encoding of focus and interrogative meaning in Mandarin intonation. *Phonetica*, 62(2-4), 70-87. doi:10.1159/000090090
- [5] Lang, A. E., & Lozano, A. M. (1998). Parkinson's Disease. *New England Journal of Medicine*, 339(16), 1130-1143. doi:10.1056/nejm199810153391607
- [6] Goberman, A. M., & Coelho, C. (2002). Acoustic analysis of Parkinsonian speech I: Speech characteristics and L-Dopa therapy. *NeuroRehabilitation*, 17(3), 237-246.
- [7] Ackermann, H., Konczak, J., & Hertrich, I. (1997). The temporal control of repetitive articulatory movements in Parkinson's disease. *Brain and Language*, 56(2), 312-319.
- [8] Ackermann, H., & Ziegler, W. (1991). Articulatory deficits in parkinsonian dysarthria: an acoustic analysis. *Journal of Neurology, Neurosurgery & Psychiatry*, 54(12), 1093-1098.
- [9] Ludlow, C. L., Connor, N. P., & Bassich, C. J. (1987). Speech timing in Parkinson's and Huntington's disease. *Brain and Language*, 32(2), 195-214.
- [10] Sidtis, J. J., & Sidtis, D. V. L. (2012). Preservation of relational timing in speech of persons with Parkinson's disease with and without deep brain stimulation. *Journal of medical speech-language pathology*, 20(4), 140

TEMPORAL REVERSION OF PHONATION INSTABILITY IN PARKINSON'S DISEASE BY NEUROACOUSTICAL STIMULATION

Gerardo Gálvez-García¹, Andrés Gómez-Rodellar^{1,2}, Daniel Palacios-Alonso³, Guillermo de Arcas-Castro¹, Pedro Gómez-Vilda²

¹Neuroacoustics Laboratory, Universidad Politécnica de Madrid, Campus Sur, Ctra. de Valencia km. 7, 28031 Madrid, Spain

²Neuromorphic Speech Processing Laboratory, Center for Biomedical Technology, Universidad Politécnica de Madrid, Campus de Montegancedo, s/n, 28223 Pozuelo de Alarcón, Madrid, Spain

³Escuela Técnica Superior de Ingeniería Informática - Universidad Rey Juan Carlos, Campus de Móstoles, Tulipán, s/n, 28933 Móstoles, Madrid, Spain

gerardo.galvez@i2a2.upm.es, andres.gomez@ctb.upm.es, daniel.palacios@urjc.es, garcas@sec.upm.es, pedro@fi.upm.es

Abstract: This paper investigates the use and effects of neuroacoustical stimulation of Parkinson's Disease patients on the stability of phonation as a means of assessing improvements in their neuromotor condition. Four patients have been submitted to active and inert signal mixtures including binaural beats and pink noise. A positive effect (reduction in phonation features tracking instability) has been found in all the patients studied. The effects seem to be statistically relevant, although the experimental framework did not allow to assess the duration of the stabilization effects. More research is needed to generalize the results and establish the optimal signal configurations and stimulation protocols.

Keywords: Neuromotor diseases, Parkinson, Phonation instability, neuroacoustical stimulation, binaural beats

I. INTRODUCTION

The effects of neurodegenerative diseases on speech are well known. Phonation, articulation, prosody and fluency are speech characteristics strongly affected by Parkinson's Disease (PD) [1]. On the other hand, it is also well known that neuroacoustical stimulation based on *binaural beats* may modify the brain activity measured in the cortex by electroencephalography (EEG) [2]. Binaural beats stimulation is a perceptual phenomenon where two sine waves are presented to a subject separately through each ear [3]. This study is intended to test if neuroacoustical stimulation is able to modify the phonation stability of PD patients by the evaluation of phonation features derived from the acoustic analysis of PD patients phonation before (pre) and after (post) stimulation. Relevant aspects as the signal combinations used in active and inert stimulation are presented and discussed, as well as the order in which stimuli are applied. The statistical analysis of data is also presented and discussed.

Results seem to avail the feasibility of this methodology in improving neuromotor stability in PD patients.

II. METHODS

Four volunteer speakers, two women, 56 and 66 years old, and two men, both 64 years old, in stage 2 and 3 of Hoehn & Yahr scale (H&Y) were recruited from the Parkinson's Disease Association of Madrid. These were labeled as F1/66, F2/56, M1/64 and M2/64, respectively. A summary of their specific conditions is given in Table 1.

Table 1 Description of the participants' H&Y stage.

Code	Gender	Age	Time since 1 st diagnostic (y)	H&Y
F1/66	F	66.47	9	2
M1/64	M	64.66	5	3
F2/56	F	56.22	5	2
M2/64	M	64.43	3	2

Each patient was submitted to neuroacoustical stimulations and speech recordings. The stimuli included active (supposedly exerting a modification in the functional performance of the patient to approach normativity) and inert or neuter (supposedly not exerting such positive modification of functional performance). The active stimulations consisted of pink noise and a pair of sine waves of 154 and 168 Hz, applied to the left and right ears during 10 minutes to generate a perceptual beat in the upper auditory pathways of 14 Hz, as this frequency is supposed to be associated to the EEG activity decay detected in PD [4]. The inert stimulations were based on the application of simple pink noise during the same amount of time. The specific protocol is shown in Figure 1. In a first session (top) patients were randomly selected to receive an active or an inert stimulation. In a second session separated at least seven days from the

first one, the patient received an inert or an active stimulation (the reverse than the one in the first session). The stimulation session (bottom) consisted of a voice recording task (pre-stimulus), followed by a binaural stimulation task lasting 10 minutes, to end with another voice recording task (post-stimulus). Recording tasks of at least two-seconds of the vowel [a:] were conducted immediately before (pre) and after (post) each stimulation session. These were analyzed following the phonation feature extraction described in [5]. The whole study was approved by the Ethical Committee of Universidad Politécnic de Madrid, and it adhered to the Declaration of Helsinki. All the volunteer participants signed an informed consent document.

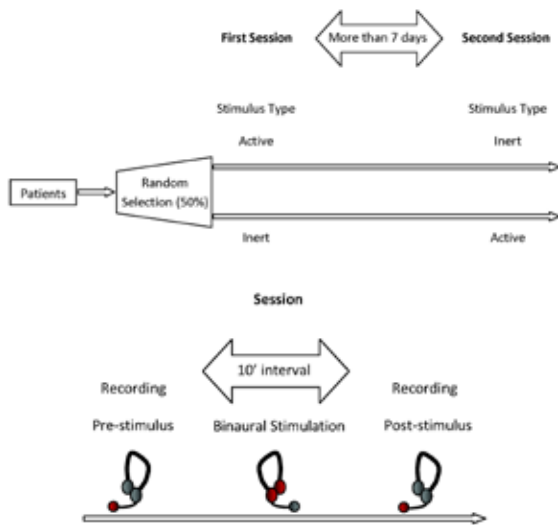


Figure 1. Top: Description of the neurostimulation session sequencing. Bottom: Timing of a session.

The stability of phonation was described by certain types of features as distortion, biomechanical instability and tremor. The distortion features used were jitter and shimmer. Both are seen as relative fluctuations of the periodicity and amplitude of the glottal source, given as the instability index

$$I_k = 2 \frac{F_k - F_{k-1}}{F_k + F_{k-1}} \quad (1)$$

where k is the glottal cycle index and F_k is the feature implicit (for jitter F_k is the cycle period, and for shimmer it is the amplitude of the glottal source cycle). The biomechanical instability features are evaluated following I_k when the features F_k are the mass and stiffness associated to the vocal folds estimated from a two-mass biomechanical model [6].

Finally, other important correlates of phonation instability are tremors. Classically tremor in voice has been divided into three bands, known as physiological (between 2-4 Hz), neurological (5-8 Hz) and flutter (9-12 Hz) [7]. Tremor may be derived from the distributed body stiffness [8]. The most relevant tremor features are the amplitudes associated to each of the three bands and the root-mean square value of tremor. As the minimum band to be estimated (physiological) may be within 2-4 Hz, at least 500 ms long windows should be used, covering most of the vowel phonation except the leading and trailing edges (first and last 200 ms).

III. RESULTS

Ten features from the analysis of the maintained vowel [a:] corresponding to the two sessions (pre- and post- both examined at intervals of 500 ms) were stored in their respective matrices (\mathbf{X}_{pr} and \mathbf{X}_{ps}) by feature and estimation session. The features are the ones seen in Figure 2, corresponding to S1-pre, S1-post, S2-pre and S2-post: jitter relative (2), shimmer relative (3), body mass and stiffness unbalance (38 and 40), body cover and stiffness unbalance (44 and 46), physiological tremor amplitude (67: 2-4 Hz band), neurological tremor amplitude (69: 5-8 Hz), flutter band amplitude (71: 9-12 Hz) and overall RMS tremor amplitude (72). The ordinals are relative to the features extracted by BioMet®Tools [9].

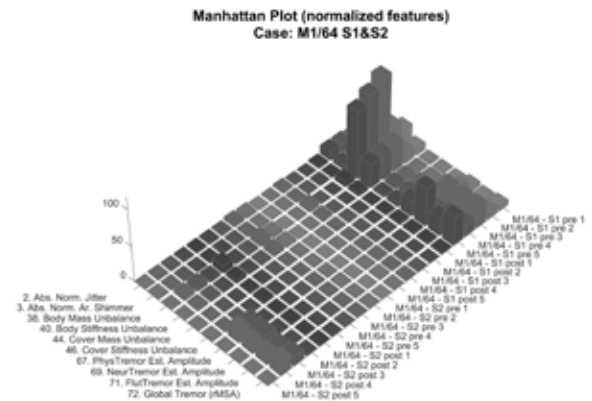


Figure 2. Plot of the normalized phonation features from subject M1/64 for the 5-pre and 5-post tests of session 1 (S1: active stimulation) and session 2 (inert stimulation).

Features are represented normalized to their averages on a population of 50 male normative subjects in the range from 20-60 years old, recruited at the ENT services of Hospital Universitario Gregorio Marañón of Madrid. A similar normative database is used for female subjects. It must be stressed that larger values mean more deviation from the normative values. It

may be seen that the five S1 pre-samples show large deviations in features 38 and 40 (vocal fold body biomechanical unbalances: differences in estimations on neighbor phonation cycles) and in features 67-72 (tremor). These deviations have disappeared in the five S1 post-samples, more aligned with the normative database. This same situation is observed in the five S2 pre-samples, showing only slight body unbalances (38 and 40) in samples S2 pre 1 and S2 pre 2, whereas S2 post 3 and S2 post 4 show a larger deviation affecting the four tremor features. This behavior will be reflected in the log-likelihood ratios shown in Table 3, expressing an improvement in S1 and a worsening in S2.

Due to the low number of subjects implied in the experimentation, the use of parametric and non-parametric methods was preferred to offer higher contrast robustness in the subsequent statistical analysis. The statistical relevance of the neurostimulation was evaluated by means of t-Student, Kolmogorov-Smirnov and Mann-Whitney U tests on \mathbf{X}_{pr} and \mathbf{X}_{ps} under the null hypothesis of equal means. The p-values of the three tests (pvSt: t-Student, pvKS: Kolmogorov-Smirnov, pvMW: Mann-Whitney) of the four patients mentioned are given in Table 2.

Table 2 Results of the relevance tests.

Patient	pvSt	pvKs	pvMW
F1/66	9.2835e-05	7.8922e-09	7.5609e-11
F2/56	0.1590	9.2799e-04	8.9078e-04
M1/64	4.2094e-09	1.2940e-14	9.1501e-15
M2/64	3.4648e-05	6.8543e-06	5.8886e-07

These results avail statistical significance to the rejection of the null hypothesis for the four cases under Kolmogorov-Smirnov and Mann-Whitney tests, whereas t-Student tests avail the rejection for F1/66, M1/64 and M2/64. These results have to be interpreted in the sense that apparently the pre- and post-stimulus features differ among themselves enough to consider that there has been some effect in the neuromotor function of the patient after the stimulation, with the exception of F2/56 based on the t-test, which fails in rejecting the null hypothesis, thus putting some doubt on the effects of stimulation in this case, although non-parametric tests would avail the effects of stimulation. Although, these tests show the different statistical behavior of data distributions before and after stimulation, they do not show if the post-stimulus function is acting to regress the neuromotor function to the normative behavior or not. To infer if there has been any improvement of the post-stimulus feature estimations (\mathbf{X}_{ps}) with respect to the pre-stimulus estimates (\mathbf{X}_{pr}), relative pre-post log-likelihood ratios are to be used, defined as

$$\lambda = \log \left\{ \frac{p(\mathbf{X}_{ps} | \Gamma)}{p(\mathbf{X}_{pr} | \Gamma)} \right\} \quad (2)$$

where $\Gamma = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ is the before-mentioned normative male or female model respectively, in terms of feature averages and covariance matrices. The respective log-likelihoods are supposed positive if there is an improvement, and negative otherwise. The log-likelihood ratios for the active (λ_A) and inert (λ_I) stimulations are given in Table 3.

Table 3 Log-likelihood ratios for active and inert stimulations.

Patient	λ_A	λ_I
F1/66	74.261	-31.521
F2/56	7.627	278.487
M1/64	167.695	-31.877
M2/64	91.755	-13.191

Now, it seems that a relevant improvement in neuromotor stability has been observed in F1/66, M1/64 and M2/64 when active stimulation was used, whereas there has been some worsening after inert stimulation. The case of F2/56 shows exactly the reverse situation: the improvement after session 1 is almost not relevant, whereas the improvement after session 2 is strongly relevant. As the order of stimulation (active/inert) was at random, the only explanation for this apparent paradox is that the inert stimulation was used in session 1, whereas the active one was used in session 2. The review of the experiment log confirmed this assumption.

IV. DISCUSSION

The issue of neuroacoustical stimulation to change the functional behavior of phonation in PD patients is quite recent and subject to all kinds of controversy. It is quite unclear if these changes are long lasting or temporary and confined to some minutes after active stimulation has been applied, although results seem to support that some improvements can be appreciated when the active stimulus of the kind described is applied. It is also intriguing if some effects could be perceived in other motor functions, and in cognitive activity as well, these effects being under test using electroencephalography (EEG), although data analytics become more complicate and results are pending. Another relevant question is the design of the active and inert signals. A first approach based on the combination of pink noise and binaural beats in the band 150-160 Hz seems to be efficient, but other combinations of signals producing beats not only in the band of 14 Hz, but in others related with specific EEG

cycles are to be tested as well. A mention to the configuration of the inert stimulation is also due. It seems that pink noise is behind the activation or enhancement of certain cognitive processes [10], although more research work is needed in this respect. What seems more intriguing is that the use of pink noise as the inert stimulation produced the reverse effect in three of the four subjects tested, i.e., phonation stability turned worse after inert stimulation. Needless to say, this seems a quite exciting and challenging topic.

V. CONCLUSIONS

From the results shown, it could be assumed that a certain positive effect is produced by active neuroacoustical stimulation in the phonation neuromotor stability in three out of four patients subject to the specified stimulation conditions and protocol. It seems also clear that the fourth patient benefited from active neurostimulation when applied in the reverse order (a week after the inert one). It has to be better established if this effect is permanent or transitory, and in this second case, for how long. Anyway, the results shown suggest that the stimulation produces a measurable effect, although a deeper, wider and longer study on other concomitant factors is still pending.

ACKNOWLEDGMENTS

This research has been funded by grants TEC2016-77791-C4-4-R (MINECO, Spain) and CENIE_TECA-PARK_55_02 INTERREG V-A Spain – Portugal (POCTEP). The authors would like to thank ‘Asociación de Parkinson de Madrid’, and Dr. J. C. Martínez-Castrillo from Hospital Universitario Ramón y Cajal of Madrid for their help and advice, and especially to the Foundation for the Promotion of Industrial Innovation (F2I2) as the funding body in part of this research.

REFERENCES

- [1] L. Brabenec, et al., “Speech disorders in Parkinson's disease: early diagnostics and effects of medication and brain stimulation”, *J. Neural Transm.*, Vol. 124, No 3, 2017, pp. 303–334.
- [2] R. F. Hink, K., Kodera, O., Yamada, K., Kaga, and J. Suzuki, “Binaural Interaction of a Beating Frequency-Following Response”, *Audiology*, Vol. 19, No. 1, 1980, pp. 36–43.
- [3] G. Oster, “Auditory beats in the brain”, *Sci. Am.*, Vol. 229, No. 4, 1973, pp. 94–102.
- [4] M. Y. Neufeld, et al., “EEG frequency analysis in demented and nondemented parkinsonian patients”, *Dementia*, Vol. 5, No 1, 1994, pp. 23–28.
- [5] G. Gálvez, et al., “Neuroacoustical Stimulation of Parkinson's Disease Patients: A Case Study”, *Proc. of the IWINAC 2019, Lecture Notes in Computer Science*, Vol. 11487, 2019, pp. 329–339.
- [6] P. Gómez, et al.: “Glottal Source biometrical signature for voice pathology detection”, *Speech Communication*, Vol. 51, 2009, pp. 759–781.
- [7] F. Viallet, A. Ghio and S. Skodda, “Vocal tremor analysis via AM-FM decomposition of empirical modes of the glottal cycle length time series”, *Proc. Interspeech*, 2015, pp. 766–770.
- [8] P. Gómez, et al., “Parkinson's disease monitoring by biomechanical instability of phonation”, *Neurocomputing*, Vol. 255, 2017, pp. 3–16.
- [9] P. Gómez, et al., “BioMet@Phon: A System to Monitor Phonation Quality in the Clinics”, *Proc. of eTELEMED 2013*, 2013, pp. 253–258.
- [10] R. Blomberg, et al., “Speech Processing Difficulties in Attention Deficit Hyperactivity Disorder”, *Frontiers in Psychology*, 05 July 2019, doi: 10.3389/fpsyg.2019.01536

A NEUROMECHANICAL MODEL OF JAW-TONGUE ARTICULATION IN PARKINSON'S DISEASE SPEECH

A. Gómez¹, A. Tsanas², P. Gómez¹, D. Palacios³, A. Álvarez¹, R. Martínez¹

¹Neuromorphic Speech Processing Laboratory, Center for Biomedical Technology, Universidad Politécnica de Madrid, Campus de Montegancedo, s/n, 28223 Pozuelo de Alarcón, Madrid, Spain

²Usher Institute of Population Health Sciences and Informatics, Medical School, University of Edinburgh, UK

³Escuela Técnica Superior de Ingeniería Informática - Universidad Rey Juan Carlos, Campus de Móstoles, Tulipán, s/n, 28933 Móstoles, Madrid, Spain

Email addresses: andres.gomez@ctb.upm.es, Athanasios.Tsanas@ed.ac.uk, pedro@fi.upm.es, daniel.palacios@urjc.es, agustin@junipera.datsi.fi.upm.es, rmolalla@fi.upm.es

Abstract: This study extends previous work modeling the jaw-tongue biomechanical system to further investigate neuromotor activity in muscular activity during certain speech gestures to model hypokinetic dysarthria in Parkinson's Disease (PD) patients. The objective of this study is to estimate the parameters of an inverse model for the characterization of neuromotor activity on the masseter from 3D accelerometry (3DAcc) and speech kinematics. Model parameter estimation is based on an iteration of initial kinematic data on the sagittal plane obtained from multiple regression with respect to formant dynamics derived from acoustical analysis. Preliminary results point to the feasibility of this methodology in populating model parameters from both healthy controls and PD patients.

Keywords: Neuromotor Dysarthria, Parkinson's Disease, Speech Formants, Neurodegenerative Diseases

I. INTRODUCTION

It is well known that Hypokinetic Dysarthria (HD) is one of the manifestations of Parkinson's Disease (PD) in the speech of affected patients [1][2][3][4]. Therefore, the modeling of HD is of great interest for the detection and monitoring of PD using speech. This study builds upon our previous work modeling the jaw-tongue biomechanical system investigating dysarthric speech to model HD in PD patients [5]. The objective of this study is to propose an inverse model for the estimation of neuromotor activity on the masseter from 3DAcc and speech kinematics using diadochokinetic exercises by healthy participants and PD patients to check if this model can be applied to PD speech. The model may be used in establishing biomarkers for monitoring PD symptom trajectory progression in treatment and rehabilitation using speech as the main vehicular trait. Estimations obtained on data from a

healthy control and four PD male patients have been used and compared.

II. FUNDAMENTALS

The study is based on a simplified jaw-tongue articulation model [6] which is known to be representative of PD dysarthria [7] relating acoustic and kinematic variables as the first two formants $F = \{F_1, F_2\}$ and the horizontal and vertical coordinates $G = \{g_{hr}, g_{vr}\}$ of a joint Jaw-Tongue Reference Point (P_{JTr}) in the sagittal plane, which may be seen as the center of moments in the biomechanical system integrated by the mandible bone, tongue and facial tissues associated [8] (see Figure 1).

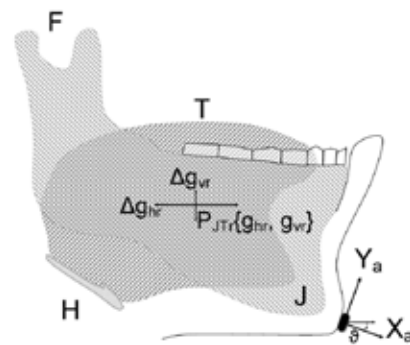


Figure 1. Jaw-tongue biomechanical model. The kinematic variables $G = \{g_{hr}, g_{vr}\}$ are the horizontal (g_{hr}) and vertical (g_{vr}) PJTr coordinates and their relative displacements with respect to the origin $\Delta G = \{\Delta g_{hr}, \Delta g_{vr}\}$. $\{X_a, Y_a\}$ are the tangential and normal components of acceleration in the sagittal plane.

The model assumes that a Linear Time-Invariant relationship (LTI) may be established between the P_{JTr} sagittal coordinates and the first two formants, which may be summarized as

$$\Delta G = W \times \Delta F; W = \{w_{ij}\}_{i=1,2}^{j=1,2} \quad (1)$$

where $\Delta\mathbf{G} = \{\Delta\mathbf{g}_{hr}, \Delta\mathbf{g}_{vr}\}$ is the vector of the horizontal ($\mathbf{g}_{hr}=\mathbf{g}_1$) and vertical ($\mathbf{g}_{vr}=\mathbf{g}_2$) P_{JT_r} displacements in the time domain, which may be obtained from the rotation and integration of the tangential and normal acceleration components $\{\mathbf{X}_a, \mathbf{Y}_a\}$ on the subject's chin [9], \mathbf{W} in (1) is a 2x2 matrix expressing the LTI, and $\Delta\mathbf{F}$ is the relative displacement with respect to the first two formant means in the time domain, as the first two formants are known to be more strongly associated with articulation kinematics than higher formants [6], respectively

$$\Delta\mathbf{F} = \{\mathbf{F}_1 - \text{mean}(\mathbf{F}_1), \mathbf{F}_2 - \text{mean}(\mathbf{F}_2)\} \quad (2)$$

The ultimate purpose of the model described in (1) is to allow the estimation of the model weight matrix \mathbf{W} , which is to be populated using individual weight values w_{ij} from healthy controls and subjects diagnosed with PD, to establish possible regression models on the kinematic components of the relative speed of the P_{JT_r} exclusively from acoustic estimates ($\Delta\mathbf{F}$) derived from the speech signal. The methodology proposed is based on solving for the model parameters \mathbf{W} using standard multivariate regression between the observed variables $\Delta\mathbf{G}$ and $\Delta\mathbf{F}$. It may be formulated as the minimization of the cost function E

$$\mathbf{W}_{est} = \underset{\mathbf{W}}{\text{argmin}}\{E\}; \quad (3)$$

$$E = (\Delta\mathbf{G} - \mathbf{W} \times \Delta\mathbf{F})(\Delta\mathbf{G} - \mathbf{W} \times \Delta\mathbf{F})'$$

where the notation ($'$) denotes matrix transposition. The estimation method is based on a first approximation of w_{i10} and w_{ij0} by simple regression to estimate these initial values, and an iteration using multiple regression to refine the joint estimates of the weights in \mathbf{W} as

$$w_{iik} = \frac{\Delta\mathbf{g}_i \Delta\mathbf{F}'_i - w_{ijk-1} \Delta\mathbf{F}_i \Delta\mathbf{F}'_j}{\Delta\mathbf{F}_i \Delta\mathbf{F}'_i} \quad (4)$$

$$w_{ijk} = \frac{\Delta\mathbf{g}_i \Delta\mathbf{F}'_j - w_{iik-1} \Delta\mathbf{F}_i \Delta\mathbf{F}'_j}{\Delta\mathbf{F}_j \Delta\mathbf{F}'_j}$$

where k is the iteration step. Practical convergence using minimum least squares is typically achieved within a few iterations.

III. MATERIALS AND METHODS

Materials: Eight male and female PD patients (16 subjects) were recruited from a PD patient association in the metropolitan area of Madrid (Asociación de pacientes de Parkinson de Alcorcón y Móstoles, APARKAM). The data used in this study are from a subset integrated by four male PD patients (stage 2 in H&Y scale) from the initial cohort of 16 patients examined, and from a healthy control male, included in the study for comparison purposes. The study was approved by the Ethical Committee of UPM

(MonParLoc, 18/06/2018). The voluntary participants signed an informed consent. The principles stated in the Declaration of Helsinki were strictly followed. Figure 2 presents an example of the process of multiple signal collection from PD patients. The equipment used in the simultaneous recording of surface electromyography (sEMG), 3DAcc, and speech is illustrated. The sEMG from the two attachments of the masseter complex to the jaw and skull, and the 3DAcc signals from an accelerometer attached to the chin were digitized and collected with a Biopac MP150 EMG100 at 2 kHz and 16 bits (SEMG). A Sennheiser cardioid wireless microphone (ew320 g2) on a MOTU Traveler MK1 sound card was used to record speech at 40 kHz and 32 bits. Speech was downsampled to 8 kHz (a compromise between accuracy and computational requirements, which allowed to extend the results to data acquired on a telephone channel over an e-Health platform) and the first two formants were estimated by adaptive lattice inverse filters [10] every 2 ms. Consequently, sEMG and 3D accelerometer signals were downsampled to 500 Hz. Recordings comprise the sustained vowels $[a:]$, $[e:]$, $[i:]$, $[o:]$ and $[u:]$, the fast repetition of the syllables $[pa:]$, $[ta:]$ and $[ka:]$ and the diphthong $[...aja...]$. In the present study only this last recording was used in the estimation of \mathbf{W} by multiple regression, as this diphthong produces the widest sweeps of formant dynamical patterns associated to the high-low and forward-backward displacement of P_{JT_r} .

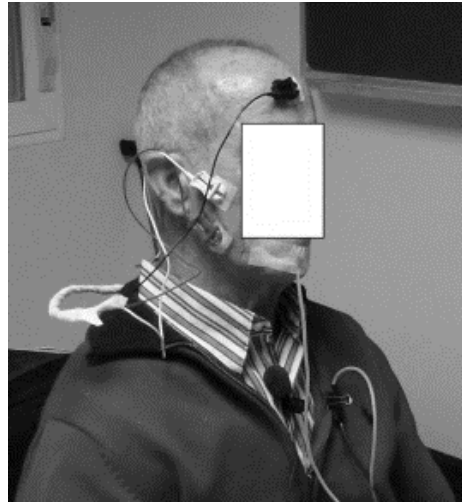


Figure 2. Signal acquisition framework. Three sEMG electrodes are fixed on the extremes of the masseter (differential pair) and on the forefront (reference). The 3D accelerometer is fixed on the chin. A cardioid clip microphone is fixed on the collar.

The speech signal, the sEMG and the 3 acceleration channels from two repetitions of the $[...aja...]$ by one of the PD patients included in the study are shown in Figure 3 as an example.

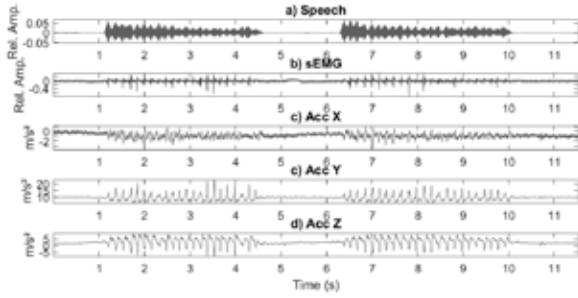


Figure 3. Signal acquisition example from the repetition of the phonetic sequence [...aja...] by a PD patient.

The repetition of [...aja...] as fast as possible has been used in the present study to estimate the model parameters summarized in matrix \mathbf{W} in kinematic terms.

Methods: The data used in this study are the first two formants derived from the speech utterance produced by the fast repetition of [...aja...] unbiased and smoothed to be compared with the jaw-tongue reference displacements obtained after rotation and integration of the acceleration signals [9]. The alignment among formant trajectories and displacements in the case referred in Figure 3 may be seen in Figure 4.

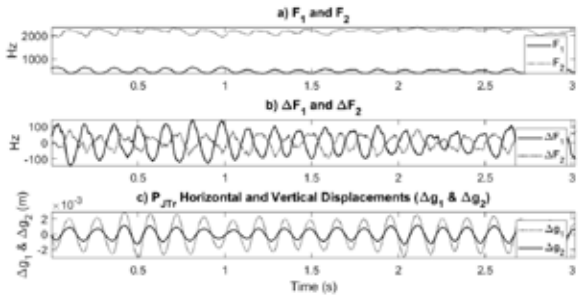


Figure 4. Formant deviations and reference point displacements obtained from the data in Figure 3, corresponding to a PD patient: a) formants F1 and F2; b) formant deviations ΔF ; c) reference point displacements ΔG .

It may be said that the estimations of formant deviations and reference displacements ($\Delta \mathbf{F}$ and $\Delta \mathbf{G}$) from PD patients (PD1-4), compared to the healthy control (HC) are smaller in range. $\Delta \mathbf{F}$ and $\Delta \mathbf{G}$ are used in estimating the initial values of the model weights w_{ij0} in (4). The model weights w_{ij} after 20 iterations of (4) are given in Table 1. Spearman coefficients R_{Sij} associated to each weight estimate (p-values < 0.001) are also given.

Table 1. Model weights and correlation coefficients (R: Spearman coefficient, p-value < 0.001); $\times 10^{-4}$ cm/Hz

Subject	HC	PD1	PD2	PD3	PD4
Age (y)	28	73	71	76	70
w_{11}	-0.171*	-0.065*	-0.197*	-0.088*	-0.025*
w_{12}	0.083*	0.029*	0.336*	0.050*	0.031*
w_{21}	-0.338*	-0.227*	-0.075*	-0.129*	-0.053*
w_{22}	0.238*	0.101*	0.130*	0.079*	0.067*
R_{S11}	-0.796	-0.76	-0.832	-0.750	-0.578
R_{S12}	0.506	0.674	0.820	0.583	0.713
R_{S21}	-0.840	-0.765	-0.796	-0.769	-0.637
R_{S22}	0.777	0.675	0.791	0.585	0.775

IV. RESULTS AND DISCUSSION

A brief review of the results in terms of w_{21} and w_{22} shows that the weights obtained for the control subject (HC) are the largest in absolute value. A large weight magnitude means that small sweeps in formants are associated with large displacements in the reference point, otherwise, small displacements in the reference point are required to produce large sweeps in formants if the weight magnitudes are small. In this case, it may be hypothesized that the effective oral cavity is also reduced, therefore a small change in its cross section could produce a substantial change in the formants. This would be the case of PD4 compared to the other speakers. A very interesting analysis comes from the normalized weights for each speaker (by columns), i.e., when subtracting the averages and dividing by the standard deviation. These normalizations are shown in Table 2.

Table 2. Normalized model weights per speaker

Subject	HC	PD1	PD2	PD3	PD4
w_{11}	-0.484	-0.173	-1.046	-0.648	-0.560
w_{12}	0.508	0.490	1.227	0.705	0.489
w_{21}	-1.132	-1.316	-0.529	-1.049	-1.070
w_{22}	1.108	0.999	0.347	0.992	1.141

In this case it may be seen that in four of the cases (HC, PD1, PD3 and PD4) the absolute values of the normalized weights w_{11} and w_{12} are considerably smaller than w_{21} and w_{22} , stressing that the largest relationship is to be found between the vertical displacement and the first two formants, contrary to what happens in the case of PD2, where it seems that the horizontal displacement is more strongly related to the first two formants than the vertical displacement. Analyzing also the lower part of Table 1, it seems that the large values of the correlation coefficient shown by

the healthy control and PD patients' between kinematic and acoustic correlates indicate that some statistical association is evident. This relationship points towards developing causative interpretable human voice production models explaining differences in PD vs healthy controls in a further study. The largest correlation values correspond mainly to the weights w_{11} and w_{21} , expressing the influence of the horizontal and vertical displacements $\Delta \mathbf{g}_1$ and $\Delta \mathbf{g}_2$ on the first formant variation ΔF_1 . On the other hand, inter-speaker variability observed in the biomechanical model may suggest that the matrix \mathbf{W} is a potential biomarker to be taken into account in developing future neuromarkers based on acoustical analysis of speech signals. Another pending study is the extension of the proposed first-principle biomechanical model of voice production, using high-quality speech signals sampled at larger sampling rates.

V. CONCLUSIONS

In the present study a LTI model relating kinematic and acoustic variables from a healthy control and PD patients has been hypothesized as a possible description of formant-to-displacement kinematics. The feasibility and reliability of the model have been assessed and estimated. The relationship between both types of variables (displacements and formants) is well supported by the computed correlation coefficients. Although the small number of speakers points to tentative interpretation and generalization of the results, it seems that these preliminary findings may support the feasibility of using these models in signal inversion to produce reliable estimations of kinematic correlates from acoustic recordings both in healthy controls as well as in PD patients. In this sense we aim to extend current analysis with this first-principles mechanistic model on larger databases such as the Parkinson's Voice Initiative (PVI) [11] data to explore how findings generalize and gain potentially new insights into PD speech.

ACKNOWLEDGMENTS

This research has been funded by grants TEC2016-77791-C4-4-R (MINECO, Spain) and CENIE_TECA-PARK_55_02 INTERREG V-A Spain – Portugal (POCTEP). The authors thank Asociación de Parkinson de Alcorcón y Móstoles (APARKAM), and Azucena Balandín (director) and Zoraida Romero (speech therapist) for their help and advice.

REFERENCES

[1] S. Skodda, W. Grönheit, N. Mancinelli and U. Schlegel, "Progression of Voice and Speech

- Impairment in the Course of Parkinson's Disease: A Longitudinal Study", *Parkinson's Disease*; Article ID 389195, 2013.
- [2] S. Sapir, "Multiple Factors Are Involved in the Dysarthria Associated With Parkinson's Disease: A Review With Implications for Clinical Practice and Research", *J. Speech, Lang. and Hear. Res.*, Vol. 57, 2014, pp. 1330-1343.
- [3] J. Ruzs. et al., "Imprecise vowel articulation as a potential early marker of Parkinson's disease: effect of speaking task", *J. Acoust. Soc. Am.*, Vol. 134, 2013, pp. 2171–2181.
- [4] J. Mekyska et al., "Robust and complex approach of pathological speech signal analysis", *Neurocomputing*, Vol. 167, 2015, pp. 94-111.
- [5] P. Gómez et al., "Articulation Dynamics in Parkinson Dysarthria", *Proc. of MAVEBA 17*, Firenze University Press, December 13-15, 2017, pp. 81-84.
- [6] C. Dromey, G. O. Jang. and K Hollis, "Assessing correlations between lingual movements and formants", *Speech Comm.*, Vol. 55, 2013, pp. 315-328.
- [7] J. A. Whitfield and A. M. Goberman, "Articulatory-acoustic vowel space: Application to clear speech in individuals with Parkinson's disease", *J. Comm. Disord.*, Vol. 51, 2014, pp. 19-28.
- [8] P. Gómez et al. "Characterization of Parkinson's disease dysarthria in terms of speech articulation kinematics", *Biomed. Signal Proc. and Control* Vol. 52, 2019, pp. 312-320.
- [9] P. Gómez et al. "Neuromechanical Modelling of Articulatory Movements from Surface Electromyography and Speech Formants", *International Journal on Neural Systems*, Vol. 29, No. 2, Article ID: 1850039, 2019.
- [10] J. R. Deller, J. G. Proakis and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. Macmillan, NewYork, 1993.
- [11] S. Arora, L. Baghai-Ravary and A. Tsanas, "Developing a large scale population screening tool for the assessment of Parkinson's disease using telephone-quality speech", *Journal of Acoust. Soc. Am.*, Vol. 145, No 5, 2019, pp. 2871-2884.

PARKINSON'S DISEASE CLASSIFICATION BASED ON VOWEL SOUND

D. Hemmerling¹, D. Sztahó²

¹ AGH University of Science and Technology, Department of Measurement and Electronics, Al. Mickiewicza 30, 30-059 Krakow, Poland

² Laboratory of Speech Acoustics, Budapest University of Technology and Economics, Department of Telecommunications and Media Informatics, Magyar tudósok krt. 2, 1117 Budapest, Hungary
hemmer@agh.edu.pl, sztaho@tmit.bme.hu

Abstract: Parkinson's disease (PD) is the second most frequent neurodegenerative disorder that causes decisive deterioration of the quality of life through severe motor and cognitive dysfunctions. In this study we present the usage of vowel signals to accurately detect PD by using machine learning methods. The data set consists in total of 198 recordings of vowel /a/ phonated in sustained manner, where 50% of data was assigned as a representation of Parkinson's disease state. The voice signals are described by the set of features extracted from time, frequency and cepstral domains applied to Principal Component Analysis (PCA) and nonlinear Support Vector Machine (SVM) to distinguish between PD patients and healthy control group. The results ensure 93.43% of classification accuracy.

Keywords: Parkinson's disease, voice analysis, pathological speech, voice signal analysis

I. INTRODUCTION

Up to 89% of all patients with Parkinson's disease (PD) reveal speech disorders [1]. Voice disorders, including speech volume reduction, problems of articulation and fluency are often one of the first symptoms. Speech disorders in people with PD appear due to: laryngeal function deficit, impaired performance of facial muscles, decreased vital capacity of the lungs and decreased speech drive. These changes cause many abnormalities in voice and speech signals such as: volume reduction, lowering the tone of the voice, limited modulation (monotonous speech), difficulty with loudness changes, reduction of vocal fold tension, rough and hoarse tone, inappropriate articulation (speech becomes indistinct) and change of pace of speech [2,3]. These impairments are called hypokinetic dysarthrias. Dysarthric speech is characterized by dysfunctions of phonation, articulation and prosody, which arose as a result of damage to the centers and nerve pathways responsible for the innervation of the speech organs [4]. Phonation is defined as the vibration of vocal cords, which leads to sound generation. Articulation combines the modification of position, tension and tongue arrangement involved in the

production of speech. From a clinical point of view, phonation problems are associated with incorrect vocal fold movement and incomplete closing of the vocal chords [5]. Changes in articulation are caused by reduced amplitude and speed of movements: lips, jaws and tongue. This leads to reduced accentuation, inaccurate articulation of consonants to babbling. Prosody is a sonorous speech property that takes into account the intonation, volume, accent and duration of the phoneme [6]. Abnormalities in prosody are manifested by speaking with short, accelerated phrases, monotony and limited speech volume, change of speech speed, pauses, difficulty in expressing emotions, repeating sounds or syllables [7].

Many researchers and physicians consider automated acoustic analysis as a useful non-invasive, fast in implementation tool for PD detection. Researchers implemented various speech features to create an automatic computer-aided diagnosis system to detect Parkinson's disease [8-10]. The most frequently used features include the phonation parameters, such as jitter and shimmer parameters, noise parameters [8-11]. The paper [12] describes the usage of phonation, articulation and prosody features applied to sustained vowels, vowels uttered with modulated tone, different words, phonemes and speech tasks. The dataset consists of 100 patients. The classification accuracy is 91.3% obtained for vowel /a/. The acoustic features describing the signals are applied to various classification tasks to differentiate between healthy control group and Parkinson's disease. For classification, different machine learning methods are described to detect PD such as Deep Neural Networks, Support Vector Machines, Naive Bayes, Decision Trees, Regression and Rotation Forest [9,10,13,14]. The paper [14] presents a usage of Parallel Distributed Neural Network classifier with backpropagation learning algorithm and majority voting scheme technique in its design. The detection accuracy of PD is up to 90%. The authors of [9] compare different classifiers to distinguish between healthy patients and those suffering from PD. The classifiers used in the research are: Neural Network, Decision Trees, Regression and DMneural algorithm. The highest classification rate was achieved for Neural Network,

which is 92.9%. The major problem in the classification tasks in a medical application is the size of implemented database. Mostly, the authors implement to their algorithms very small amount of data which is around 60 voice recordings. The study of [15] show classification accuracy equal to 98.6% when vowel /a/ is used. The database consisted of 33 PD patients and 10 healthy people. The paper [16] presents 92% of detection accuracy with vowel /e/ phonated by 20 persons with PD and 20 healthy ones. The database described in the paper [17] was set up from 3 languages: 170 German speakers, 100 Spanish speakers and 35 Czech speakers. For classification the authors used a read text and calculated energy content in the transitions between voiced and unvoiced segments, which enabled classification accuracy at the level of 91-98%, depending on the language. The same database, as presented in this paper, was used in publication [18]. For signal analysis the authors used measures of: voice quality, pitch, intensity, articulation, nonlinear measures, prosodic measures and the age. The best results were obtained using support vector machines with 89.3% accuracy.

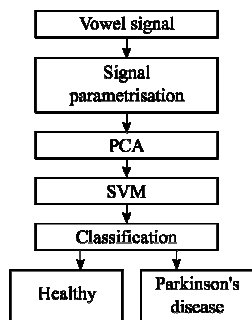


Fig.1 Diagram presenting classification process of patients with Parkinson's disease and healthy persons.

II DATABASE

A Hungarian speech database was created and used that contains speech samples from 83 PD patients. The audio footage was done in two health institutes in Budapest: Virányos Clinic and Semmelweis University. For healthy control (HC) population, speech from 33 subjects were also recorded. Signed consent form from each subject was collected to record his or her voice. For the experiment USB sound card (Terratec 6fire USB) with good quality A/D converter and low noise level (audio coding: PCM, sampling rate: 44.1 kHz, quantization: 16-bit) was used as external recording device along with a clip-on condenser microphone (Audio-Technika ATR3350). The recordings were created in quiet office environment (medical office). The text material consisted of a vowel /a/ recording. The examination was repeated 3 times for each of the patients. For our research, we have used 99 recordings from 33 PD

patients and 99 recordings from 33 healthy control speakers. In total, we have used 198 recordings from 66 patients. Speech from speakers of healthy control population (HC) (with same age distribution) has been also recorded using the same text material and recording environment. In order to balance the H-Y distribution of the database, 33 healthy subjects (17 male speaker with mean age of 69.06 ± 8.53) and 16 female speakers with mean age of $59.31 (\pm 7.67)$) were recorded labeled with 0 H-Y score. The HC subjects had no known diseases and were not under any medical treatment.

III. METHODS

Diagnostics of the vocal tract can be performed based on selected voices, allophones, phonemes and vowels. Vowels are the group most often used in medical diagnostics due to the steady state that occurs during their phonation, enabling determination of stable acoustic signal parameters over time. In order to obtain information important from a diagnostic point of view, it was necessary to select such features of the signal and its frequency and mel-cepstral spectrum, which determine the manner of its perception by the recipient, thus allowing to conduct a diagnostic, control or prognostic process. The choice of diagnostic features describing the speech signal consisted in finding such features, using which, it was possible to differentiate between pathological and healthy cases. Individual features may vary in unit of measure and have a different range of variation. In case of many classification algorithms, this can have a significant impact on the final result. Therefore, it is recommended to pre-process data including their scaling and normalization, that is, bringing the characteristics of the signal to a dimensionless form with a standardized range of variation. The normalization was done in the range between 0 and 1 for each acoustic parameter.

In order to quantify the process of phonation, its parametric description is necessary. The following mean values of the parameters of this process were considered: fundamental frequency, jitter and shimmer coefficient, five-point and eleven-point disturbance coefficient, five-point amplitude disturbance coefficient, energy, zero, first, second and third spectral moment, kurtosis, relative power factor, amplitude and frequency 1-, 2-, 3- control, maximum value, 12 mel-cepstral coefficients, harmonic to noise ratio. These parameters describe the periodicity, articulation, noise content and non-linearity of the phonation process.

The vector describing the vowel signal was set up from many parameters. For feature extraction we used PCA. This method enables orthonormally data transformation so that the variance of the data remains

constant, but is concentrated in the lower dimensions. As a result, we obtained a set of principal components, whereas the first principal component covered the highest variance of the data, and the last principal component covered the lowest variance of the data. The data had been reduced and the principal components values extracted, we have experimentally chosen 19 out of 33 parameters without any loss of information (100% of variation was held).

IV. RESULTS

One of the most popular methods of machine learning is the support vector machine (SVM). One of the main stages in the implementation of the SVM is to find a hyperplane in a multidimensional space that optimally separates points belonging to different classes with the largest possible margin of confidence. The larger the margin, the greater the distance to the nearest point of the given class on each side of the hyperplane. The goal is to find the maximum margin and achieve the highest class separation efficiency. In medical applications, it is impossible to separate the data in linear manner. In such situations SVM can use kernel trick to transform the data and construct separating hyperplane in new space. In this paper, we implement kernel-SVM formulation with a cubic polynomial kernel functions.

To evaluate the classification, we used 10-fold cross-validation in which we randomly selected 90 % of the data for learning and the remaining 10% of data for testing the algorithm. The recordings of the same speaker were never assigned to the train and test set simultaneously. The procedure was iterated 10 times. The results from cross-validation are summarized in confusion matrix in Tab. 1.

Tab. 1 Classification results calculated using confusion matrix

	Sensitivity [%]	Precision [%]	F1-score
HC	94.79	91.92	0.93
PD	92.16	94.95	0.94
Mean value	93.48	93.44	0.93
Accuracy [%]		93.43	
MCC		0.87	
AUC		0.97	

To assess the quality of the classification process we calculated accuracy, sensitivity, precision, F1 score, Matthews correlation coefficient (MCC) and AUC. The accuracy explains what percentage of parameters from the test set were correctly assigned to their respective classes and obtained at the stage of testing. The sensitivity meant the test's ability to identify

positive results. The precision is defined as ratio of the number of results classified correctly to given group to all classified by the system to the same group.

The F1 score is defined as the harmonic average of the precision and recall. The F1 score can take values from 0 and 1 range, where best value is at 1 meaning perfect precision and recall and the worst at 0.

MCC is a correlation coefficient between target and classification results. It returns a value between -1 and +1. -1 appear when there is perfect disagreement between target and classification results. Value of 1 means a perfect agreement between target and classification results. MCC involves values of all the four quadrants of a confusion matrix, that is why it is considered as a balanced measure.

The results are also presented using the receiver operating characteristic (ROC) curve.

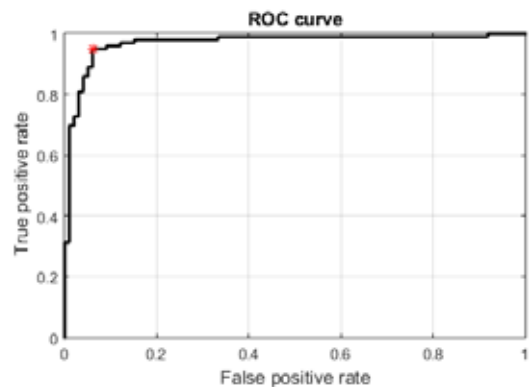


Fig. 2 ROC curve obtained from the classification, True positive class is represented by the healthy control group, false positive rate. AUC=0.97.

The horizontal axis of this curve describes the false positive rate, which means the rate of samples erroneously detected as pathological, and the vertical axis describes the true positive rate, which means the rate of samples correctly classified as pathological. To measure the performance of binary classification system we calculated the area under the ROC curve (AUC). AUC ranges from 0 to 1, where 1 means perfect classification. The ROC curve is presented in Fig 2.

For the experiment, recordings of vowel /a/ were used. Based on presented results, the system more accurately classified patients with PD rather than healthy patients. The precision for PD is 94.95% and for healthy group is 91.92%. The accuracy of the presented system is 93.43%. The MCC value is 0.87%, whereas AUC is 0.97%.

The data reduction to 19 parameters had no influence on classification accuracy. It means that the PCA enabled the reduction of the number of parameters and ensured the same quality as original feature vector.

V. DISCUSSION

In this paper we present an integrated, automatic acoustical analysis of deformed pathological speech in Parkinson's disease. The analysis was performed among patients and showed that speech impairments caused by PD can be computed and detected using acoustical methods. In order to accelerate and simplify the calculations, we used Principal Component Analysis, that enabled parameters reduction from 33 to 19 without any loss of classification accuracy. The results showed that the calculated parameters and their reduction made it possible to distinguish acoustic features for pathological and healthy speech. We used different methods for validating a created model that had a confirmation of health status or appearance of Parkinson's disease. The obtained results show that the PCA and SVM classifier achieved the classification accuracy 93,43% using only vowel /a/ at sustained phonation. The phonation of vowels is easy, fast and recurrent. This is a huge advantage, because the patients with Parkinson's disease are in different conditions and sometimes it might be difficult for them to read any text or to repeat different words. The usage of vowels might be very useful in the context of remote monitoring of PD patients. This result is the highest classification accuracy achieved for presented database. The results are promising and we expect that developed tool will be a great help to diagnostic health care in neurological disease as Parkinson's disease.

VI. ACKNOWLEDGEMENTS

The audio samples were recorded at Virányos Clinic (under the supervision of Dr. István Valálik) and Neurology Department of Semmelweis University (head neurologist: Dr. Annamária Takáts), Budapest.

The research has been partly funded by project no. K128568, that has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the K_18 funding scheme. The research was partly funded by the CELSA (CELSA/18/027) project titled: "Models of Pathological Speech for Diagnosis and Speech Recognition".

REFERENCES

[1] L. Ramig, C. Fox, and S. Sapir, "Speech treatment for parkinsons disease," *Expert Review of Neurotherapeutics*, vol. 8, no. 2, pp. 297–309, 2008.

[2] I. Ramig, C. Fox, and S. Sapir, "Speech disorders in parkinson's disease and the effects of pharmacological, surgical and speech treatment with emphasis on lee silverman voice treatment," *Handbook of clinical neurology*, vol. 83, pp. 385–399, 2007.

[3] J. A. Logemann, H. B. Fisher, B. Boshes, and R. Blonsky, "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of parkinson patients," *Journal of Speech and hearing Disorders*, vol. 43, no. 1, pp. 47–57, 1978.

[4] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," *Journal of speech and hearing research*, vol. 12, no. 2, pp. 246–269, 1969.

[5] K. S. Perez, L. O. Ramig, M. E. Smith, and C. Dromey, "The Parkinson larynx: tremor and videostroboscopic findings," *Journal of Voice*, vol. 10, no. 4, pp. 354–361, 1996.

[6] S. Skodda, W. Visser, and U. Schlegel, "Vowel articulation in parkinson's disease," *Journal of voice*, vol. 25, no. 4, pp. 467–472, 2011.

[7] J. Ruzs, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated parkinsons disease," *The journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 350–367, 2011.

[8] M. Shahbakhi, D. T. Far, E. Tahami et al., "Speech analysis for diagnosis of parkinsons disease using genetic algorithm and support vector machine," *Journal of Biomedical Science and Engineering*, vol. 7, no. 4, pp. 147–156, 2014.

[9] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for highaccuracy classification of parkinson's disease," *IEEE Transactions on biomedical engineering*, vol. 59, no. 5, pp. 1264–1271, 2012.

[10] H. Guruler, "A novel diagnosis system for parkinsons disease using complex-valued artificial neural network with k-means clustering feature weighting method," *Neural Computing and Applications*, vol. 28, no. 7, pp. 1657–1666, 2017.

[11] J. Ruzs, R. Cmejla, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, and E. Ruzicka, "Acoustic markers of speech degradation in early untreated parkinsons disease," vol. 61, no. 12, pp. 58–08, 2011.

[12] M. Cernak, J. R. Orozco-Arroyave, F. Rudzicz, H. Christensen, J. C. Vasquez-Correa, and E. Noth, "Characterisation of voice quality of parkinsons disease using differential phonological posterior features," *Computer Speech & Language*, vol. 46, pp. 196–208, 2017.

[13] L. A. Passos, C. R. Pereira, E. R. Rezende, T. J. Carvalho, S. A. Weber, C. Hook, and J. P. Papa, "Parkinson disease identification using residual networks and optimum-path forest," in *2018 IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI)*. IEEE, 2018, pp. 000 325–000 330.

[14] M. Can, "Neural networks to diagnose the parkinsons disease," *SouthEast Europe Journal of Soft Computing*, vol. 2, no. 1, 2013.

[15] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for highaccuracy classification of parkinson's disease," *IEEE Transactions on biomedical engineering*, vol. 59, no. 5, pp. 1264–1271, 2012.

[16] E. A. Belalcazar-Bolanos, J. R. Orozco-Arroyave, J. Vargas-Bonilla, J. D. Arias-Londono, C. G. Castellanos-Dominguez, and E. Noth, "New cues in low-frequency of speech for automatic detection of parkinsons disease," in *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Springer, 2013, pp. 283–292.

[17] J. R. Orozco-Arroyave, F. H'onig, J. D. Arias-Londo'no, J. Vargas-Bonilla, S. Skodda, J. Ruzs, and E. N'oth, "Voiced/unvoiced transitions in speech as a potential bio-marker to detect parkinson's disease," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[18] D. Sztah, I. Vallik, and K. Vicsi, "New hungarian database for analysis of speech of people suffering from parkinsons disease" *DOGS 2017*. At Novi Sad, Serbia, 2017.

ANALYSIS OF PHONATORY FEATURES FOR THE AUTOMATIC DETECTION OF PARKINSON'S DISEASE IN TWO DIFFERENT CORPORA

Laureano Moro-Velázquez^{1,*}, Jorge A. Gómez-García^{2,†}, Najim Dehak¹, Juan I. Godino-Llorente^{2,‡}

¹ Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA

² Escuela Técnica Superior de Ingeniería y Sistemas de Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain
{laureano, ndehak3}@jhu.edu, {jorge.gomez.garcia, ignacio.godino}@upm.es

Abstract: In this study, several automatic detectors of Parkinson's Disease (PD) based on phonatory aspects were analyzed employing two different corpora containing speech from speakers with PD. The features employed to characterize phonation were jitter, shimmer, noise measurements, complexity, modulation spectrum features and perceptual linear predictive coefficients. To differentiate between speakers with and without PD a gaussian mixture model classification scheme was used. Then, the approach providing the best results was combined with a scheme using articulatory information of the speech in order to assess the complementarity between phonatory and articulatory aspects in the automatic detection of PD. Cross-validation trials (k-folds) employing exclusively phonatory information provided accuracies between 64% and 71%, with AUC between 0.68 and 0.80 depending on the corpus. Results suggest that a combination of all the analyzed features with a PCA dimensionality reduction produce the best accuracy, AUC and sensibility. Also, results indicate that phonatory approaches tend to be less accurate in PD detection than other articulatory approaches proposed in previous studies. Finally, results suggest the complementarity between the studied articulatory and phonatory approaches is low.

Keywords: Parkinson's Disease, Modulation Spectrum, GMM, Complexity, Noise.

I. INTRODUCTION

Within all the works found in literature assessing phonatory aspects of the voice of speakers with Parkinson's Disease (PD), the most recent use phonatory features as the input of automatic detectors to distinguish between the voices of people with and without PD by means of machine learning techniques.

In one of the first studies [1], authors employ dysphonic features including jitter, shimmer, noise, complexity measurements and Mel-Frequency Cepstrum Coefficients (MFCC) to train a binary classifier. An accuracy of 98.6% was reached in the binary detection using only 10 features obtained after feature selection (8 MFCC values and two noise features). Although authors followed a k-folds cross-

validation scheme with 10 folds, the use of the same speakers in training and testing partitions during cross-validation could be providing over-optimistic results.

A subsequent work [2] analyzes the detection of PD using five vowels phonated at different levels characterized by means of jitter, shimmer, noise or complexity measures among others. After a feature selection and employing random forests, a 90% accuracy (employing cross-validation) was obtained.

Other works provide different detection accuracies ranging between 76% and 91% employing similar features and classification methods [3]–[5]. However, none of the studies compare the features separately and using several corpora in the same study.

The present study explores the properties of multiple characterization methods combined with a classification scheme providing automatic detectors of PD using sustained phonation (vowel /ah:/). In particular, five types of characterizations were used: noise, amplitude and frequency perturbations (jitter and shimmer), complexity measurements, Perceptual Linear Predictive (PLP) coefficients and Modulation Spectrum (MS) features. All of these features measure the irregularities caused by the symmetric and non-symmetric rigidity of the muscles controlling the vocal folds, incoordination of movements in the phonatory system and tremor, common in speakers with PD [6].

The main purpose of this study is to compare the efficiency of these features in the automatic detection of PD with two different corpora and how these detectors based on phonatory aspects can complement other articulatory-oriented approaches.

II. MATERIALS AND METHODS

Four different approaches were analyzed in this study, attending to the type of features used to characterize the sustained phonation of the speakers. Then, the method providing the best results was combined with a scheme described in a previous study [7] in order to evaluate the complementarity between phonatory and articulatory aspects for PD detection.

A. Materials

Two different subsets containing speech from patients suffering from PD and controls were employed in this work. The first subset was extracted from the Neurovoz corpus [8] and contains a sustained vowel /ah:/ and six

* orcid.org/0000-0002-3033-7005; † orcid.org/0000-0002-6060-387X

‡ orcid.org/0000-0001-7348-3291

Study funded by grant DPI2017-83405-R from Government of Spain and MITS Global Seed Funds.

Text-Dependent Utterances (TDU) from 47 parkinsonian and 32 control speakers whose mother tongue is Spanish Castillian. The second one, containing a sustained vowel /ah:/ and six TDU, was extracted from the GITA corpus [9] composed by 50 patients with PD and 50 age- and sex-matched control speakers whose native language is Spanish Colombian.

Additionally, Albayzin corpus [10] was used to create an Universal Background Model (UBM) in the experiments using the articulatory aspect of speech.

B. Methods

Each of the four phonatory approaches employed a different type of features, while all of them used the sustained vowel /ah:/ as input. In the front-end the voice signals, sampled at 44.1 kHz, were normalized and windowed employing a Hamming window. After obtaining the correspondent features, the classification technique employed in the four phonatory approaches was Gaussian Mixture Model (GMM), with number of gaussians, G , varying in powers of 2 within the range [4, ..., 256]. A k-folds cross validation scheme was applied with $k=11$. The decision threshold during scoring was obtained using the Equal Error Rate (EER) point from the training folds.

First approach. Noise, jitter, shimmer and complexity:

In the first approach, noise, jitter, shimmer and complexity measures were used as input features of the detector. Jitter measurements are aimed to characterize the frequency perturbations of the glottal pulse [11], [12] while, shimmer quantifies the differences between the amplitudes of the peaks of the different glottal cycles within a signal. In this study, the jitter relative to the average fundamental frequency, $\%jitt$ and the shimmer relative to the average amplitude of the signal, $\%shim$, were employed.

On the other hand, Noise measurements quantify the additive noise present in the phonatory signal which can be caused by an incomplete closure of the vocal folds and uneven movements of these, directly related with the rigidity in the muscles involved in phonation, caused by PD. In this approach, the following noise features were used: Normalized Noise Energy (NNE) [13], Harmonics to Noise Ratio (HNR) [14] and Glottal to Noise Excitation ratio (GNE) [15].

In the same sense, complexity features were employed to characterize non-linearity behaviors in the vocal folds as well as the dynamics of the phonatory system. The features used in this approach were Correlation Dimension (D_2), Largest Lyapunov Exponent (LLE) [16], Hurst Exponent (HE) [17], [18], Detrended Fluctuation Analysis (DFA) [17], [18], Recurrence Period Density Entropy (RPDE) [17], Approximate Entropy (ApEn) [19], Sample Entropy (SaEn) [20], modified Sample Entropy (mSaEn) [21], Gaussian Kernel Entropy (GKE) [22], Fuzzy Entropy (FuzzyEn) [23] and Permutation Entropy (PE) [24].

A more detailed explanation about the calculation of the features in this approach can be found in [25].

The resulting feature vectors allowed to quantify the irregular vibration of the vocal folds that can be a reflection of non-symmetric movements as a consequence of rigidity, des-coordination or tremor in the larynx muscles due to PD.

To compute the noise and complexity features, the employed frame length was 55 ms as recommended in [26] with an overlapping of 50%. For $\%jitt$ and $\%shim$, all the cycle peak points were calculated for the whole signal and the local values were computed for the correspondent 55 ms frames with 50% overlapping.

Second Approach. Modulation Spectrum features:

In the second approach, the MS features proposed in [27], [28] were utilized in the front-end of the PD detector in order to characterize the voice signal. MS provides information about the energy at modulation frequencies that can be found in the carriers of a signal. Some representative features can be extracted from a MS to feed a further classification stage. The features employed in this approach were: Low Modulation Ratio (LMR), Cumulative Intersection Level (CIL), Ratio Above Linear Average (RALA), MS Percentiles (MSP), Contrast (MSW) and Homogeneity (MSH).

To compute all the MS features, signals were resampled to 25 kHz (after filtering and downsampling the original signal), and then windowed with frame length 180 ms and overlapping of 50%, as in [28].

Third approach. Rasta PLP:

Rasta-PLP coefficients were considered in this study too due to they provide Area Under the Curve (AUC) values over 0.8 in a previous study [5] employing sustained vowels. To compute the coefficients, the signal was filtered and downsampled to 16 kHz and then windowed, with frame length of 10 ms and 50% overlapping. The values used for the number of coefficients, N , were 10 and 12. This configuration was the one providing the best results in [5].

Fourth approach. All features:

In the fourth approach, all of the features included in the first, second and third approach are appended in order to create a single feature vector per frame containing all the features. All of them were calculated as described previously, changing only the frame shift that was adjusted to 27.5 ms in all cases in order to obtain the same number of frames per type of feature and recording. In this approach two techniques were utilized separately to reduce the feature vector dimension before the classification stage. The first one consisted in a dimensionality reduction technique employing Principal Components Analysis (PCA). The PCA transformation vectors were generated for each fold using only the training data and were applied to the testing data. The second one was a feature selection technique: Maximum Information Maximization (MIM) [29]. The number of features obtained after applying this technique were 6, 10 and 14.

Combination with an articulatory approach:

To end, the scores per speaker from the approach providing the best results in the phonatory experiments were fused with the scores per speaker from the articulatory scheme described in [7] by means of logistic regression. Basically, the articulatory scheme consisted in the allophonic distillation, Rasta-PLP characterization and subsequent training of a UBM employing Albayzin. Then, the UBM was adapted by the Rasta-PLP coefficients from a corpus containing parkinsonian speech (Neurovoz or GITA) by means of Maximum a Posteriori (MAP) adaptation. The speech tasks for the adaptation where TDU, the allophonic distillation type for Albayzin was plosive when using GITA for adaptation and fricative when employing Neurovoz since those provided the best results in [25]. The combination of the articulatory scores per speaker with those obtained in phonatory experiments was aimed to take advantage of the complementarity of the phonatory and articulatory aspects for PD detection.

III. RESULTS

Tables 1 and 2 include the results of the four phonatory approaches for GITA and Neurovoz respectively. Best results are marked in bold.

Table 3 contains the results employing the articulatory scheme described in [7] while Table 4 includes the results of the combination of the scores per speaker from that scheme with the scores from the phonatory approach combining all the features (fourth approach) with PCA dimensionality reduction to 14 features. Fig. 1 shows the ROCCH-DET curves of this combination. All accuracy results include Confidence Interval (C. I.) calculated as indicated in [30].

IV. DISCUSSION AND CONCLUSIONS

In this study, several approaches to automatically detect PD using a sustained vowel were analyzed. Different features previously analyzed in literature such as noise, jitter, shimmer, complexity and MS features along with PLP coefficients were analyzed employing a GMM classification scheme. The MS features provided better AUC than the rest of the features in the Neurovoz corpus (0.66 vs 0.61 and 0.63), however these results are not reproduced in the GITA corpus (0.75 vs 0.76 and 0.81). The combination of noise, jitter, shimmer and complexity provides the best accuracy results in GITA although Rasta-PLP and the combination of all features provide better AUC. In general, results are dissimilar in both corpora and only the combination of all features with PCA dimensionality reduction is within the better results in terms of accuracy and AUC in both corpora. This analysis suggest that none of these measurements in combination with GMM can provide a high accuracy and reliability in the detection of PD.

Table 1. Best results in the four approaches for GITA.

Features	Accuracy ± C. I.	AUC	Sens.	Spec.	G
Noise, %Jitt, %Shim, Complexity	74±9	0.76	0.8	0.68	4
MS	69±9	0.75	0.74	0.64	4
Rasta- PLP+Δ+ΔΔ	72±9	0.81	0.72	0.72	8
All features +PCA	71±9	0.8	0.72	0.7	16
All features +MIM	74±9	0.75	0.76	0.72	16

Table 2. Best results in the four approaches for Neurovoz.

Features	Accuracy ± C. I.	AUC	Sens.	Spec.	G
Noise, %Jitt, %Shim, Complexity	63±11	0.61	0.67	0.58	4
MS	59±11	0.66	0.69	0.45	4
Rasta- PLP+Δ+ΔΔ	64±10	0.63	0.75	0.48	4
All features +PCA	64±10	0.68	0.75	0.48	16
All features +MIM	60±11	0.64	0.63	0.58	4

Table 3. Results of articulatory approach in GITA and Neurovoz (no combination with phonatory approach).

Corpus	Accuracy ± C. I.	AUC	Sens.	Spec.
GITA	85±7	0.91	0.82	0.88
Neurovoz	89±7	0.93	0.87	0.91

The results obtained in this study indicate that the phonatory approaches analyzed provide less differentiation properties than others based in articulatory aspects such as those presented in [5], [7]. This conclusion is supported by the fact that some phonatory perturbations or dysphonia are found too in elder neurologically healthy subjects. In the same sense, a signal with a relatively simple structure as a sustained vowel, contains much less information about motor problems than other more complex signals such as those obtained with running speech. This is directly related with the lower accuracy and AUC obtained in the phonatory approaches respect to those from articulatory approaches such as the one used in this study (Table 3).

Finally, the combination of the articulatory and phonatory approaches provide a small increment of the AUC in the Neurovoz corpus. The results in the GITA corpus after the combination remain with the same values obtained using only the articulatory approach,

indicating small to none complementarity between the two schemes. This low complementarity can be justified by the fact that many articulatory approaches, like the one used in this study and in [7], could be already capturing phonatory information too.

Table 4. Results of combination of phonatory and articulatory approaches in GITA and Neurovoz.

Corpus	Accuracy \pm C. I.	AUC	Sens.	Spec.
GITA	85 \pm 7	0.91	0.82	0.88
Neurovoz	87 \pm 7	0.94	0.85	0.91

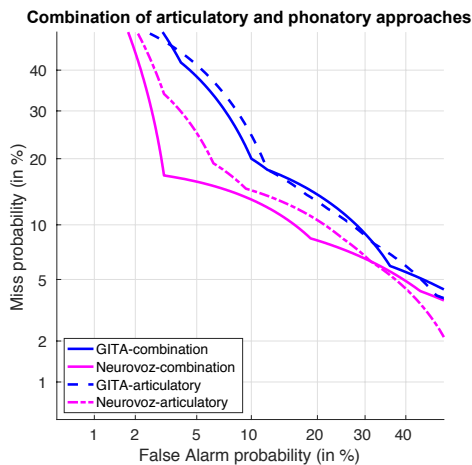


Fig. 1. ROCCH-DET curves of the combination of articulatory and phonatory approaches.

REFERENCES

- [1] A. Tsanas, M. a. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1264–1271, 2012.
- [2] J. Mekyska *et al.*, "Assessing progress of Parkinson's disease using acoustic analysis of phonation," in *IWOBI 2015*, pp. 111–118.
- [3] A. Benba *et al.*, "Voice analysis for detecting persons with parkinson's disease using plp and VQ.," *Journal of Theoretical & Applied Information Technology*, vol. 70 (3), 2014.
- [4] A. Benba, A. Jilbab, and A. Hammouch, "Detecting Patients with Parkinson's disease using Mel Frequency Cepstral Coefficients and Support Vector Machines," *International Journal on Electrical Engineering and Informatics*, vol. 7, no. 2, pp. 297–308, 2015.
- [5] L. Moro-Velázquez, J. A. Gómez-García, J. I. Godino-Llorente, J. Villalba, J. R. Orozco-Arroyave, and N. Dehak, "Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson's Disease," *Applied Soft Computing J.*, vol. 62, pp. 649–666, 2018.
- [6] J. Duffy, *Motor speech disorders: substrates, differential diagnosis, and management*, 2nd ed. St. Louis, MO: Elsevier, 2013.
- [7] L. Moro-Velázquez *et al.*, "Study of the automatic detection of Parkinson's Disease based on speaker recognition technologies and allophonic distillation," in *40th International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018.
- [8] L. Moro-Velázquez *et al.*, "A forced Gaussians based methodology for the differential evaluation of Parkinson's Disease by means of speech processing," *Biomed. Signal Process. Control*, vol. 48, pp. 205–220, 2019.
- [9] J. Orozco-Arroyave and J. Arias-Londoño, "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease," *LREC*, 2014.
- [10] A. Moreno *et al.*, "Albayzín speech database: Design of the phonetic corpus," *Eurospeech 1993. Proc. 3rd Eur. Conf. Speech Commun. Technol.*, vol. 1, pp. 175–178, 1993.
- [11] R. Baken and R. Orlikoff, *Clinical measurement of speech and voice*. Cengage Learning, 2000.
- [12] S. Feijoo and C. Hernández, "Short-term stability measures for the evaluation of vocal quality," *J. Speech Hear. Res.*, vol. 33, no. 2, pp. 324–34, Jun. 1990.
- [13] H. Kasuya, "Normalized noise energy as an acoustic measure to evaluate pathologic voice," *J. Acoust. Soc. Am.*, vol. 80, no. 5, p. 1329, Nov. 1986.
- [14] G. De Krom, "A Cepstrum-Based Technique for Determining a Harmonics-to-Noise Ratio in Speech Signals," *J. Speech Hear. Res.*, vol. 36, no. April 1993, pp. 254–266, 1993.
- [15] D. Michaelis, "Glottal-to-noise excitation ratio a new measure for describing pathological voices," *Acta Acust. united with Acust.*, vol. 83, no. 4, pp. 700–706, 1997.
- [16] H. Kantz and T. Schreiber, "Nonlinear time series analysis," *Cambridge Univ. Press*, vol. 7, 2003.
- [17] M. a Little, P. E. McSharry, S. J. Roberts, D. a E. Costello, and I. M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *Biomed. Eng. Online*, vol. 6, p. 23, Jan. 2007.
- [18] C. K. Peng, S. Havlin, H. E. Stanley, and A. L. Goldberger, "Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series," *Chaos*, vol. 5, no. 1, pp. 82–87, Mar. 1995.
- [19] S. M. Pincus, "Approximate entropy as a measure of system complexity," *Proceedings of the National Academy of Sciences*, vol. 88, pp. 2297–2301, 1991.
- [20] J. Richman and J. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *Am. J. Physiol. Circulatory Physiol.*, vol. 278(6), pp. H2039–H2049, 2000.
- [21] H. Xie, W. He, and H. Liu, "Measuring time series regularity using nonlinear similarity-based sample entropy," *Phys. Lett. A*, vol. 372, no. 48, pp. 7140–7146, 2008.
- [22] L. Xu, K. Wang, and L. Wang, "Gaussian kernel approximate entropy algorithm for analyzing irregularity of time-series," *Proc. 2005 Int. Conf. Mach. Learn. Cybern.*, vol. 9, pp. 5605–5608, 2005.
- [23] W. Chen, Z. Wang, H. Xie, and W. Yu, "Characterization of surface EMG signal based on fuzzy entropy," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 15, no. 2, pp. 266–272, 2007.
- [24] M. Zanin, L. Zunino, O. A. Rosso, and D. Papo, "Permutation Entropy and Its Main Biomedical and Econophysics Applications: A Review," *Entropy*, vol. 14, pp. 1553–1577, 2012.
- [25] L. Moro-Velázquez, "Towards the differential evaluation of Parkinson's Disease by means of voice and speech processing," Universidad Politécnica de Madrid, 2018.
- [26] G. Arias-Londono, J. D., Godino-Llorente, J. I., Sáenz-Lechón, N., Osma-Ruiz, V., & Castellanos-Dominguez, "Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 2, pp. 370–379, 2011.
- [27] L. Moro-Velázquez, J. A. Gómez-García, J. I. Godino-Llorente, and G. Andrade-Miranda, "Modulation Spectra Morphological Parameters: A New Method to Assess Voice Pathologies according to the GRBAS Scale," *Biomed Res. Int.*, vol. 2015, 2015.
- [28] L. Moro-Velázquez, J. A. Gómez-García, and J. I. Godino-Llorente, "Voice pathology detection using modulation spectrum-optimized metrics," *Front. Bioeng. Biotechnol.*, vol. 4, no. JAN, 2016.
- [29] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *Journal of machine learning research*, vol. 3, 1415–1438, 2003.
- [30] N. Saenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, and P. Gomez-Vilda, "Methodological issues in the development of automatic systems for voice pathology detection," *Biomed. Signal Process. Control*, vol. 1, no. 2, pp. 120–128, 2006.

JOINT ANALYSIS OF VOCAL JITTER, FLUTTER AND TREMOR IN VOWELS SUSTAINED BY NORMOPHONIC AND PARKINSON SPEAKERS

J. Schoentgen², A. Kacha¹, F. Grenez²

¹ Laboratoire de Physique de Rayonnement et Applications, University of Jijel, Jijel, Algeria

² B.E.A.M.S., Université Libre de Bruxelles, Brussels, Belgium

jschoent@ulb.ac.be, akacha@ulb.ac.be, fgrenez@ulb.ac.be

Abstract: The object of the presentation is a joint analysis of vocal cycle length perturbations in different frequency bands. The purpose is a comparison of the sizes of vocal jitter, flutter and tremor relative to each other, which may enable situating their physiological causes. Indeed, simulations suggest that the variability of the inter-spike intervals of the active motor neurons as well as of their firing rate influence unevenly fast and slow perturbations of the instantaneous frequency of vibration of the vocal fold body, whereas the relative amplitudes of vibration of the fold body and cover influence observed jitter, flutter and tremor proportionally.

Keywords: Vocal cycle length perturbations, TA-muscle tension, body-cover model of the vocal folds

I. INTRODUCTION

The objective is to present an analysis method that jointly estimates vocal jitter, flutter and tremor and enables comparing their sizes relative to each other. Jitter, flutter and "neurological" tremor designate vocal frequency perturbations the respective frequency bands of which are (>20Hz), (10-20Hz) and (2-10Hz) [1]. Ultra-slow perturbations in the (0-2Hz) band are caused by breathing, pulsatile blood flow and frequency drift and are not discussed. A corpus of control and Parkinson speakers is used to illustrate the analysis method.

The presentation focuses on type I speech signals exclusively. Type I signals are pseudo-periodic and monophonic. In type I signals, a possible cause of observed vocal perturbations is the fluctuations of the tension of the thyro-arytenoid (TA) muscle [2,3]. Physiological parameters that fix the size of the fluctuations are the number of active motor neurons, their dead time, their average firing rate and variability of inter-spike intervals (i.s.i) as well as the latency and rise time of the muscle twitches.

A reason to analyze vocal perturbations jointly in distinct frequency bands is that the ratios of the size of flutter or tremor relative to the size of jitter may be as informative as their absolute sizes. Indeed, jitter, flutter and tremor may, but must not evolve proportionally to each other. Indeed, simulations suggest that the size of

flutter relative to the size of jitter depends on the variability of the inter-spike intervals of the motor neurons [3]. In contrast, muscle tension fluctuations contribute to but do not explain vocal tremor. Independently from fluctuations of the muscle force, vocal tremor involves slow modulations of the firing frequency of the motor neurons [8].

Cycle length perturbations obtained for a corpus of control and Parkinson speakers confirm that tremor increases for the Parkinson speakers, but they show that the ratios of the size of tremor or flutter relative to the size of jitter stay the same for the two types of speakers, which would suggest that jitter, flutter and tremor have evolved proportionally. The presentation therefore focuses on fold-internal mechanisms that would modulate the size of muscle tension fluctuations irrespective of their cause and frequency band.

Two models are discussed, which explain the link between fluctuations of the TA-muscle tension owing to neural and muscular activity on the one hand and observed vocal perturbations on the other. The models distinguish between the body and cover of the vocal fold, which respectively designate the muscle and the lamina propria together with the epithelium. The models predict that the amplitudes of vibration of the body and cover relative to each other modulate the size of jitter, flutter and tremor proportionally.

II. METHODS

A. Corpus

The corpus comprises German vowels [a] sustained by 129 male and 76 female Parkinson speakers and 42 male and 32 female control speakers. The age of the Parkinson speakers ranges from 41 to 83 years (67.6 ± 13.7 years) and the age of the control speakers from 26 to 83 years (60.9 ± 13.7 years). The duration since diagnosis ranges from one year to 30 years (6.6 ± 4.9 years). The stimuli are .wav sampled at 44100 Hz and recorded via a condenser microphone in a quiet room at the Department of Neurology of the Knappschafts-Krankenhaus, Ruhr-University Bochum, Germany

B. Analysis of vocal cycle length time series

The vocal cycle lengths are estimated in the time domain using dynamic programming and the prominence of the speech cycle amplitudes [4]. The decomposition of a cycle length perturbation time series into a jitter, flutter and tremor time series is a straightforward generalization of the analysis of vocal jitter. Indeed, jitter is obtained by smoothing the cycle length time series by means of a running average, followed by subtracting the smoothed from the original time series and assigning the difference to jitter [5].

The decomposition into jitter, flutter, tremor and a residue involves carrying out the previous step three times. Each step involves smoothing (i.e. low-pass filtering) by a running average, subtracting the smoothed from the un-smoothed time series, storing the difference and replacing the un-smoothed by the smoothed time series before the next step. The successive subtractions guarantee that the decomposition is exact, that is, the sum of the stored differences and the residue is exactly equal to the raw time series. The jitter, flutter, tremor and residue time series are obtained by fixing the cut-off frequency of the running average low-pass filter to 20Hz, 10Hz and 2Hz respectively [1]. The residue is the filtered time series after step 3.

Before decomposition, the cycle length time series are upsampled by interpolation. Upsampling replaces the (feebly) variable raw cycle lengths by a constant time step and enables a more fine-grained selection of the length of the running average, which must be an odd number of samples.

Acoustic features are the coefficients of variation of the jitter, flutter and tremor time series, as well as the ratios of the size of flutter or tremor relative to the size of jitter.

III. MODELS

A. A model of the neurological causes of perturbations of the vocal cycle lengths

The objective is to model jitter and flutter via a simulation of the fluctuations of the tension of the thyro-arytenoid (TA) muscle. The model is inspired by a proposal by Titze, but drops problematic *a priori* assumptions of the latter.

The tension of a skeletal muscle is the outcome of the concurrent activity of many motor units of which each comprises a motor neuron that innervates a group of muscle fibers that contract after the arrival of an action potential, i.e. an electrical spike emitted by the neuron. The duration between two adjacent spikes is the inter-spike interval. The simultaneous contraction of several muscle fibers that are under the control of a single neuron is called a muscle twitch. The muscle tension is the outcome of the superposition of many

muscle twitches in space because of the concurrent firing of many motor neurons and in time because of the rapid succession of the action potentials of a single motor neuron.

In Titze's model, the average duration between successive action potentials that initiate muscle twitches is equal to one over the average firing rate of the neuron [2]. Shift values drawn from Gaussian distributions then randomly move the emitted spikes from their periodic default positions.

Randomly repositioning periodic spikes by normally distributed shifts may be problematic because of the symmetry and infinite support of the Gaussian distribution. Indeed, spikes may be shifted forward and backward in time and spike n may occur before spike $n - 1$ or later than spike $n + 1$ when the variability of the inter-spike intervals is large. In [2], unacceptable shifts have been avoided by keeping the coefficient of variation of the inter-spike intervals small (≤ 0.15).

We have updated Titze's model by following Deger et al.'s [6] advice and simulated inter-spike intervals by means of a Gamma distribution that generates positive inter-spike intervals for any firing rate and any coefficient of variation, without risking violating causality.

B. Models of the fold-internal modulation of the instantaneous fluctuations of the tension of the TA-muscle

Hereafter, we discuss two models of the link between the perturbations of the instantaneous tension of the body of the vocal folds (i.e. the TA-muscle) and observed vocal perturbations. The fold-inherent modulation modifies identically the fluctuations of the tension of the TA-muscle, whatever their frequency band or cause.

Model I. One model is kinematic and is based on the assumption that the body and cover of the vocal fold vibrate sinusoidally at the same average frequency, but at different amplitudes A_b and A_c . The instantaneous phases Φ_b and Φ_c are assumed to be the same up to a perturbation θ_{pert} of the phase of the fold body. The instantaneous phase of the fold cover is assumed to be unjittered. When the folds do not touch, the movement of the fold edge x_{edge} is the sum of the sinusoidal motions of the body and cover and a constant abduction.

$$x_{edge} = x_{abd} + A_b \sin(\phi_b) + A_c \sin(\phi_c) \quad (1)$$

Assuming for arguments's sake that $\theta_{pert} = \Phi_b - \Phi_c$ is small, that is, $\cos \theta_{pert} \approx 1$, $\sin \theta_{pert} \approx \theta_{pert}$ and $\tan \theta_{pert} \approx \theta_{pert}$ and applying an elementary trigonometric rule [9], one obtains the following approximate expression.

$$x_{edge} \approx x_{abd} + A \sin(\phi_c + \frac{A_b}{A_b + A_c} \times \theta_{pert}) \quad (2)$$

The argument of the sinusoid in approximation (2) suggests that the perturbations of the frequency of vibration of the muscle are modulated by the ratio of its depth of vibration relative to the total depth of vibration A_b/A_b+A_c .

Model II. A second model is based on the link proposed by Titze between the natural frequency of vibration F_o and the passive σ_p and active stress σ_a of the vocal folds (3). Symbol L_m designates the vibrating length of the vocal folds, ρ the density of the tissue, A_b the depth of vibration of the TA-muscle and A the total depth of vibration.

$$F_o = \frac{1}{2L_m} \sqrt{\frac{\sigma_p}{\rho}} \left(1 + \frac{A_b}{A} \frac{\sigma_a}{\sigma_p}\right)^{1/2} \quad (3)$$

The purpose of model (3) is to explain that the anatomical mass of the vocal folds is an irrelevant control parameter of F_o [7]. Expression (3) can however be turned into a model that relates perturbations of the active (i.e. TA-muscle) stress $\Delta\sigma_a$ to perturbations of the natural frequency ΔF_o by inserting expression (3) into the corresponding differential expression (4).

$$\Delta F_o = \frac{1}{4L_m^2 \rho} \frac{1}{F_o} \frac{A_b}{A} \Delta\sigma_a \quad (4)$$

Mechanical model (4) and kinematic model (2) assign a key role to the ratio of the depth of vibration of the TA-muscle relative to the total depth of vibration as a modulator of the perturbations of the instantaneous tension and/or frequency of vibration of the TA-muscle.

C. Simulations of vocal jitter and vocal flutter by means of a model of the fluctuations of the tension of the TA-muscle and a kinematic body-cover model of the vocal folds

Perturbations of the instantaneous phase of the body of a kinematic body-cover model have been simulated by means of the model of the fluctuations of the tension of the TA-muscle described in section III.A and in [3].

The glottal depth in the kinematic body-cover model is discretized into 10 glottal widths, each involving one expression (1). The purpose is to simulate the increasing phase delay of the fold cover from the glottal entrance to the glottal exit. Glottal closure is then simulated by means of a maximum operator that zeroes negative glottal widths, the smallest of which is then selected by means of a minimum operator to transform the glottal volume into the glottal area. Parameter values that have been kept constant are the average vocal frequency (120Hz), the vibrating fold length (1cm), the abduction (0.05cm)

and total amplitude of vibration (0.2cm) as well as the rise time (0.02sec) and latency (0.0015sec) of the muscle twitches

The effects of the remaining model parameters on measured glottal cycle lengths have been studied by means of 1000 simulations. Each involves randomly selecting a parameter value in the relative range (1 ± 0.5) times the average. The parameters and their absolute ranges have been the firing rate of the motor neurons (15Hz-45Hz); the coefficient of variation of the inter-spike intervals (0.15-0.30); the number of active neurons (50-150); the dead time of the motor neurons (0.00125-0.00375sec); the hemi-phase delay of the cover between the glottal entrance and exit ($\pi/12$ - $3\pi/12$ rad) and the amplitude of vibration of the body of the vocal folds (0.05cm-0.15cm).

The phase of model (1) may become numerically unstable when $A_c \sim A_b$ [9]. Outliers in simulated jitter, flutter and tremor have therefore been replaced by the next value in the array of 1000 simulations.

IV. RESULTS AND DISCUSSION

A. Vocal jitter, flutter and tremor in a corpus of normophonic and Parkinson speakers

The coefficients of variation of the jitter, flutter and tremor time series, as well as the values of the ratios of the size of flutter or tremor relative to the size of jitter have been obtained for the corpus of control and Parkinson speakers. The quartiles are reported in Table 1. Tests of the statistical significance of the differences between control and Parkinson speakers have been carried out by means of a one way ANOVA. The differences have been statistically significant ($p < 0.05$) for vocal flutter and tremor, but not for vocal jitter as well as the ratios of the size of flutter or tremor relative to the size of jitter ($p > 0.1$).

The lack of statistically significant differences between control and Parkinson speakers of the corresponding ratios would suggest that jitter, flutter and tremor have increased congruently in Parkinson speakers. A prosaic explanation would be uneven recording conditions that bias the estimates of the glottal cycle lengths. Hereafter, we discuss an alternative, which is the fold-internal modulation by the relative vibration depth of the TA-muscle.

B. Simulated vocal jitter, flutter and tremor

Table 2 shows the coefficients of variation of the simulated perturbations and the ratios of flutter or tremor relative to jitter and Table 3 shows the linear regression weights of the z-normalized model parameters described in section III.C predicting the z-

normalized values of jitter and flutter as well as the ratio of jitter relative to flutter.

Table 1: Quartiles of the jitter, flutter and tremor in % and of the ratios of the size of flutter or tremor relative to the size of jitter for a corpus of vowels sustained by control and Parkinson speakers.

		Min	25%	50%	75%	Max
Control	Jitter (%)	0.18	0.3	0.41	0.66	2.49
	Flutter (%)	0.12	0.2	0.31	0.45	1.35
	Tremor (%)	0.28	0.51	0.79	0.99	2.42
	Flutter/Jitter	0.14	0.45	0.66	0.95	2.11
	Tremor/Jitter	0.24	1.15	1.72	2.32	5.2
Park.	Jitter (%)	0.14	0.35	0.48	0.84	3.34
	Flutter (%)	0.12	0.27	0.37	0.49	2.49
	Tremor (%)	0.31	0.67	0.89	1.16	8.09
	Flutter/Jitter	0.17	0.55	0.77	0.97	1.64
	Tremor/Jitter	0.26	1.26	1.71	2.47	5.48

The parameter called 'activity' is the z-normalized product of the firing rate of the motor neurons with the number of motor units. The second line in Table 3 reports the weight of the amplitude of vibration of the cover relative to the total amplitude of vibration.

Table 2: Quartiles of the simulated jitter, flutter and tremor in % as well as of the flutter/jitter and tremor/jitter ratios.

	Quartile 1	Median	Quartile 3
Jitter (%)	0.14	0.21	0.34
Flutter (%)	0.09	0.14	0.22
Tremor (%)	0.15	0.25	0.36
Flutter/Jitter	0.58	0.71	0.84
Tremor/Jitter	0.84	1.14	1.47

Table 3: Statistically significant regression coefficients of the z-normalized model parameters described in section III.C predicting the z-normalized sizes of jitter and flutter as well as the ratio flutter/jitter

	Jitter	Flutter	Flutter/Jitter
1-A _p /A	-0.46	-0.44	+0.06
Activity	-0.53	-0.43	n.s.
CV i.s.i.	+0.14	+0.38	+0.33

Comparing observed (Table 1) and simulated (Table 2) ratios suggests that flutter may be explained by the fluctuations of the tension of the TA-muscle,

whereas the latter contribute to, but do not explain vocal tremor.

The regression weights in Table 3 indicate that simulated jitter and flutter (i) decrease with an increase of the activity of the TA-muscle; (ii) increase with an increase of the variability of the inter-spike intervals of the motor neurons; (iii) increase with a decrease of the amplitude of vibration of the cover relative to the muscle and (iv) the size of flutter relative to jitter depends on the variability of the inter-spike intervals.

The vibration depths of the body and cover of the vocal folds relative to each other are the missing link between observed vocal jitter and flutter and the fluctuations of the TA-muscle tension. The relative amplitudes of vibration of the muscle and cover modulate muscle tension fluctuations proportionally, whereas muscle activity and i.s.i. variability influence jitter, flutter and tremor unevenly.

REFERENCES

- [1] E. H. Buderand and E. A. Strand, "Quantitative and graphic acoustic analysis of phonatory modulations: the modulogram," *J. Speech, Language, Hearing Res.*, vol. 46, pp. 475–490, 2003.
- [2] I. R. Titze, "A model for neurologic sources of aperiodicity in vocal fold vibration," *J. Speech, Hearing Res.*, vol. 34, pp. 460–472, 1991.
- [3] J. Schoentgen and Ph. Aichinger, "Analysis and synthesis of vocal flutter and vocal jitter," in *Proceedings Interpeech*, Graz, 2019.
- [4] C. Mertens, F. Grenez, L.-C. Buchman and J. Schoentgen. "Reliable tracking based on speech sample salience of vocal cycle length perturbations." in *Proceedings Interspeech*, Makuhari, Japan, 2010, pp. 2566–2569.
- [5] P. Boersma and D. Weeninck, *Praat: doing phonetics by computer [Computer program]*, 2014, [Version 5.4.04, retrieved 2014 from <http://www.praat.org>].
- [6] M. Deger, M. Helias, C. Boucsein and S. Rotter, "Statistical properties of superimposed stationary spike trains," *J. Comput. Neurosci.*, vol. 32, 443–463, 2012.
- [7] I. R. Titze, "Vocal fold mass is not a useful quantity for describing Fo in vocalisation." *J. Speech Lang Hear Res.*, vol. 54(2), 2011, 520-5221.
- [8] I. R. Titze, B. Story, M. Smith and R. Long, "A reflex resonance model of vocal vibrato." *J. Acoust. Soc. Am.*, vol. 111(5), 2002, 2272-2282.
- [9] H. S. Black, *Modulation Theory*. Van Nostrand, 1953, page 221.

THE EFFECTS OF DEEP BRAIN STIMULATION ON SPEECH ARTICULATION AND VOCALIZATION IN PARKINSON'S DISEASE

J. J. Sidtis^{1,2}, D. Van Lancker Sidtis^{1,3}, R. Ramdhani⁴, M. Tagliati⁵

¹ Nathan Kline Institute/Brain and Behavior Laboratory, Orangeburg NY, USA

² New York University Langone Medical School/Psychiatry, NY NY, USA

³ New York University/Department of Neurology, NY NY, USA

⁴ Northwell Health/Neurosciences, Great Neck NY, USA

⁵ Cedars-Sinai Medical Center/Department of Neurology, Los Angeles CA, USA

John.Sidtis@nyu.edu

Diana.Sidtis@nyu.edu

Ritesh.A.Ramdhani@gmail.com

Michele.Tagliati@cshs.org

Abstract: High frequency deep brain stimulation (DBS) of the subthalamic nucleus (STN) has become an effective and widely used tool in the treatment of Parkinson's disease (PD). This intervention has varied effects on speech in PD. The present study examined the effects of different stimulation frequencies in the intelligibility of spontaneous speech produced by individuals with PD who are being treated with DBS. Questions remain about preferred DBS stimulation settings for optimal outcome. Seven right-handed, native speakers of English with PD were studied off medication at three stimulation conditions: OFF, 60 Hz (LFS), and the clinical setting of 185 Hz (HFS). Spontaneous speech was recorded at each frequency and excerpts were prepared for transcription (intelligibility) and difficulty judgements. Intelligibility for spontaneous speech was reduced at both high frequency stimulation (HFS) and low frequency stimulation (LFS) compared to OFF, with speech produced at HFS more intelligible than that produced at LFS. Both voice quality and articulation were judged to be more abnormal with stimulation. However, voice quality ratings during stimulation were correlated with those without stimulation. This was not true for articulation ratings, suggesting that STN-DBS exacerbation existing voice abnormalities but may have introduced new articulatory abnormalities. **Keywords:** Parkinson's disease, deep brain stimulation (DBS), speech, voice, dysarthria.

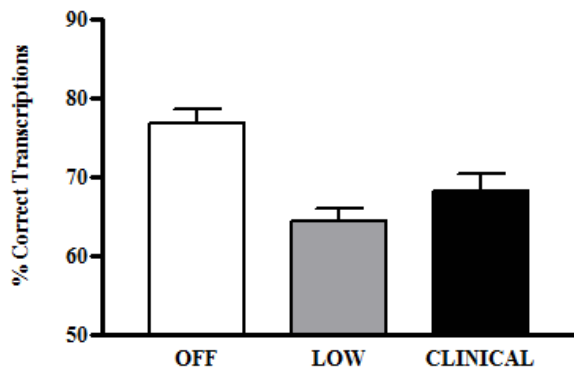
I. INTRODUCTION

Progressive loss of the ability to speak normally is a characteristic feature of Parkinson's disease (PD). Both the ability to articulate properly and the control

of one's voice are affected. An increasingly common form of therapy involves high frequency electrical stimulation of deep brain structures. This treatment acts like the more common pharmacological therapy with Levodopa. Unfortunately, neither form of treatment improves speech or voice, and the electrical stimulation has actually been associated with increasing difficulties in these areas [1,2]. Research has demonstrated varied effects of deep brain electrical stimulation on speech in PD. In some studies vowel production improved, but many individuals exhibit reduced speech intelligibility or complain of increased difficulty with speech articulation. Questions remain about the ideal parameters of deep brain stimulation for optimal clinical outcome and minimal side effects [3]. Some have suggested that the potential negative consequences of electrical brain stimulation can be reduced by lowering the frequency of the stimulation. While these questions are being pursued, there are opportunities to observe the relative effects of deep brain electrical stimulation on the articulatory and vocal components of speech.

II. METHODS

Seven right-handed, native speakers of English with PD treated with bilateral deep brain stimulation of the subthalamic nucleus were studied off medication at three conditions: stimulators off, stimulation at 60 Hz, and the clinical setting of stimulation at 185 Hz. Spontaneous speech was elicited as the effects of PD and DBS are more pronounced in this mode compared to reading and repetition [4,5]. Spontaneous speech was recorded in each condition and excerpts were prepared for transcription (intelligibility) and difficulty judgements. Separate excerpts were prepared for listeners to rate abnormalities in voice,



Deep Brain Stimulation Setting

articulation, fluency, and rate. Excerpts prepared for the intelligibility task were 5-10 words in length, with 1-5 syllables per word. Utterances differed across conditions, eliminating the possibility of order or practice effects. Difficulty ratings provide additional information about the experience of the listener in the process of determining the intelligibility of the spoken utterances. During the intelligibility phase, listeners were asked to listen to the phrases, which were 5-10 words in length as described above, and write down what they heard using pen or pencil on a numbered answer sheet provided. For each item they were also asked to rate the level of difficulty they experienced in transcribing the utterance using a scale of 1 (easy) to 5 (difficult) and circling one of these numbers on the answer sheet for each item. No linguistic support was provided for the transcriptions; answer sheets consisted only of a number and a blank line for each stimulus item.

III. RESULTS

Intelligibility for excerpts from spontaneous speech changed with DBS frequency [$F(2,28) = 24.03$; $p < 0.001$]. Intelligibility was reduced at both high frequency stimulation (HFS) [$t(14) = -3.63$; $p = 0.003$] and low frequency stimulation (LFS) [$t(14) = -8.46$; $p < 0.001$] compared to when the stimulation was off, with speech produced at HFS more intelligible than that produced at LFS [$t(14) = 2.5$; $p = 0.025$]. The average transcription accuracy scores are presented in Fig. 1. The difficulty of transcribing the utterances also increased with stimulation [$F(2,28) = 16.35$; $p < 0.001$]. This is depicted in Fig. 2. Both voice quality and articulation were judged to be more abnormal with stimulation on, but the voice abnormalities represented an exacerbation of the abnormalities

present without stimulation. In contrast, stimulation appeared to introduce articulatory problems not present without stimulation.

Fig. 1. The average percentage of words accurately transcribed from spontaneous speech produced during each of three deep brain stimulation settings. The voltages and amperages were constant in the two stimulation conditions: only the frequencies were changed.

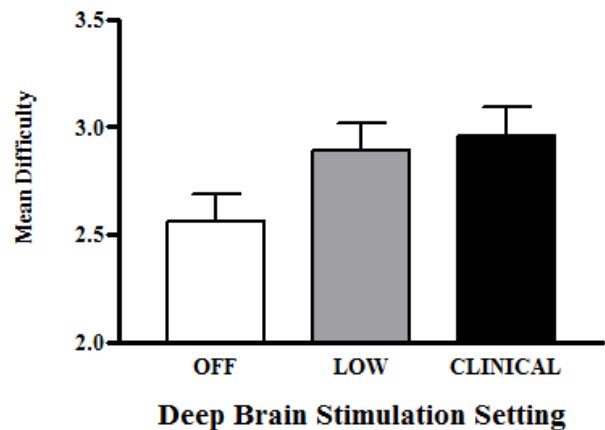


Fig. 2. The average difficulty ratings of performing the transcriptions for spontaneous speech during each of three deep brain stimulation settings.

IV. DISCUSSION

Reducing frequency at which the brain is stimulated in treating PD does not necessarily improve the intelligibility of spontaneous speech. Voice quality and articulation, as rated by listeners, were both negatively affected. Voice quality abnormality ratings during stimulation were correlated with the ratings obtained while the stimulators were turned off. This suggests that the stimulation worsened voice abnormalities that were present in the absence of stimulation. In contrast, the ratings of articulatory abnormalities during stimulation were not associated with articulation when the stimulators were turned off.

V. CONCLUSIONS

These results suggest that the stimulation is introducing new abnormalities in articulation [6,7] and exacerbating pre-existing Parkinsonian abnormalities [8]. While there is substantial clinical evidence that vocalization during speech and singing reflect different brain systems, the observations of speech intelligibility and its characteristics with and

without deep brain stimulation suggests that vocalization and articulation operate with some independence during spoken language as well.

REFERENCES

- [1] G. Deutchel, J. Herzog, G. Kleiner-Fisman, C. Kubu, A. M. Lozano, K. E. Lyons, M. C. Rodriguez-Oroz, F. Tamma, A. I. Troster, J. L. Vitek, J. Volkmann, V. Voon, "Deep brain stimulation: Postoperative issues," *Mov. Disord*, vol 21 (Suppl 14), S219-237, 2006.
- [2] S. Skodda, "Effect of deep brain stimulation on speech performance in Parkinson's disease," *Parkinsons Dis*, Article ID 850596, 2005.
- [3] D. Aldridge, D. Theodoros, A. Angwin, A.P. Vogel, "Speech outcomes in Parkinson's disease after subthalamic nucleus deep brain stimulation: A systematic review," *Parkinsonism Relat. Disord*, vol. 33, 3-11, 2016.
- [4] D. Van Lancker Sidtis, T. Rogers, V. Godier, M. Tagliati, J.J. Sidtis, "Voice and fluency changes as a function of speech task and deep brain stimulation," *J. Speech Lang. Hear. Res*, vol. 53, 1167-1177, 2010.
- [5] D. Van Lancker Sidtis, K. Cameron, L. Bonura, J.J. Sidtis, "Speech intelligibility by listening in Parkinson speech with and without deep brain stimulation," *J. Neurolinguistics*, vol. 25, 121-132, 2012.
- [6] E. Eklund, J. Qvist, L. Sandström, F. Viklund, J. van Doorn, F. Karlsson, "Perceived articulatory precision in patients with Parkinson's disease after deep brain stimulation of subthalamic nucleus and caudal zona incerta," *Clin. Linguist. Phon*, vol. 29, 150-166, 2015.
- [7] M.V. Sauvageau, J.P.Roy, L. Cantin, M. Prud'Homme, L. Langlois, J. Macoir, "Articulatory changes in vowel production following STN DBS and levodopa intake in Parkinson's disease," *Parkinsons Dis*, e382320, 2015.
- [8] S. Skodda, G. Wenke, U. Schlegel, "Impairment of vowel articulation as a possible marker of disease progression in Parkinson's disease," *PlosONE*, vol 7(2), e32132, 2012.

BIOMEDICAL SPEECH SIGNAL INSIGHTS FROM A LARGE SCALE COHORT ACROSS SEVEN COUNTRIES: THE PARKINSON'S VOICE INITIATIVE STUDY

Athanasios Tsanas¹, Siddharth Arora²

¹ Usher Institute, Edinburgh Medical School, University of Edinburgh, Edinburgh, UK

² Mathematical Institute, University of Oxford, Oxford, UK

A. Tsanas: atsanas@ed.ac.uk, tsanasthanasis@gmail.com. S. Arora: arora@maths.ox.ac.uk

Abstract: Previous work has demonstrated the enormous potential of speech signals collected under highly controlled acoustic conditions in biomedical applications. These include accurately differentiating people diagnosed with Parkinson's Disease (PD) from Healthy Controls (HC), and longitudinal telemonitoring of PD symptom severity. The generalizability and scalability of these findings need to be investigated when speech signals are not recorded under optimal, carefully controlled acoustic conditions. In this regard, we recently completed the Parkinson's Voice Initiative (PVI) study collecting data from a very large cohort comprising 1483 PD and 8300 HC participants from seven countries. Specifically, we collected 19,303 sustained vowel /a/ recordings: 144 (Argentina), 227 (Brazil), 1521 (Canada), 75 (Mexico), 573 (Spain), 4088 (UK) and 12,675 (USA). We acoustically characterized these recordings using 307 dysphonia measures which we had previously investigated in related PD studies. We draw comparisons against previous studies which processed high quality speech data, and their generalizability in this large-scale cohort. We found that many of the state-of-art nonlinear dysphonia measures do not differentiate PD and HC sufficiently well, likely because of the reduced signal bandwidth. These exploratory findings provide new insights into understanding the challenges in the PVI dataset, underlining the need for further speech signal processing development.

Keywords: Parkinson's Disease (PD), Parkinson's Voice Initiative (PVI), speech signal processing, sustained vowel phonations

I. INTRODUCTION

The potential of capitalizing on acoustic analysis of speech signals to develop decision support tools across diverse biomedical applications has received considerable research attention in the last 10-15 years. Many successful applications of analyzing speech signals in neurodegenerative disorders, such as Parkinson's Disease (PD) have been reported. Indicatively, we had previously used sustained vowels to demonstrate: (a) almost 99% differentiation of people diagnosed with PD from Healthy Controls (HC) [1], (b) accurate replication of the standard PD symptom

severity metric, Unified Parkinson's Disease Rating Scale (UPDRS) [2], [3], and (c) automatic assessment of voice rehabilitation after people diagnosed with PD completed Lee Silverman Voice Treatment (LSVT) [4]. Furthermore, we reported that speech may be a biomarker towards distinguishing people with Leucine-Rich Repeat Kinase 2 (LRRK2) mutations, idiopathic PD, and HC [5], and identifying people at risk of developing PD who have been diagnosed with rapid eye movement sleep behavior disorder [6]. Other tasks such as the diadochokinetic test and continuous speech have been used to assess dysarthria in the context of PD assessment [7], [8]. There is also considerable work focusing on developing speech articulation kinematic models based on first-principles, which may provide tentative insights into the underlying vocal production mechanism and problems associated with neurodegenerative disorders [9], [10].

Typically, studies focusing on speech signal analysis report findings on carefully screened cohorts devoid of potential comorbidities, using a limited number of phonations which are collected under highly controlled acoustic settings [11], [12]. Hitherto, it is not clear whether the developed algorithmic tools would work sufficiently well in settings where speech recordings are not collected in acoustically controlled settings and using ubiquitous recording equipment.

We have recently completed the Parkinson's Voice Initiative (PVI) study [13], aiming to collect speech data in a community setting under free-living conditions and uncontrolled acoustic environments. The PVI was conceived to assess the robustness and generalizability of speech signals and the methods we had previously developed [1] towards developing a mass population screening tool for PD. With limited financial resources, we have collected the largest PD speech database to-date within a few months. In our first, recent PVI study we reported early results towards capitalizing on telephone-quality sustained vowels for differentiating PD from HC [13]. The overall balanced accuracy was less than 70%, but this is still a promising finding given the uncontrollable factors in the PVI data collection process.

This study aims to summarize the key lessons learned from this exercise, the insights that we have gained comparing data collected in this community study versus data collected under controlled acoustic conditions in [1], and ultimately maps out plans to further mine this unique, extremely rich resource for biomedical applications.

II. DATA

The PVI project was set in seven major geographical locations where participants were invited to contribute their voice samples along with basic demographic information (age, gender). The participants were self-selected and were asked to self-report whether they had been given a PD clinical diagnosis. They were instructed to sustain vowel /a/ for as long and as steadily as possible, following standard speech protocols [11]. The phonations were sampled at 8 kHz and stored on secure cloud servers. For each participant we collected two phonations to account for potential problems in the recording. Ultimately, we collected 19,303 voice recordings: 144 (Argentina), 227 (Brazil), 1521 (Canada), 75 (Mexico), 573 (Spain), 4088 (UK) and 12,675 (USA). Participants could find further details about the study by optionally pressing a button when making the call before any data collection; they were notified that by continuing the call they would be providing informed consent for their data to be used in this research project. For further details on the PVI study please refer to [13].

III. METHODS

A. Data Preprocessing

We screened out non-usable recordings (too noisy or very short, failure to record speech), and processed 2759 recordings from 1483 PD participants, and 15,321 recordings from 8300 HC.

B. Acoustic characterization of phonations

We applied a range of classical and nonlinear speech signal processing algorithms to extract 307 dysphonia measures. These speech signal processing algorithms aim to extract intricate properties quantifying deviation from periodicity (jitter and shimmer variants expressing deviations in steady vocal fold closure patterns), energy-related concepts and the creation of turbulent noise, and subtle changes in the placement of the articulators.

Further details on the physiological underpinning and algorithmic expressions for the computation of the dysphonia measures were previously described in detail in our previous studies [12], [14], [15]. The fundamental frequency (F0), which is one of the key characteristics of speech [11], was computed using the SWIPE

algorithm [16] which we had previously demonstrated is the most accurate F0 estimation algorithm in sustained vowel /a/ signals, including a setting with lower quality data [17]. The Matlab source code for the computation of the dysphonia measures and the ensemble computation for very accurate F0 estimation in the context of sustained vowels is freely available on the first author's website: <https://www.darth-group.com/software>.

Overall, applying these speech signal processing algorithms to the PVI dataset led to the computation of a 18,080×307 feature matrix.

C. Visualization and statistical exploration

We used the top-10 ranked features reported in [1] to investigate how well these dysphonia measures generalize in the setting with lower quality data. We computed Spearman correlation coefficients between the selected feature subset and the binary outcome (PD vs HC), aiming to compare findings using the PVI dataset against our previous study where we had used high quality speech recordings [1]. Moreover, we provide a succinct image to facilitate visual exploration of the data using the 10 investigated dysphonia measures. The long diagonal presents the univariate histograms of the 10 dysphonia measures and the off long-diagonal are the colored scatter plots.

IV. RESULTS

Table I provides the statistical comparison of the ranked features using Spearman correlations in [1] and this study. We observe that the statistical correlations are considerably lower in this study compared to our previous findings [1].

Table I: Statistical associations between dysphonia measures and the binary outcome (PD vs HC)

Dysphonia measure	Spearman correlations	
	Tsanas et al. [1]	This study
VFER _{entropy}	-0.388	-0.014
VFER _{NSR,TKEO}	-0.379	-0.018
11 th MFCC	0.369	0.046
VFER _{NSR,SEO}	-0.365	-0.030
4 th delta MFCC	-0.363	0.055
VFER _{mean}	-0.321	0.010
RPDE	0.292	0.014
DFA	0.287	0.023
Shimmer _{PQ11}	0.285	0.030
HNR _{mean}	-0.285	-0.066

We provide the correlation coefficients reported in [1] for convenience and for reference with the current study. VFER stands for Vocal Fold Excitation Ratio, MFCC for Mel Frequency Cepstral Coefficient, RPDE for Recurrence Period Density Entropy, DFA for Detrended Fluctuation Analysis, and HNR for Harmonics to Noise Ratio.

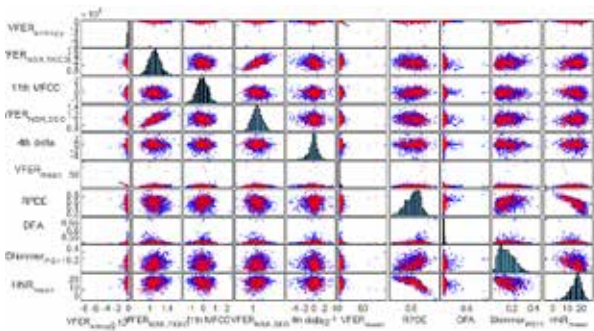


Figure 1: Colored scatter plots with embedded histograms for the 10 dysphonia measures used. Blue indicated HC and red PD.

Figure 1 presents the colored scatter plot: visually there is no obvious pattern discerning the PD and HC groups. This suggests that unlike [1] these dysphonia measures appear not to be able to univariately or pairwise jointly discriminate groups in the PVI dataset.

V. DISCUSSION

This study provides further insights into the PVI data towards understanding the complexity and challenges in the collected data, building on our previous work [13], [18]. We investigated the 10 dysphonia measures which previously exhibited fairly strong [19] statistical association with the binary outcome (PD vs HC) in [1]. We found that these dysphonia measures appear to be weakly correlated with the outcome in the PVI dataset.

We remark that the set of dysphonia measures explored here was developed specifically for biomedical applications and we have previously shown it may generalize well in other settings with high quality data [20]. Many of the advanced nonlinear methods rely on high quality data, and indeed methods such as VFER belong to the wider family of signal-to-noise ratio approaches which appear to work very well in biomedical applications [3], [12], [21]. This family of dysphonia measures aims at capturing information at the higher end of the spectrum; empirically we had previously suggested that 2.5 kHz could be used as a threshold to denote ‘noise’ in the signal [3], [12].

Given the objective constraints in terms of data bandwidth of the phone network system, the higher end of the frequencies are not captured and the nonlinear dysphonia measures appear to be very sensitive to the noisy environments thus losing their typical edge in this application when having high quality recordings [1], [12]. Overall, this emphasizes the need to develop more robust dysphonia measures which generalize well in low quality signals (sampled at 8 kHz with likely different distortions in their frequency responses introduced due to the recording equipment). Our findings indirectly support the well-established practical endorsement of using at least 20 kHz sampling rate whenever possible

[11], which is likely particularly required for assessment of disorders which leave a voice imprint.

In the results reported we processed collectively data from seven geographical locations. It is possible there is some way to stratify the data and repeat the analysis which might improve current results. We had repeated the exploratory analysis for individual cohorts, or focusing e.g. only on the US cohort [19] but did not find any clear patterns. This merits further investigation which we are currently working on. Recently, we reported on projecting selected dysphonia measure subsets in two dimensional spaces towards exploring relationships in differentiating groups [19] which facilitates new ways to explore this dataset.

We previously demonstrated using the PVI data that we could differentiate PD versus HC [13] with sensitivity and specificity that was considerably lower than the almost 99% accuracy reported when using high quality speech signals recorded under carefully controlled acoustic environments [1]. The PVI data where participants self-enrol and participate using their own devices comes with all associated difficulties of standardizing equipment and the data collection setting (different background acoustic noise, different microphones and response curves in the recording mechanisms, signal quality issues). The PVI study is, to the best of our knowledge, the first study that attempts to achieve this aim at such a large scale across different countries and under non-controlled acoustic conditions and therefore may potentially lead to a disruptive practical intervention [22].

The PVI findings overall support the use of speech for biomedical applications at scale, although careful consideration needs to be exercised in terms of the voice quality degradation when data is collected under non-controlled acoustic settings. It also highlights the need to develop new acoustic measures which are more robust to noisy acoustic environments. We are currently further exploring the use of the dataset stratifying cohorts to investigate differences in acoustic measures across languages and cohorts aiming to report both on normative data and dysphonia measure effects which are characteristic for different languages.

ACKNOWLEDGMENT

We are grateful to M.A. Little who led the Parkinson’s Voice Initiative where the data for this study was collected, to Ladan Baghai-Ravary for developing the tools for data collection, and to Aculab for hosting the data cloud server. We would like to extend our thanks to all participants in the PVI study. The study was made possible through generous funding via an EPSRC-NCSML award to AT and SA.

REFERENCES

- [1] A. Tsanas, M.A. Little, P.E. McSharry, J. Spielman, L.O. Ramig: “Novel speech signal processing algorithms for high-accuracy classification of Parkinson’s disease,” *IEEE Transactions on Biomedical Engineering*, Vol. 59, pp. 1264-1271, 2012
- [2] A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig: “Accurate telemonitoring of Parkinson’s disease progression by non-invasive speech tests”, *IEEE Transactions on Biomedical Engineering*, Vol. 57, pp. 884-893, 2010
- [3] A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig: “Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson’s disease symptom severity,” *Journal of the Royal Society Interface*, Vol. 8, pp. 842-855, 2011
- [4] A. Tsanas, M.A. Little, C. Fox, L.O. Ramig: “Objective automatic assessment of rehabilitative speech treatment in Parkinson’s disease”, *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 22, 181-190, 2014
- [5] S. Arora et al.: “Investigating voice as a biomarker for leucine-rich repeat kinase 2-associated Parkinson’s disease: a pilot study,” *Journal of Parkinson’s Disease*, Vol. 8(4), pp. 503-510, 2018
- [6] S. Arora, et al.: “Smartphone motor testing to distinguish idiopathic REM sleep behavior disorder, controls, and PD”, *Neurology* Vol. 91 (16), pp. e1528-e1538, 2018
- [7] J. I. Godino-Llorente, S. Shattuck-Hufnagel, J.Y. Choi, L. Moro-Velazquez, J.A. Gomez-Garcia, “Towards the identification of idiopathic Parkinson’s disease from the speech. New articulatory kinetic biomarkers”, *PLOS One*, Vol. 12(12): e0189583, 2017
- [8] J.R. Orocz-Aroyave et al. “Automatic detection of Parkinson’s disease in running speech spoken in three different languages”, *Journal of Acoustical Society of America*, Vol. 139, pp. 481-500, 2016
- [9] P. Gómez-Vilda et al.: “Phonation biomechanics in quantifying Parkinson’s disease symptom severity”, in *Recent Advances in Nonlinear Speech Processing*, Springer, pp. 93-102, 2016
- [10] P. Gomez-Vilda, J. Mekyska, A. Gomez, D. Palacios, V. Rodellar, A. Alvarez: “Characterization of Parkinson’s disease dysarthria in terms of speech articulation kinematics”, *Biomedical Signal Processing and Control*, Vol. 52, pp. 312-320, 2019
- [11] I.R. Titze: *Principles of Voice Production*. National Center for Voice and Speech, Iowa City, US, 2nd printing, 2000
- [12] A. Tsanas: *Accurate telemonitoring of Parkinson’s disease symptom severity using nonlinear speech signal processing and statistical machine learning*, Ph.D. thesis, Oxford Centre for Industrial and Applied Mathematics, University of Oxford, 2012
- [13] S. Arora, L. Baghai-Ravary, A. Tsanas: “Developing a large scale population screening tool for the assessment of Parkinson’s disease using telephone-quality speech,” *Journal of Acoustical Society of America*, Vol. 145(5), 2871-2884, 2019
- [14] A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig: “New nonlinear markers and insights into speech signal degradation for effective tracking of Parkinson’s disease symptom severity”, *International Symposium on Nonlinear Theory and its Applications (NOLTA)*, pp. 457-460, Krakow, Poland, 5-8 September 2010
- [15] A. Tsanas: “Acoustic analysis toolkit for biomedical speech signal processing: concepts and algorithms”, *8th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, pp. 37-40, Florence, Italy, 16-18 December 2013
- [16] A. Camacho, J.G. Harris, “A sawtooth waveform inspired pitch estimator for speech and music”, *Journal of the Acoustical Society of America*, Vol. 124, 1638-1652, 2008
- [17] A. Tsanas, M. Zañartu, M.A. Little, C. Fox, L.O. Ramig, G.D. Clifford, “Robust fundamental frequency estimation in sustained vowels: detailed algorithmic comparisons and information fusion with adaptive Kalman filtering”, *Journal of the Acoustical Society of America*, Vol. 135, 2885-2901, 2014
- [18] A. Tsanas, M.A. Little, P.E. McSharry: “A methodology for the analysis of medical data”, in *Handbook of Systems and Complexity in Health*, Eds. J.P. Sturmburg, and C.M. Martin, Springer, pp. 113-125 (chapter 7), 2013
- [19] A. Tsanas, S. Arora: “Exploring telephone-quality speech signals towards Parkinson’s disease assessment in a large acoustically non-controlled study”, *19th IEEE International Conference on BioInformatics and BioEngineering*, Athens, Greece, 28-30 October 2019
- [20] E. San Segundo, A. Tsanas, P. Gomez-Vilda: “Euclidean distances as measures of speaker similarity including identical twin pairs: a forensic investigation using source and filter voice characteristics”, *Forensic Science International*, Vol. 270, pp. 25-38, 2017
- [21] J.I. Godino-Llorente, V. Osuma-Ruiz, N. Saenz-Lechon, P. Gomez-Vilda, M. Blanco-Velasco, F. Cruz-Roldan: “The effectiveness of the glottal to noise excitation ratio for the screening of voice disorders”, *Journal of Voice*, Vol. 24(1), pp. 47-56, 2010
- [22] A.K. Triantafyllidis, A. Tsanas: “Applications of machine learning in real-life digital health interventions: review of the literature,” *Journal of Medical Internet Research (JMIR)*, Vol. 21(4), e12286, 2019

SESSION II
SINGING VOICE

TOWARDS A SOMATOSENSORY TRAINING DIGITAL ENVIRONMENT FOR LYRIC SINGING PEDAGOGY

E. Angelakis¹, A. Georgaki¹

¹ Laboratory of Music Acoustics and Technology (LabMAT), University of Athens, Athens, Greece
angelakisv@gmail.com, georgaki@music.uoa.gr

Abstract: Lyric-voice quality level has been repeatedly reported as declining during the last few decades, both by scholars and distinguished opera professionals. Many believe that the current vocal pedagogy for this genre is to be blamed for this effect and have reported a need for vocal education reform. Meanwhile, scientific research of the vocal mechanism has provided knowledge on an interdisciplinary level, in fields such as Singing Acoustics and Cognition, while new real-time visual feedback technologies and software tools, that can be of immediate use to the vocal pedagogic system, are evolving constantly. In this paper we propose an assistive feedback environment for the pedagogy of lyric singers, which includes a multi-window interface for testing postural and breathing habits, diaphragm activity, glottal activity and formant tuning.

Keywords: lyric singing, EGG, somatosensory, singing pedagogy, visual feedback

I. INTRODUCTION

Scientific research on singing vocal production, emission and projection has its origins in antiquity and it is mostly related to medicine, acoustics of theaters and pedagogy [1-3]. More precisely, vocal pedagogy “phonaskia” - emerged and was delivered to singers, actors and rhetors by professional teachers, the “phonaskous” [4]. After this, since the 1600s and the Renaissance period, the re-birth of arts and creation of the Opera genre, demanded singers of a very high proficiency level, and thus singing training came to prosper. During the following 300 years many famous vocal pedagogues (Giulio Caccini, Pier Francesco Tosi, Nicola Porpora, Manuel García II, Mathilde Marchesi, to list but a few prominent teachers) have developed singing and teaching techniques to ameliorate their trainee’s vocal abilities such as range, volume, stamina, agility and artistry [5]. The pinnacle of this movement was the famous “school of bel canto”. This school and its branches catalyzed the bringing up of many generations of legendary operatic singers, the vocal qualities of the latest of which can be attested, not only by writings of their contemporaries, but also through the recordings they have left behind.

Subsequent to this large period of vocal pedagogy flourish, the last 100 years have been a time when operatic voice education has been discussed negatively ever more often. First signs of its decline are first reported by, singer and vocal teacher, Lilli Lehmann as early as 1902 [6], while reports of the same issue have been more often since 1960 [7-11]. It suffices to quote the 7 years preliminary report of an ongoing Princeton University survey, in which 95% of the experts interviewed observes problems such as “decline in the number of great spinto and dramatic singers, especially of Verdi, beginning around 1980”, and also to refer to the same research conclusion that “For the first time since Verdi was alive, most opera professionals believe that it is impossible to cast these [late Verdi] operas at a level that would once have been considered minimally acceptable” [10].

II. THE SINGING VOICE RESEARCH IN THE DIGITAL ERA

Singing, and lyric singing in particular, is an art form of interdisciplinary interest, as its multifaceted nature can be examined through many, seemingly unrelated, sciences. Musicology professor Kay Norton, describes the possible benefits of singing in her book «*Singing and Wellbeing: Ancient Wisdom, Modern Proof*» (2015) by combining theories and sources taken from the fields of Ethnomusicology, Cultural Anthropology, Anatomy, Phoniatrics, Acoustics, Psychoacoustics, Psychology, Philosophy, Theology, Sociology, Neuroscience, Biochemistry and Music Therapy [12]. Presently, the most relevant question seems to be how vocal education could employ this knowledge and research, blend it with music neuro-aesthetics and the tradition of empiric pedagogy and eventually fuse it into a new model for operatic vocal training.

Emerging technologies for the production, and performance control of the singing voice inaugurate a new era for dealing with the understanding the voice quality and mechanisms during singing. Recent research in this field combine techniques for computational and digital technology, such as the analysis of the voice sound signal, with the physiological and acoustical study of the vocal instrument [13-14], the analysis of singing performance through emotions, as also the investigation

of the cognitive mechanism of singing, through the prism of behavioral and neurophysiological studies [15-17].

But in which way can we approach the understanding of the singing voice phenomenon through a more objective, definitive and unified way, combining different sciences methodologies? At the intersection of Music Informatics, Music Technology, Music Cognition and Music Acoustics, signal processing software tools have already made their appearance for the study, analysis, and understanding of the voice mechanisms.

However, although scientific studies are advancing, the empirical conservatory model of lyric singing pedagogy remains practically unchanging, with just a minor percentage of institutions utilizing innovative technologies and knowledge. Vocal trainers between them serve conflicting positions, opinions, and vocal goals. Yet, how can a music school or vocal student know which the objectively optimal method for vocal training is? It seems that singing pedagogy would greatly benefit from an update which will encompass methods, knowledge and tools from science and research, in order to avoid repeating prior mistakes. Almost 140 years of voice recording and reproduction equipment have passed, but this utility has not, yet, been incorporated into an organized singing educational system. Today, we have advanced to also have interactive digital systems, created for portable devices and computers, which are easily accessible to the average student and teacher.

III. ON VOCAL PEDAGOGY METHODS

Appelman [8], addresses the problem directly and suggests a possible solution as early as half a century ago: "Vocal pedagogy cannot survive as an independent educational entity if the physiological and physical facts, which comprise its core, remain subjects of sciolism (superficial knowledge). Researchers must uncover and interpret scientific facts about voice, voice quality and vocal pedagogy so that they might become realistic pedagogical tools that may be employed by future teachers of voice." It is, perhaps, time for development of more specialized tools, addressed specifically to the teacher and student of lyric singing.

A 2018 cataloguing of published research reports at least 754 studies on voice Acoustics since 1949 [18]. The same study concluded that «Until 2010, great importance was given to the voice quality of singers and their occupational demands. Acoustic analysis was widely used to study the effects of training. Since 2010, the concern with functionality is increasing, rather than the organic voice structures. Musical perception studies have been a trend, as well as the use of electroglottography.» On their 2011 review paper, Kob et al [13] report that voice research has advanced greatly during the last few decades, due to the development, or advancement, of tools like the electroglottograph (EGG) or the endoscope,

while also stressing the significance of the evolution of voice acoustic modeling.

Parallel to that, research on cognitive mechanisms during singing, has shown proof that professional lyric singers tend to have an increased volume of gray matter in the ventral primary somatosensory cortex of the brain, in particular at the centers corresponding to organs associated with the mechanism of the voice [15]. This, combined with the observation of a great activation of the above brain regions of the singers during singing, suggests a particularly large kinesthetic perception of the "vocal instrument" [17].

This ability seems to stem from the need of professional singers to gradually learn to rely more on somatosensory rather than their auditory feedback. This indicates that they learn to better control their vocal mechanism in order to create the physical conditions that will allow the production of the desired sound [15-17]. This level of control is necessary, as somatosensory control of voice onset is activated from the moment when sound preparation begins, while auditory feedback exists only if a vocal sound has already begun to be produced and we are then able to hear it. Professional lyric singers must be constant about pitch precision and quality accuracy of produced sounds. They cannot afford to wait for their sounds to be heard in order to then correct or change them. Furthermore, a direct correlation between increase of the above properties and the experience and level of vocal training of the lyric singer is evident [17].

IV. ARCHITECTURE OF THE FEEDBACK SOMATESENORY PLATFORM

The above research on singing cognition, acoustic analysis and understanding of the physiology of voice, can become the core components for creating new approaches to the human voice. These are the principal elements we employ in order to design our methodology at the team "Aedos" of the Laboratory of Music Acoustics and Technology at the University of Athens (LabMAT). During the last five years the team has developed several tools for the pedagogy of the singing voice, which have been tested in numerous cases.

Among the most popular of these software tools are 1) the micro-tone, multi-modal vocal pitch training system "Fonaskein" [19] and 2) "Match Your Own Voice!", a software which aim to bridge the gap between the vocal lesson and unsupervised vocal practice sessions [20]. This is achieved by means of real time visual representation comparison between a "reference", target vocal spectrum of the student (chosen by the professional vocal trainer for each note/vowel) and the same student's practice sounds. This visual feedback software tool has been experimentally evaluated and has appeared to accelerate a) a more precise kinesthetic control of the vocal mechanism and b) adoption of the "vocal technique" proposed by the trainer [20].

In the near future, we intend for these tools to be connected to an Artificial Intelligence recommender system, such as the Web 4.0 platform, to provide more features, such as pitch and lyric singing training.

The project we present here is a platform which will be designed to give the user the ability to “view” several aspects of the use of his/her vocal mechanism in real-time. For this platform LabMAT will collaborate with the “Athena” research and innovation center in information, communication and knowledge technologies. The innovation of this project is the use of multiple sensors and the cross-referencing of these different inputs in order to suggest guidelines for an alternative way of vocal production. A present, initial design of the platform adopts sensors such as:

- a precision microphone for acoustic analysis of the voice, giving feedback regarding the accuracy of formants tuning, vowel clarity, spectral energy, pitch correctness,
- cameras designed for the estimation of body and neck stance, lip/jaw movement and mouth aperture,
- an accelerometer for measuring diaphragmatic movement’s direction and speed,
- an Electroglottograph (EGG), for the non-invasive examination of the user’s glottal activity and open and close quotient.

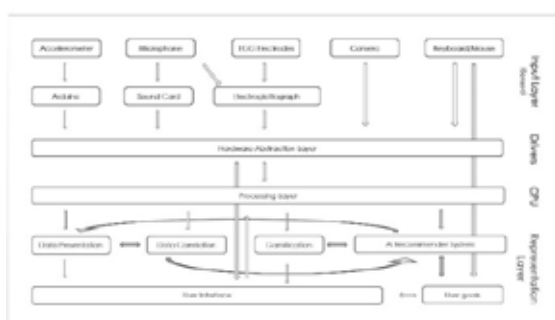


Fig. 1, Diagram of an Architecture Scenario for our Multi-Sensor Feedback Platform for Singers.

The user will be able to connect any number of the compatible sensors and receive data from them during singing tasks. The software will then be able to present the user with the data input from the available sensors through a multi-window interface, designed to help identify correlations between measurements from different measures. Also, AI suggestions on experimentation for possible improvements, such as different body alignment, pitch precision, formant tuning, spectral energy, glottal behavior, or even breathing habits, will be available at the user’s request. The platform will by no means pretend to have a knowledge of a “correct” vocal technique, but will be in the position to offer exercises, in the form of “games” or “challenges”. Such trials can help users find out what

helps them the most or measure the effects of a certain “vocal technique” modification in other measured parameters. For example, the user may notice that a different neck alignment helps in the production of more precise note pitch, or that a newly suggested mouth opening results in a clearer vowel or a denser vocal energy spectrum. Moreover, functions of “Match Your Own Voice” software will be implemented in similar tasks, designed according to “gamification theory” specifications for effective learning.

IV. RESULTS

The platform in discussion can find use in both research and pedagogy of the singing voice. In the research field it could be useful in studying correlations between a singer’s breathing technique, glottal behavior patterns and vocal spectrum results. Furthermore, additional accelerometers could be used to provide information for the movement of any number of other muscle groups, relevant to singing control, such as chest, shoulder, neck, back, pelvis, jaw or head. As a means of pedagogical assistance, the platform can be utilized in order to help users discover aspects of their singing mechanism that could be used in a more productive or healthier way. It can also find application as an instrument for the vocal teacher, to help explain and demonstrate vocal technique aspects to students, in a more objective and explicit way, leaving no room for speculation, misconception or misunderstanding of the vague singer terminology. It could likewise be used 1) as an agent that provides an objective comparative view of a singer’s progress in time, 2) as a guide to Extended Vocal Techniques experimentation, but could also be 3) a powerful somatosensory singing practice aid.

VI. DISCUSSION

Current software for the voice are addressed mostly to amateurs or children, yet the group of people that aspire to be professionals in the field of lyric singing do not have a large selection of specialized tools in their disposal. Lyric singing demands precise coordination of a large portion of the body and its functions and a high degree of accurate control over muscle groups which most people do not know how to willfully control. The above platform is intended to fill a part of this gap by having a wide range of functions, in order to better serve the additional needs of a student of lyric singing.

As with most technologies, a wider range of functions can bring on some restrictions. In this occasion the use of multiple sensors can dramatically increase cost and mobility (especially the use of an Electroglottograph). For this reason the user will have the choice to use any number of the supported sensors.

Parallel to the above, a separate project for the creation and experimental evaluation of another software

platform, comprising of interactive software tools for primary education students, which will promote somatosensory singing training for children, is already underway. This will hopefully help raise a new generation of young “singers”, who will then have an inherit advantage over current adults.

VII. CONCLUSION

We propose that a new, generation of technological software tools and environments be created [21], which will draw information from more inputs than solely the vocal acoustic signal. Such tools, which will be able to provide the user with a multi-faceted perspective of the voice, along with proper training of vocal teachers, could potentially have a drastic positive impact on vocal pedagogy and lyric-voice quality.

REFERENCES

- [1] Aristotle, *On voice, Opera omnia: Aristotelis Problemata. Vol. 14* (pp. 110-130). Lipsiae: O. Holtze, 1870.
- [2] Hippocrates, *Opera omnia*, ed. curavit C.G. Kühn. Vol. 26, 1825.
- [3] Galen, *Claudii Galeni opera omnia*. Kühn, K. G. (Ed.). Lipsiae : C. Knobloch, Vol. 17, 1821
- [4] K. Melidis, “The profession of ‘phonaskos’ as revealed in ancient inscriptions and medical texts,” *Rudiae Ric. sul mondo Class.*, vol. 22–23, pp. 28–30, 2010.
- [5] B. Coffin, *Historical vocal pedagogy classics*. Scarecrow Press, 1989.
- [6] L. Lehmann, *How to Sing*. The Macmillan Co, 1902.
- [7] G. B. Shaw, *How to Become a Music Critic*. Edited, with an introduction by Dan H. Lawrence. London: Rupert Hart-Davis. 1960. (p.99)
- [8] D. R. Appelman, *The science of vocal pedagogy: Theory and application* Indiana University Press. Vol. 378, 1967.(p.9)
- [9] C. L. Osborne, “Where Have All the Aidas Gone? The Crisis in Opera Training,” *Music Educ. J.*, vol. 66, no. 2, pp. 50–53, 1979.
- [10] A. Moravcik, “Where have the Great Big Verdi Voices Gone?,” in *Europaischen Musiktheater-Akademie Isolde Schmid-Reiter, Poetischer Ausdruck der Seele”: Die Kunst, Verdi zu singen. ConBrio Verlagsgesellschaft*, pp. 83–128, 2016.
- [11] Chapman, J. L. (2017). *Singing and teaching singing: A holistic approach to classical voice*. Plural Publishing. (p. xviii-xix)K. Norton, *Singing and wellbeing: Ancient wisdom, modern proof*. Routledge, 2015.
- [12] K. Norton, *Singing and wellbeing: Ancient wisdom, modern proof*. Routledge, 2015.
- [13] M. Kob, N. Henrich, H. Herzel, D. Howard, I. Tokuda, and J. Wolfe, “Analysing and understanding the singing voice: recent progress and open questions,” *Curr. Bioinform.*, vol. 6, no. 3, pp. 362–374, 2011.
- [14] J. Wolfe, D. T. W. Chu, J.-M. Chen, and J. Smith, “An Experimentally Measured Source--Filter Model: Glottal Flow, Vocal Tract Gain and Output Sound from a Physical Model,” *Acoust. Aust.*, vol. 44, no. 1, pp. 187–191, 2016.
- [15] B. Kleber, R. Veit, C. V. Moll, C. Gaser, N. Birbaumer, and M. Lotze, “Voxel-based morphometry in opera singers: Increased gray-matter volume in right somatosensory and auditory cortices,” *Neuroimage*, vol. 133, pp. 477–483, 2016.
- [16] B. Kleber, R. Veit, N. Birbaumer, J. Gruzelier, and M. Lotze, “The brain of opera singers: experience-dependent changes in functional activation,” *Cereb. Cortex*, vol. 20, no. 5, pp. 1144–1152, 2009.
- [17] J. M. Zarate, “The neural control of singing,” *Front. Hum. Neurosci.*, vol. 7, p. 237, 2013.
- [18] P. M. Pestana, V. F. Susana, and M. C. Manso, “Trends in Singing Voice Research: An Innovative Approach,” *J. Voice*, 2018.
- [19] F. Moschos, A. Georgaki, and G. Kouroupetroglou, “Fonaskin: An interactive software application for the practice of the singing voice,” in *Proceedings of the Sound and Music Computing (SMC) Conference*, pp. 326–331 (2016).
- [20] E. Angelakis, Kosteletos, G., Andreopoulou, A., & Georgaki, A. (2018, October). Development and Evaluation of an Audio Signal Processing Educational Tool to Support Somatosensory Singing Control. In *Audio Engineering Society Convention 145*. Audio Engineering Society.
- [21] V. Katsouros, E. Fotinea, R. Frans, E. Andreotti, P. Stergiopoulos, M. Chaniotakis, & Martín-Albo, D. (2018). iMuSciCA: Interactive Music Science Collaborative Activities for STEAM Learning. In *Designing for the User Experience in Learning Systems* (pp. 123-154). Springer, Cham

ASSESSING ARTICULATION DURING A VOWEL MATCHING EXERCISE USING ARTICULOGRAPHY: A PILOT STUDY

H. Daffern¹, A.J. Gully²

¹ AudioLab, Department of Electronic Engineering, University of York, York, UK

² Department of Language and Linguistic Science, University of York, York, UK
helena.daffern@york.ac.uk amelia.gully@york.ac.uk

Abstract: This paper examines the articulatory strategies utilised by choir singers to aid choral blend through the use of electromagnetic articulography (EMA). Whilst vowel matching is considered a beneficial acoustic strategy in achieving blend there is little known about the articulatory changes that occur in practice. Five choir singers were recorded singing sustained vowels with six EMA electrodes placed on the tongue and the lips. Subjects sang a sustained note, during which a pre-recorded stimulus (same pitch and vowel) was played over headphones to which they were asked to blend. In a control condition no stimulus was played. D4, F#4 and A4 to vowels /a, i, u/ were presented in each condition in a randomised order. Fundamental frequency and the first three formants were estimated from the acoustic signal, as well as the sensor positions. The absolute difference for these features between baseline and matching condition was measured across each subject, feature and condition. Results are highly variable with significant differences observed only for one subject. Patterns of increased variability in the matching condition are observed for most subjects, indicating that the singers were doing something to achieve vowel matching, although not with a consistency to make these findings significant.

Keywords: Singing, articulography, vowel matching, blend, choir

I. INTRODUCTION

A key performance goal for an individual singing in a choir is to ‘blend’ their voice into the texture and not be heard as an individual. In the context of choral singing, vowel production is considered an important contributing factor to choral blend: practitioners focus on ‘vowel matching’ when training and directing choirs based on the premise that ‘intonation and overall intelligibility rely on matching vowel sounds, uniformity is crucial to producing the kind of blend that results in a truly beautiful choral tone’ [1]. As described by Ternström [2] in a review of the limited research in this area, empirical studies also indicate that a unified vowel is desired for a good choral sound. Hunt asserts that “the problem of unity of vowel is one of intonation of formant frequencies” [3], and a

number of studies have assessed formant frequency changes in solo versus singing choral singing modes alongside other factors including fundamental frequency, intensity and evidence of a singer’s formant cluster [4, 5, 6, 7].

Goodwin conducted a study whereby individuals sang along and then ‘blended’ to a pre-recorded choral ensemble delivered over headphones. Results showed a stronger intensity of fundamental frequency and first formant and weaker second and third formant in a choral condition [8]. Studies conducted to date tend to focus on the acoustics rather than the articulatory strategies being used to achieve ‘blend’.

It is now possible to track points on the articulators (tongue, lips, mouth) in real-time as a participant sings, with the emergence of technologies including electromagnetic articulography (EMA) [9]. This study makes use of EMA and acoustic recordings to examine the alterations singers make as they perform a vowel matching exercise. This pilot study seeks to determine whether tongue height and lip rounding contribute to vowel matching between singers, with the additional aims of assessing whether EMA provides useful data for measuring articulation during singing, and whether these can be mapped to acoustic parameters.

II. METHODS

A. Recording set-up

Subjects were sat within a Carstens AG501 articulograph in a rectangular room with no acoustic treatment. In both recording scenarios, EMA sensors were placed in the midsagittal plane on the upper and lower lips (sensors UL and LL), and tip and back of the tongue (sensors TT and TB); in addition, a sensor was positioned on the side of the tongue (TS) and at the corner of the mouth (LC). A small lavalier microphone (DPA 4062-BM) was taped to the subject’s nose and open-back headphones were worn (Beyerdynamic DT990 Pro). Subjects were also fitted with electrolaryngograph (Lx) electrodes to allow parallel measurement of vocal fold activity (to be reported in a later study).

B. Target stimuli

Target stimuli were recorded of two biologically defined female singers, one novice and one classically trained experienced choral singer, sustaining vowels /a, i, u/ at three pitches (D4, F#4, A4) for five seconds. The recording set-up described above was used, except these singers did not wear headphones.

C. Subjects

Five experienced female choral singers aged 24-37, who did not record the stimuli, took part. All subjects sang in choirs with membership by audition at least weekly. Subject S2 regularly performs in professional choirs and S2, S3 and S5 each had two degrees in music with voice as their first study. Subject S4 had an undergraduate music degree but no formal singing training, and subject S1 had five years singing lessons but no formal music qualifications.

D. Tasks

The participants sang each vowel/note combination for six seconds. They were instructed on the vowel and given a reference pitch from an electronic pitch pipe before each note. Stimuli were presented at a level calibrated to match that of the target singer in the same space. The initial three seconds provided baseline data (condition B). After three seconds they heard either no external sound (condition M0), the novice singer (condition M1) or the trained singer (condition M2) and were asked to blend, with a focus on matching their vowel, to what they heard. Condition, note and vowel order were randomized.

E. Analysis

Approximately one second of audio (comprising an integer number of vibrato cycles) was selected from the steady-state portions of the recorded signal before (B) and during the test conditions (M0/M1/M2), along with corresponding EMA data. Fundamental frequency (f_0) and the first three formants (F_1 , F_2 , F_3) were calculated automatically using Praat [10] with heuristically-determined subject-specific parameters. In addition, 3D position of each EMA sensor (UL, LL, LC, TT, TS, TB) were calculated for each segment.

F. Ethics

Ethical approval was granted from the University of York Physical Sciences Ethics Committee.

III. RESULTS

Absolute differences were calculated between baseline measures (B) and test conditions (M0/M1/M2) to account for the fact that different vowels might require opposite articulatory matching strategies which

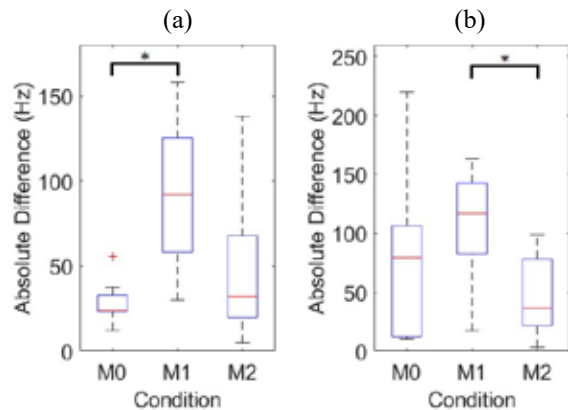


FIGURE 1: Differences in median F_3 for subjects (a) S1 and (b) S2.

could skew signed results. Results were not compared across subjects due to the small number of participants, and because subjects may use different strategies to achieve the same target.

Significance values reported below were calculated using a two-sample Kolmogorov-Smirnov (KS) test.

A. Acoustic Analysis

The absolute difference in mean f_0 between the test condition and the baseline ($f_{0,diff}$) was calculated for each sample. Significant differences in $f_{0,diff}$ were observed between conditions M0–M2 for subjects S1 ($p=0.0257$), S3 ($p=0.0005$), and S5 ($p=0.0239$), and for S4 between M0–M1 ($p=0.0187$). Variance of $f_{0,diff}$ was consistently higher in conditions M1 and M2 compared to M0 for all subjects.

There were no significant findings for absolute difference in median frequency of F_1 and F_2 for any subject, although there was an increase in variability of median formant difference in both matching conditions (M1/M2) compared to M0 for subjects S3 and S4. Significant differences were observed for median F_3 difference for subject S1 (between M0–M1, $p < 0.05$) and S2 (between M1–M2, $p < 0.05$), as illustrated in Fig. 1.

B. Articulatory Analysis

The absolute difference between the mean baseline position and test positions was calculated for all samples and sensors in 3 dimensions: x (posterior-anterior), y (left-right) and z (inferior-superior); e.g. the x position of the TT sensor is denoted TT_x . Some significant changes in sensor positions were observed in all subjects, although only for one sensor and one condition for subjects 3 and 4. Tables 1 and 2 present all significant differences in sensor position.

A common pattern observed for all subjects is an increase in variance in the matching conditions (M1/M2) compared to M0 for most of the features: the difference from baseline tends to be the smallest in M0

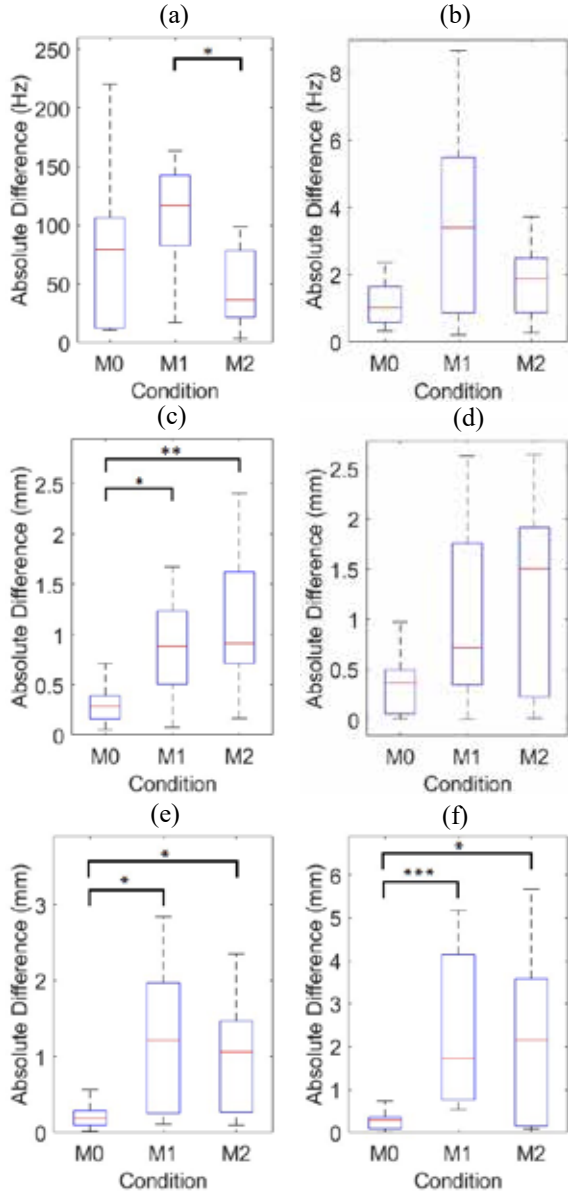


FIGURE 2: Difference from baseline condition for subject 2 for features: a) F_3 , b) f_0 , c) TT_x , d) TT_z , e) LL_x , f) LL_z .

with the smallest standard deviation compared to M1 and M2. This can be seen in Fig. 2, which profiles some of the significant findings observed for subject 2.

IV. DISCUSSION

A. Overall Findings

The higher variability observed in differences for most features in the matching conditions (M1/M2) compared to the control (M0) indicates that the subjects were responding to the stimuli in some way, supported by the significant differences observed. The combination of vowels and pitches included in the study may

TABLE 1: showing significant changes in lip sensor position (*= $p<0.05$, **= $p<0.01$, ***= $p<0.001$).

Sensor	S1	S2	S5
UL_x		M0-M1*	
LL_x		M0-M1* M0-M2*	M0-M2*
UL_y	M1-M2**		
LL_y	M0-M1* M0-M2**		
UL_z		M0-M1* M0-M2*	
LL_z	M0-M2*	M0-M1*** M0-M2*	
LC_x		M0-M1*** M0-M2***	M0-M2**
LC_y	M0-M2* M1-M2*	M0-M1* M0-M2**	M0-M1**
LC_z		M0-M1* M0-M2**	

TABLE 2: Significant changes in tongue sensor position (*= $p<0.05$, **= $p<0.01$, ***= $p<0.001$).

Subject	TT_x	TT_y	TS_x	TS_z
S1			M1-M2*	
S2	M0-M1* M0-M2**			
S3		M0-M1*		
S4			M0-M2*	
S5			M0-M2*	M0-M1**

contribute to this increased spread. Owing to few subjects it was not possible to interrogate the data at the vowel and pitch level with any confidence.

The lack of significant movement of the back tongue sensor reflects the lack of significant difference in F_2 . Combined with the stability of F_1 this might suggest that whilst being instructed to ‘blend with a focus on matching the vowel’, in fact the articulatory and acoustic positioning of vowels is idiomatic to the subject and that features other than formant position are utilised to achieve blend.

The significant change in f_0 identified for all subjects except S2 in matching conditions might suggest that these singers are using phonatory rather than articulatory strategies, and focussing on pitch as their main blending ‘feature’.

B. Subject 2

Subject 2, the most experienced singer, is an interesting case study; unlike the other singers, they produced significant differences for most articulatory features but not for f_0 .

The very stable F_1 values across condition for this subject are unexpected considering the significant movements of the lips, which might suggest a change in vocal tract length (which would in turn be likely to move the first formant). This might be explained by the

movements being too small to greatly impact the formant, but they may still contribute some small timbral change. The results for lip movement may also be most salient for the /u/ vowel due to the lip rounding involved in its production.

The differences observed for F_3 between M1–M2 but not M0 are the only observations indicating a significant difference *between* the matching conditions for S2. This suggests that whilst the subject made alterations to match the stimuli, it is likely that the nature of the movements was different for each stimulus, perhaps resulting in a shift of F_3 . This is unsurprising due to the differences between the recorded stimuli; they were recorded as solo performances and therefore the stimuli presented in condition M2 (trained singer), have a deep, regular vibrato and were louder than those presented in M1 (novice singer).

C. Limitations

The greatest limitation of this study is the sample size. In addition to more subjects, it would be useful to collect repetitions of the same tasks from the same subjects. Due to the time-consuming protocol of attaching sensors to the subjects, and the limited time that sensors remain attached, this wasn't possible using the current protocol (which required 1.5 hours per subject). It is likely that the differing backgrounds of the singers impacted the results - the professional subject (subject S2) presented the most significant articulatory differences during matching, but without richer data it isn't possible to theorise that this may be due to training.

A constant high-pitched hum could be heard over the headphones once in the articulograph due to the magnetic field. However, the use of headphones was necessary to isolate the stimulus recording from that of the participants. Subjects were given time to listen to the background noise and assess its impact on their performing the task. Despite claims that it wasn't a problem this cannot be ignored as a mitigating factor.

The room in which the experiment took place is not acoustically treated and is quite 'lively'. To mitigate against this to some extent, the stimuli were also recorded in the same room and the microphone was mounted as close as possible to the performers' mouths (attached to the nose) to minimise the effect of the room on the recordings.

Formant extraction was performed automatically in Praat, using subject-specific settings to obtain the best performance possible. However, certain pitches present challenges for tracking the first formant as the first harmonic may be identified instead; furthermore, the possibility that the singers were using formant tuning strategies cannot be ruled out.

For this pilot study, sensor positions were chosen to explore a range of articulatory areas that might be of interest for future study. In future, sensor placement will be focussed on specific areas of articulation; for instance, using four sensors on the midsagittal line of the tongue, working backwards from tip as a reference point. This would also improve the accuracy of the placement of the sensors on the correct point on the tongue, which was quite variable across subjects in the current study due to there being no absolute point of reference.

V. CONCLUSION

EMA has been shown to be a viable and promising tool in assessing articulatory positions in vowel matching, however numerous complications need to be considered, especially the delivery of the vowel matching stimuli and choice of sensor positions.

REFERENCES

- [1] J. Tschiggfrie, *Teaching Tricky Diphthongs to Your Choir*. 2015. Accessed on 06.29.2019. Available: <https://www.smartmusic.com/blog/teaching-tricky-diphthongs-to-your-choir/>.
- [2] S. Ternström, "Choir acoustics: an overview of scientific research published to date," *Int J. Res. Choral Singing*, vol.1, pp.3-12, 2003.
- [3] W.A. Hunt, 'Spectrographic analysis of the acoustical properties of selected vowels in choral sound', EdD thesis, North Texas State Univ. 1970.
- [4] T. D. Rossing, J. Sundberg, and S. Ternström, "Acoustic comparison of voice use in solo and choir singing," *J. Acoust. Soc. Am.*, vol. 79(6), pp.1975-1981, 1986.
- [5] S. Ternström, and J. Sundberg, "Intonation precision of choir singers," *J. Acoust. Soc. Am.*, vol. 84(1), pp.59-69, 1988.
- [6] S. Ternström, and J. Sundberg, "Formant frequencies of choir singers," *J. Acoust. Soc. Am.*, vol. 86(2), pp.517-522, 1989
- [7] S. Ternström, "Physical and acoustic factors that interact with the singer to produce the choral sound," *J. Voice*, vol. 5(2), pp.128-143, 1991
- [8] A. W. Goodwin, "An acoustical study of individual voices in choral blend," *J. Res. Music Educ.*, vol. 28(2), pp.119-128, 1980.
- [9] P. W. Schönle, K. Gräbe, P. Wenig., J. Höhne., J. Schrader, and B. Conrad, "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," *Brain Lang.*, vol. 31(1), pp.26-35, 1989.
- [10] P. Boersma, "Praat: a system for doing phonetics by computer," *Glott International*, vol. 5(9/10), pp. 341-345, 2001.

TUNING TENDENCIES IN A SINGING QUINTET: EVOLUTION ACROSS REHEARSALS

S. D’Amario¹, D. M. Howard²

¹ Department of Electronic Engineering, University of York, York, UK

² Department of Electronic Engineering, Royal Holloway University of London, Egham, UK
Sda513@york.ac.uk; david.howard@rhul.ac.uk

Abstract: Tuning for musical instruments without a fixed pitch and for the human voice can vary note by note. A recent investigation analysing the evolution of tuning tendencies in a semi-professional singing quintet during a four-month study period demonstrates that the ensemble tended consistently to equal temperament tuning rather than just intonation. This study reports on a follow-up investigation, which aims to test the previous results, but this time with the same ensemble but a different piece that varies in rhythmical, harmonic, and melodic content. The ensemble rehearsed the specially arranged piece for this study across five rehearsal sessions with three repeated performances of the piece before and after each rehearsal. Audio and electrolaryngograph data of the repeated performances were collected to allow fundamental frequency evaluation of the individual voices. Results corroborate the previous study in that horizontal tuning was consistently closer to equal temperament rather than just intonation across repeated performances and rehearsals. This study suggests that tuning tendencies might typify singing ensembles and be unrelated to the musical characteristics of the piece being performed.
Keywords: Tuning, intonation, ensemble singing, equal temperament, just intonation

I. INTRODUCTION

Tuning is an important feature of music performance in a *cappella* singing ensembles as it contributes to musical excellence. Previous studies, mostly focussed on the analysis of intonation tendencies in single performance sessions, provide often inconclusive and contrasting results regarding the intonation system adopted during ensemble performances. A few studies observed that intonation was close to just intonation predictions, in which intervals result from the series of harmonics, in an *a cappella* singing quartet [1]. Conversely, other empirical investigations found that intonation tended to be closer to equal temperament, in which the octave is divided in 12 semitones of equal size, rather than just intonation predictions in 3-part

vocal ensembles performing a chord progression composed by renaissance theorist Benedetti [2]. More recently, a longitudinal study analysing tuning in a semi-professional singing quintet during five rehearsal sessions across a four-month study period observed that tuning in each singer was closer to equal temperament, than just intonation; and, these findings were consistent within and between performances and across rehearsals [3]. The present study aims to shed some light on intonation in singing ensembles, and test whether the tuning tendencies observed recently in the singing quintet differ with the piece being practiced.

II. METHODS

Participants: A newly-formed, semi-professional singing ensemble, formed as a regular quintet working towards performances and Masters exams, was recruited for the study (3 females, age Median = 23, Range = 6). The ensemble comprised a soprano (singer 1, S1), mezzo 1 (singer 2, S2), mezzo 2 (singer 3, S3), tenor (singer 4, S4), and bass (singer 5, S5). They were postgraduate students in ensemble singing at the Department of Music of the University of York.

Material: The ensemble rehearsed a chorale by J. S. Bach arranged for this singing ensemble by the first author. This piece is Piece B in D’Amario et al. [4]. Entries were systematically manipulated so that, except for the first simultaneous entry, each singer had one occasion to start the phrase on an up-beat ahead of the others. The arranged piece comprises eight phrases, to be sung legato to the vowel /i/.

Apparatus: The study took place in a recording studio of the Department of Electronic Engineering at the University of York (UK). The ensemble stood in a semi-circle of approximately 1.5 m radius, in part order S1, S2, S3, S4, and S5. Singers wore head-mounted close proximity microphones (DPA 4065), placed on the cheek of the singer at approximately 2.5 cm from the lips. Singers also wore electrolaryngograph electrodes from Laryngograph Ltd. (www.laryngograph.com), placed on the neck at the level of the larynx, either side of the thyroid cartilage. In addition, stereo and video recordings of the repeated performances were collected using a stereo condenser microphone (Rode NT4) and a

video camera (Sony MV1 Music Video recorder), respectively. The 12 outputs were then recorded in synchrony using a digital audio workstation (Reaper 5.40), set at 24-bit depth and 44.1kHz sampling frequency.

Design: This is a longitudinal study involving five rehearsal sessions across a four-month study period. The ensemble practiced the same stimulus for 10 minutes in each rehearsal; singers performed three repetitions (takes T1-T3) of the same pieces before and after each rehearsal (pre-post), resulting in 6 recordings per session. An additional piece, mostly contrasting in rhythmical content, was also similarly rehearsed and repeated performances collected, to investigate tuning outcomes and interpersonal synchronization between singers, and results reported in D’Amario et al. [3, 4]. The order of recording and rehearsing the two pieces was randomized within rehearsals.

Procedure: The ensemble was recorded between September 2017 and January 2018. Prior to the first rehearsal, participants received written and spoken instructions, and gave written informed consent. They were not aware of the purpose of the study, and were asked to work on expressiveness, towards a final public performance. This was designed to motivate and stimulate the singers during rehearsals, and record any tuning behaviour that could emerge spontaneously with practice, without explicitly asking for mastering tuning.

Analysis: A spreadsheet was prepared in which frequency ratios against the tonic of the first chord were calculated for both equal tempered and just tuning. Using the fundamental frequency of that tonic for a given performance enables direct tuning comparisons relative to equal temperament and just intonation to be made for each performance.

Three metrics of horizontal tuning were measured: i) pitch drift, as quantified by the fundamental frequency deviation from predicted models of just intonation and equal temperament; ii) intonation consistency, as quantified by the standard deviation (SD) of measured deviation from equal temperament; and, iii) intonation dispersion, as quantified by the range of measured deviation from equal temperament. The horizontal analysis was based on the whole set of notes of the piece performed.

Multilevel linear-models of the response variables (i.e., f_0 deviation from predicted values, SD and range of measured deviation) were implemented to test the primary fixed effects of rehearsal, and the fixed effects of rehearsal stage (i.e., pre-rehearsal and post-rehearsal) nested within rehearsal. Take, note and singer number were entered as random variables in the models. A Bonferroni correction was implemented for multiple multilevel linear modelling, and a P value threshold was set at $p = 016$.

III. RESULTS

A. Pitch drift

Visual inspection of tuning tendencies measured against ET and just intonation predictions indicates that tuning was closer to ET rather than just intonation. This was a distinctive behaviour, repeatable across rehearsals (R1-R5), repetitions (T1-T3), and singers (S1-S5). Fig. 1 shows examples of the tuning tendencies of singer 1 (soprano) in rehearsal 1, which demonstrate that the singer tended towards ET across repetitions. Complete pitch-drift analysis for each singer/note/take/rehearsal can be found online at doi.org/10.5281/zenodo.3406091. In light of these results, the inferential analysis was based on deviations from ET predictions, rather than just intonation.

As shown in Fig. 2, results from the multilevel linear modelling demonstrate that, compared with rehearsal 1, measured deviation from ET was sharper in rehearsal 2 [$\beta = 21.3$, $t(6375) = 14$, $p < 0.001$], rehearsal 3 [$\beta = 8.4$, $t(6375) = 5.5$, $p < 0.001$], rehearsal 4 [$\beta = -4.9$, $t(6375) = -3.2$, $p < 0.01$], and rehearsal 5 [$\beta = 26.7$, $t(6375) = 17.5$, $p < 0.001$].

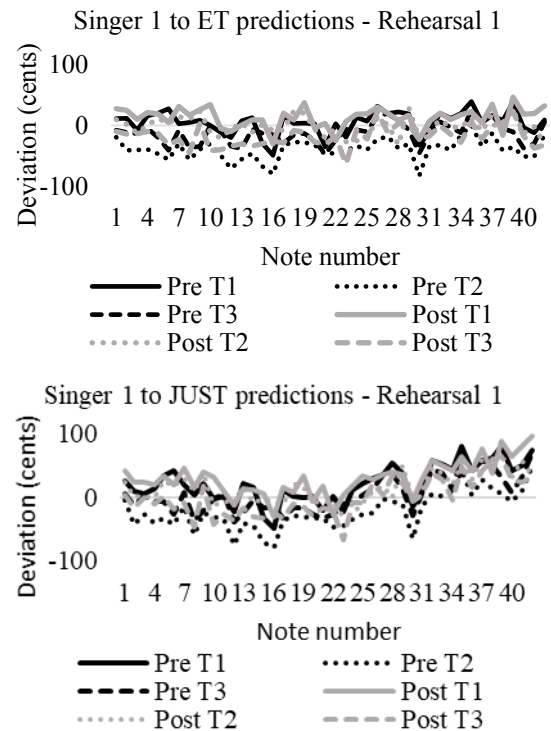


Fig 1. Measured deviation from equal temperament (ET, top) and just intonation predictions (JUST, bottom) of singer 1 (soprano), computed for each note (notes 1-42), repeated performances (T1-T3), and stage (Pre-Post), during the first rehearsal. Notes are normalized to

the first tenor note, F3, which is the tonic of the piece. Maximum and minimum values of the y-axes have been fixed for comparisons between the graphs.

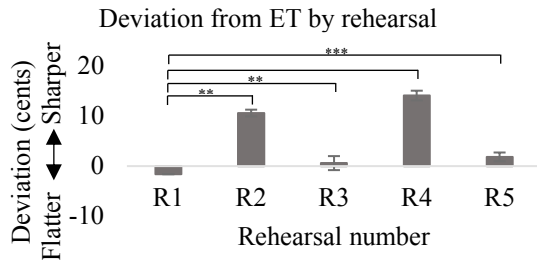


Fig. 2. Deviation from equal temperament prediction (ET) by rehearsal number (R1-R5). Error bars represent 95% CI of the mean. ** $p < 0.01$; *** $p < 0.001$.

B. Tuning consistency

As shown in Fig. 3, results from the multilevel linear modelling, based on SD of measured deviation from ET, show that tuning was more consistent in the final rehearsal session (R5) compared with the first one (R1) [$\beta = -4.8$, $t(134) = -2.8$, $p < 0.01$].

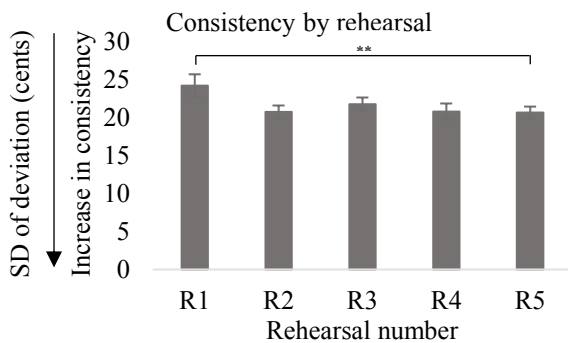


Fig 3. Consistency of tuning deviation from equal temperament (ET), by rehearsal number (R1-R5). Error bars represent 95% CI of the mean. ** $p < 0.01$.

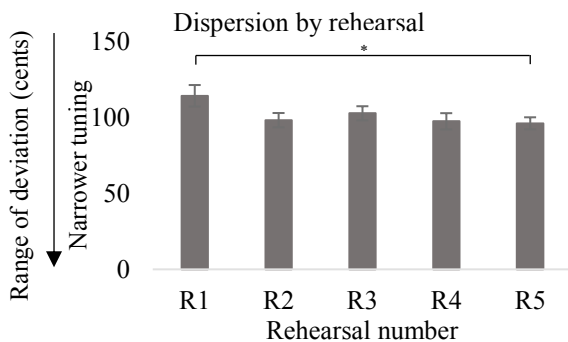


Fig 4. Dispersion of tuning deviation from equal temperament (ET), by rehearsal number (R1-R5). Error bars represent 95% CI of the mean. * $p < 0.016$.

C. Tuning dispersion

As shown in Fig. 4, the analysis of the range of tuning deviation from equal temperament system demonstrates that the range of tuning was narrower in the final rehearsal (R5), compared with the first rehearsal (R1) [$\beta = -22.5$, $t(136) = -2.5$, $p < 0.016$].

IV. DISCUSSION

Overall, during and between rehearsals and across repetitions, the ensemble performed closer to equal temperament tuning thereby avoiding pitch drift based on just intonation predictions. This result corroborates finding reported by D'Amario et al (2018), suggesting that this overall tuning tendency might typify the tuning strategy adopted by this ensemble and be independent of any specific musical characteristics of the piece performed. Deviation from equal temperament was sharper with practice across the term of study, compared with initial rehearsal. Deviation from equal temperament was also more consistent and narrower in the last rehearsal, compared with the first rehearsal. These results corroborate D'Amario et al. (2018), observing that f_0 measured deviation from equal temperament was more consistent and narrower during the course of study. Nevertheless, the role of each rehearsal seems to change according to the piece performed: the consistency and range of tuning dispersion changed significantly in the third rehearsal with the clearly homophonic piece in D'Amario et al. (2018), but it changed at the end of the term of study when rehearsing the more complex piece of the current study.

For greater ecological validity, future investigations should replicate this study with different ensembles, to test whether these tuning outcomes are generalizable across ensembles. In addition, singers in the study performed the piece to the vowel /i/. It would be of interest for further investigations to test systematically the effect of the different vowels on tuning outcomes.

V. CONCLUSION

This investigation described a longitudinal study of tuning tendencies in a semi-professional singing quintet across five rehearsal sessions during a four-month study period. Each singer showed an overall horizontal tendency to tune closer to equal temperament tuning compared to just intonation. They avoided any major pitch drift during the piece and this tuning behaviour was consistent across both repetitions and rehearsals.

Pitch accuracy represents one core element of excellence in ensemble singing, and increasing knowledge relating to how pitch accuracy develops over

time has the potential to provide choir directors and singers with strategies to use when working on tuning in ensemble singing.

REFERENCES

- [1] D. M. Howard, "Equal or non-equal temperament in a cappella SATB singing," *Logoped. Phoniatr. Vocol.* vol. 32, pp. 87–94, (2007). doi: 10.1080/14015430600865607
- [2] J. Devaney, M. Mandel, and I. Fujinaga. "A study of intonation in three-part singing using the Automatic Music Performance Analysis and Comparison Toolkit (AMPACT)," in *13th International Society for Music Information Retrieval Conference ISMIR*, pp. 511–516, 2012.
- [3] S. D’Amario, D. M. Howard, H. Daffern, and N. Pennill. "A longitudinal study of intonation in an a cappella singing quintet," *Journal of Voice*, in press. (2018). <https://doi.org/10.1016/j.jvoice.2018.07.015>.
- [4] S. D’Amario, H. Daffern, and F. Bailes. "A longitudinal study investigating synchronization in a singing quintet." *Journal of Voice*, in press. <https://doi.org/10.1016/j.jvoice.2018.07.015>.

FACE VIBRATION MEASUREMENT IN SINGING — PILOT STUDY

M. Frič, P. Dlask, V. Hruška

Musical Acoustics Research Centre, Music and Dance Faculty, Academy of Performing Arts in Prague, the Czech Republic

marek.fric@hamu.cz; pavel.dlask@hamu.cz; hruska.viktor@hamu.cz

Abstract: This study documents the possibility of using laser interferometry (ESPI) to monitor face surface oscillation based on measurements on one subject (M, 41 y. o.). The measurement was performed in three states of mutual position of the first formant and fundamental frequency. Synchronous measurements of facial vibrations employing an accelerometer confirmed the dominance of the harmonic component closest to the position of the first formant in the entire area of the head and neck. In addition to the amplitude ratios of facial surface oscillations, ESPI also distinguished the phase differences. Standing waves dominate in the oscillations of the face. The area around the corners of the mouth always oscillated most strongly and coincided with the oscillation of the upper forehead and neck. The areas around the eyes, chin, and ear vibrated in antiphase. Above the lower jaws, there was a running wave that progressed from the chin towards the angle of the mandible.

Keywords: Singing, head vibration, accelerometer, laser vibrometry, vibration velocity

I. INTRODUCTION

So far, vibrations of the head and neck surface have been measured only by means of accelerometers, laser vibrometers [1], or high-speed cameras [2]. Different oscillation regions on the face were described regarding the pitch, the vowels, and the voice register [1, 3]. Studies dealt with the energy of oscillations or their spectral representation [4]. However, the phase relations between the vibrational regions have not been thoroughly investigated.

Our previous study [5] showed the importance of the phase differences between the glottal oscillations (assessed by electroglottography [EGG]) and the vibrations of the larynx surface when resonance tubes are employed.

This study presents the amplitude and phase relations of facial vibrations. The measurements were based on electronic speckle pattern interferometry (ESPI).

II. METHODS

Preliminary measurements included singing scales using all Czech vowels in the forte dynamics of one male subject (41 y. o., amateur singer). A synchronized measurement of the EGG and the accelerometric signals (larynx, the center of the cheek) was analyzed employing wavegrams [6]. The singing of different vowels using accelerometers showed the dependence of the phase vibration of the larynx and cheek on the pitch and the type of vowel.

Based on preliminary experiments, we chose two different pitches and two vowels, at which we assumed the occurrence of phenomena interesting for a pilot ESPI measurement (see in Fig. 1):

- The vowel /u/ and the pitch F4 (347 Hz) [H1 mode] was chosen as a tone where the conjunction of the first formant and first harmonic could be observed.
- The vowel /a/ and the pitch F4 was in M1 [H2 mode], where the second harmonic was the dominant frequency of the larynx and cheek tissue vibrations, and it lies in the vicinity of the first formant.
- The vowel /a/ and the pitch A3 (220 Hz) in M1 mechanism [H3 mode], where the third harmonic was the dominant frequency of the larynx and cheek tissue vibrations.

The DANTEC DYNAMICS PulseESPI System Q-600 interferometer was used for the ESPI measurements [7] (noncontact full field analysis of dynamic phenomena). The subject had to be fixed in a stable position relative to the camera (1.2 m) and 0.3 m from the microphone (see in Fig. 2). The measurement plane was at a 45° lateral angle relative to the cranial sagittal plane. The zero-crossing of the EGG signal served as a reference for the triggering the signal governing the laser pulses.

The period of oscillation was divided into 18 phases (for pitch F4 steps 159 μs, interferometric pulse interval 100 μs, for A3 253 and 158 μs respectively). Each phase was measured ten times, and the mean values were used for the subsequent data treatment.

The measured data were processed employing the Istra software. The relative displacements for the time between the laser pulses were calculated.

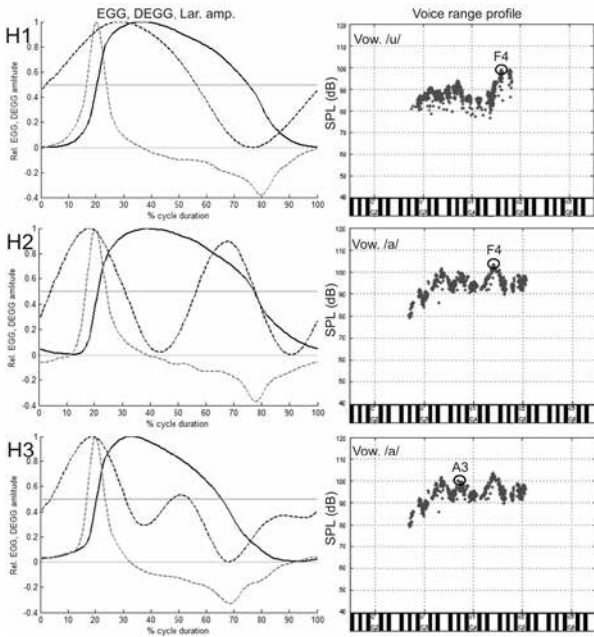


Fig. 1 H1, H2, and H3 modes from preliminary measurement. EGG, DEGG and amplitude of the surface neck accelerometer (laryngeal position) wavegrams of one period. Voice range profile.

The maps of the measured values of the relative amplitudes of the individual phases were further processed by analysis of the principal components (biorthogonal decomposition - BOD) [8], which revealed the best time covariate areas of the face.

Acoustical, EGG, accelerometric (larynx, cheek and forehead), and the Q-600 synchronization signals of all utterances of the ESPI measurement were synchronously recorded. The measured period of the EGG was found based on the synchronization signal. The instantaneous f_0 and phases of the observed period were calculated based on the distance of the first derivative of EGG (DEGG) signal peaks from the synchronization peak. Parts of the monitored signals belonging to the period found and the surrounding two periods were used for further processing. The average signal values were calculated from all measurements (for the accelerometers from the first integration with respect to time).

III. RESULTS

Every ESPI measurement at a single pitch required a total of 190 repetitions (in total lasting ca. 1.5 hour). In the post-processing analysis, the phase obtained and the repeatability of the measurements were tested against the EGG signal. Tab. 1 shows the average values of all repetitions for f_0 and SPL.

Fig. 3 shows the results of all measurements. In the first line, there are the Welch spectra from sound

recordings. The first harmonic component dominates in the vowel /u/ at the pitch F4, the first formant corresponds with the second harmonic for the vowel /a/ at the pitch F4, and with the vowel /a/ A3, the highest level is at the third harmonic.

Tab. 1 Average values and standard deviations of the fundamental frequency and SPL for measured modes

Mode	H1 - /u/ F4		H2 - /a/ F4		H2 - /a/ A3	
Par.	f_0 (Hz)	SPL (dB)	f_0 (Hz)	SPL (dB)	f_0 (Hz)	SPL (dB)
Mean	353.0	95.8	349.1	102.6	218.4	97.4
STD	2.97	1.49	2.92	2.6	15.72	4.52

Variations of the surface vibrational velocity throughout the whole period (RMS value) were calculated from the averaged relative amplitude (velocity) maps from ESPI measurements from all phases. The results are shown in the second row of Fig. 3. The dark portions show areas with relatively large velocity changes over the period, while the white portions indicate the positions of the nodes.



Fig. 2 Measurement settings. A) Detail from measurement camera with highlighted analytical area (mask). B) Position of ESPI and body with head holder and microphone.

BOD showed two essential components in all measurements. The first component always contained more than 98% of the data variability. The frequency of the oscillations of the first component matched the frequency of the dominant harmonic component of the particular tone. Phase maps of the first component (see the 3rd row in Fig. 3) are very similar among all measurements. They characterize the standing wave of facial oscillations. In the first component the vibrations close to mouth corner dominate, and the area through the face toward the ear. In the first component, the vibrations of the anterior part of the neck and upper part of the forehead are in phase. The chin and the part of the face around the eye from the mouth to the ear were oscillating in antiphase.

The face area from the chin to the ear (following the mandible) always belongs to the second component (the fourth row in Fig. 3). In F4 measurements, this component was phase-shifted by $\pi/2$ (see the fifth row

in the Fig. 3), and a hint of a running wave could be seen in a time sequence of the measured movement phases.

Some specific relative movement amplitudes (velocities) of the front part of the neck, cheek, and forehead measured by ESPI in the same positions as measured by accelerometers are depicted in the 6th row in Fig. 3. Basically, these vibrations well coincide with the accelerometer measurements (see the 7th row in Fig.3).

IV. DISCUSSION

The ESPI measurement was used for measurement of facial vibration for the first time (to the best of our knowledge) in this study. For the subject it is significantly time-consuming and mentally demanding, because it requires a large number of repetitions of the same tone with the same quality (with the head fixed in position and the eyes covered). The repetitions are required to eliminate the pitch and SPL inaccuracies that are caused by natural vibrato during singing.

The resulting data can only be compared with a small number of studies that measured the face movement velocity using single-beam scanning laser vibrometers. Our results are very consistent with the measurements made in [1, 9] in the whole facial area and in the forehead area in [10] (even though the studies measured females). However, in addition to the facial oscillation energy ratios, our measurement provides information on phase differences.

In contrast to [4], velocity energies in the cheekbone and forehead and upper neck were higher when producing the vowel /u/ than the vowel /a/. However, in [4], the speaking voice was measured. In our case, it is necessary to consider the formant tuning effect during singing.

V. CONCLUSION

The present study documents the possibility of using laser interferometry to monitor surface oscillations of the face. The initial results confirmed the presence of the dominant harmonic component in facial oscillations. BOD analysis showed in particular standing waves on the face that were very similar for all measured pitches and vocals. Because the accelerometric measurements show the relationship of

the vibration phase to the resonance events, the use of resonated voice could be better documented by this method.

REFERENCES

- [1] Kitamura T. Measurement of vibration velocity pattern of facial surface during phonation using scanning vibrometer. *Acoustical Science and Technology*. 2012, 33(2):126-128.
- [2] Frič, M, Fikejz F. Correlation between Three Facial Vibration Measurement methods. In: *International Conference on Applied Electronics, Pilsen 8-9 September 2015*, Pilsen: University of West Bohemia, 2015, pp. 41-44.
- [3] Chen FC, Ma EP, Yiu EM. Facial bone vibration in resonant voice production. *J Voice*. 2014, 28(5):596-602.
- [4] Munger JB, Thomson SL. Frequency response of the skin on the head and neck during production of selected speech sounds. *J Acoust Soc Am*. 2008 Dec;124(6):4001-12
- [5] Frič M, Hruška V. The effect of resonance tubes on facial and laryngeal vibration – A case study. *Biomedical Signal Processing and Control*. 2017, 37(Aug):50-60.
- [6] Herbst Ch. T., Fitch W. T. S., Svec J. G. Electroglottographic wavegrams: a technique for visualizing vocal fold dynamics noninvasively. *J. Acoust. Soc. Am*. 2010, 128(5): 3070-3078.
- [7] Jones, R., Wykes, C. *Holographic and speckle interferometry: a discussion of the theory, practice, and application of the techniques*. 2nd ed. New York: Cambridge University Press, 1989.
- [8] Taira K., Brunton S. L., Dawson S. T. M., Rowley C. W., Colonius T., McKeon B. J., Schmidt O. T., Gordeyev S., Theofilis V., Ukeiley L. S. Modal Analysis of Fluid Flows: An Overview. *AIAA Journal*, 2017, 55(12):4013-41.
- [9] Kitamura T., Hatano H., Saitou T., Shimokura Y. Haneishi E., Kishimoto H. A pilot study of vibration pattern measurement for facial surface during singing by using scanning vibrometer. In: *Proceedings of the Stockholm Music Acoustics Conference 2013, SMAC 2013*, Stockholm, Sweden, p. 275-278.
- [10] Kitamura T., Ohtani K. Non-contact measurement of facial surface vibration patterns during singing by scanning laser Doppler vibrometer. *Front. Psychol.*, Nov. 2015, 6: 1-8.

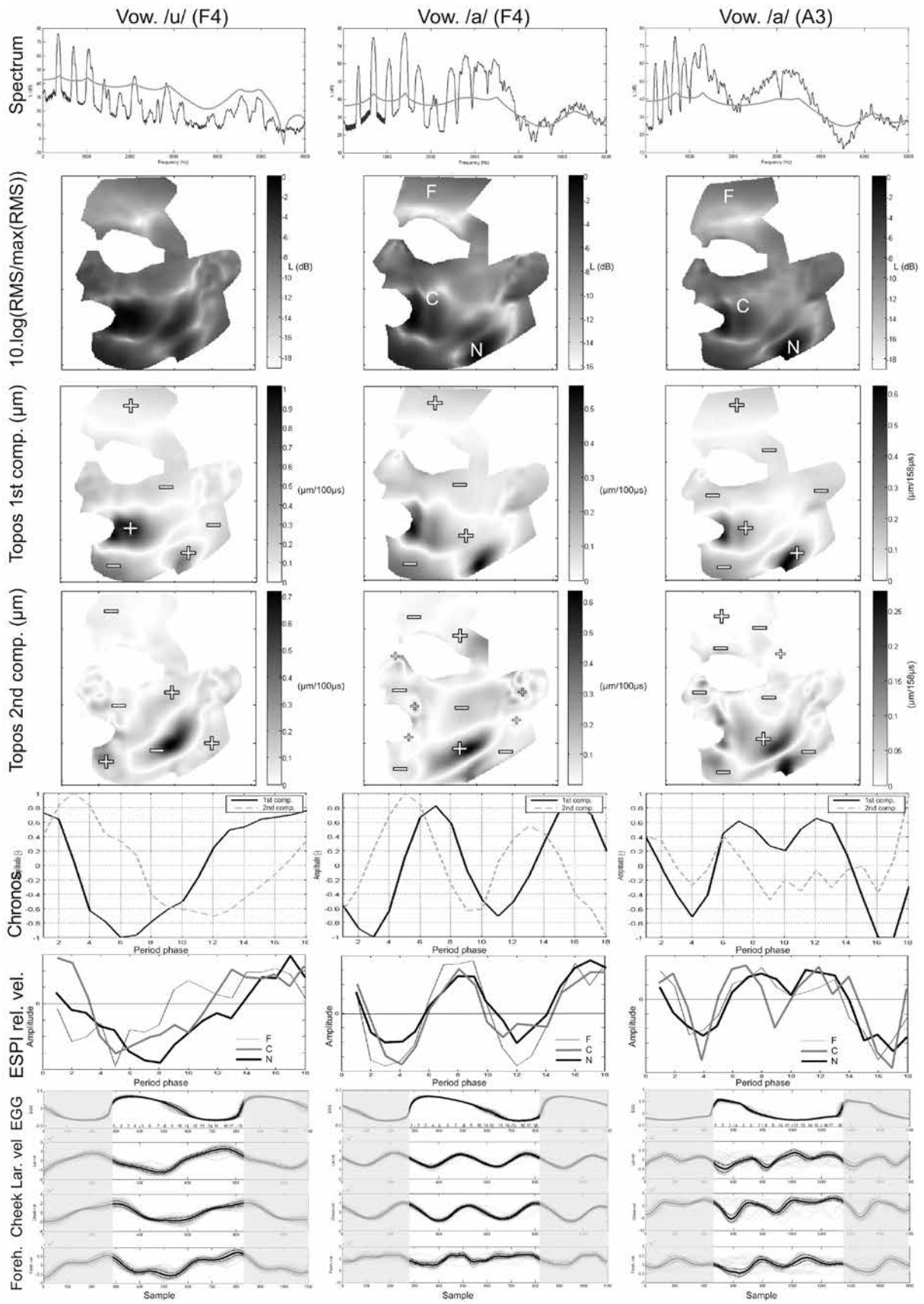


Fig. 3 Results of all measurement.

COMPARISON OF SOUND RADIATION BETWEEN POP AND CLASSICAL SINGERS

I. Podzimková, M. Frič

Musical Acoustics Research Centre, Music and Dance Faculty, Academy of Performing Arts in Prague, The Czech Republic

iva.podzimkova@seznam.cz, marek.fric@hamu.cz

Abstract: The directional characteristics of a soprano opera and pop singer in the horizontal and vertical planes were measured in an anechoic room. RMS pressure measurements were made across the whole frequency range and also in frequency bands at forty eight positions around the head at a distance of 136 cm while the singer's mouth was the center. The main aim was to describe differences in sound radiation between both singers in relative scales of sound pressure level calibrated to the reference microphone.

For deeper description, the experiment uses differences between vowels and also attempts to obtain results for the significant spectral range of 2,5 – 4 kHz. It also shows whether there is a difference in distribution of the sound while the pitch is changing. The results were compared with the only other two studies which corresponded with these types of results. It could not be compared properly because a study describing differences between singers depending on the genre was not found in the literature.

The directionality of each singer is different and differences between vowels, also in range 2,5 – 4 kHz, were found. This preliminary experiment shows that sound radiation is different while changing pitch.

Keywords: Singing, voice radiation, singing genres

I. INTRODUCTION

Only a few studies have been published about the radiation of the human voice. The first experiments showed frequency dependence in the direction of radiation of the male voice [1]. Furthermore, the characteristics of a live speaker and a modeled mannequin were compared and a significant emission was found in the direction 180° from the frontal direction.

As technology has advanced, loudspeakers have emerged as alternative sources of the human voice [3]. There are no substantial differences in radiation between men and women, even among a group of English and French speakers [4]. New methods of

measurement were developed and focused on the singing voice [5, 6]. They confirmed that directional radiation depended on the size of the head and shape of the mouth, and that vocal projection did not affect directionality. They confirmed that the differences in directional radiation between vowels were apparent only in speech, not singing. A study focused on opera singing confirmed that singers adapt their directionality in accordance with the space, specifically for larger spaces by increasing the radiation pattern to 180° [7].

Although some of these studies already deal with singing, they do not deal with the issue of different directional characteristics of singers of different genres. Such was done in this preliminary work, the aim of which is to describe radiation differences between female soprano pop and opera singers, to undertake a comparison between the different directional characteristics of vowels, and to determine the dependence of the radiation on pitch.

II. METHODS

The directional characteristics of two female singers of different genres (classical and pop singers) were measured in anechoic room (critical frequency 125 Hz ISO 3475, 5.26 x 3.75 x 2.25 m). Measurements were made using a spherical net with a radius of 1.36 m and with 48 microphones (body Sennheiser K6 and capsule Sennheiser ME 62) around the singers' heads, which was centered according to the position of the mouth within the spherical net. The advantage of the chosen method is recording at all 48 positions simultaneously in one take, with no need to move the singers or microphones and repeat the tasks. The signal was recorded at a bit depth of 24 bits with a sample rate of 48 kHz to the multi-channel Tascam X-48 recorder.

The singers sang the scale from C3 to C6 and back using the Czech vowels [8] A, E, I, O, U with mezzo-forte dynamics. The cut-out midpoints of the prolonged vowels of 500 ms were analyzed. The microphone located directly in front of the singer's mouth was used as a reference.

The calibration calculation was performed by comparing the RMS level values of the measuring

microphones to the reference. Sound radiation was calculated as the RMS of the total levels and by calculating the levels of the spectral bands observed by means of FFT.

The user's software Directivity was prepared in MATLAB to process the results. In order to compare the radiation properties of the singers, a pair of regions was prepared according to the location in the voice range profile (according to the pitch and SPL) to match the regions with the same pitch and the smallest difference in the SPL.

A comparison of the radiation between the two groups of sounds was made for each direction and for each type of vocals separately using the paired t-test method. The comparison took place both throughout the frequency range and in the individual frequency bands.

III. RESULTS

The results confirm that the directional radiation of the opera singer and the pop singer are systematically different from each other, as is shown in Fig. 1 (top row). The pop singer radiates in a narrower way than the opera singer. In particular, the most significant systematic differences are from 90° to 180° azimuth, and in elevation $\pm 30^\circ$ in 0° azimuth. These differences are more or less evident depending on vowels.

Fig. 1 (bottom row) shows the area of the frequency band between 2.5–4 kHz (the singer's formant spectral band), where the differences in radiation between individual vowels were also confirmed for the opera singer. Those trends are similar to directional characteristics of the vowels in the total row of Fig. 1.

Different directional characteristics were also observed in the opera singer, depending on the vowel sung. In contrast to the E of the opera singer, the vowels A and U radiated lower relative levels in all directions, while the vowels O and I had a similar difference from E (Fig. 2, left side Clas. row). The pop singer's differences are shown in the Pop. row and they are similar to those of the classical singer regarding the comparison of E and A. The comparisons of other vowels show differences in radiation mainly in the back part.

The comparison of vowels was observed also for the frequency band 2.5–4 kHz (also Fig. 2, right side). The vowel E radiates at relatively higher levels in almost all directions without any noticeable differences between the pop and classical singers.

Since the observed radiation characteristics were related only to the RMS level of the reference microphone, the comparison of individual vocals confirms the effect of the overall voice intensity.

The radiation of the singers depending on the pitch of both singers is shown in Fig. 3. There are groups of tones which distribute the sound pressure around the head of the singer similarly.

IV. DISCUSSION

Regarding the aspect of originality of the topic of the study, it is difficult to compare the results with those of other studies.

Data in Fig. 1 and 2 agree with the theory of the Katz's study [6], which means the sound pressure field depends on the shape of the head and mouth opening of singer.

The left and right symmetry of the field was also confirmed [3]; however there are some exceptions obviously caused by lack of attention to fixing the position of the singer's head. The singer had been given instructions, but was not checked the whole time during the process of taking measurements.

As the measurements were always calibrated to the reference, they do not show absolute values of sound pressure level; they are in relative scales. It could affect the results for silent and transient tones. Therefore, a comparison of absolute SPL values may show different results.

The work confirmed there are differences between spreading sound emissions of pop and opera singer which may refer to the singing technique used. As pop singers are supported with an amplifier (microphone), they do not need to acoustically excite the hall to support their voice. This could be one of the main reasons why the field of the pop singer is narrower. Opera singers also form the vowels in different parts of the mouth than pop singers, which can mean a different shape of the mouth and also use of different cavities.

The larger emission of the pop singer at an elevation -30° may be caused by colorful speech elements contained in singing. This theory was formed on basis of Chu's study of different kinds of speech [4].

The main point of the discussion is that it was proved that the pitch significantly correlated to SPL, which means the differences between tones and their radiation can be caused by SPL changes in addition to the voice range profile of the singer.

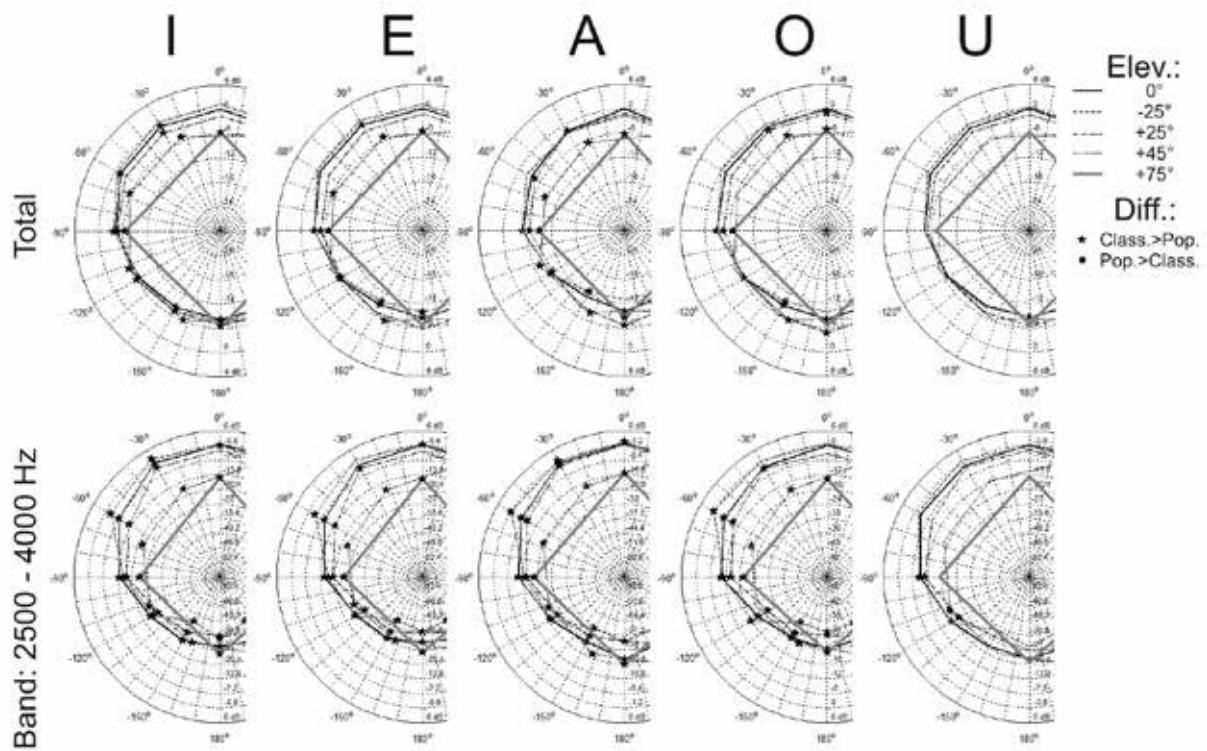


Fig. 1: Comparison of sound radiation between classical and pop singer for whole frequency range (total row) and for spectral band 2500–4000 Hz.

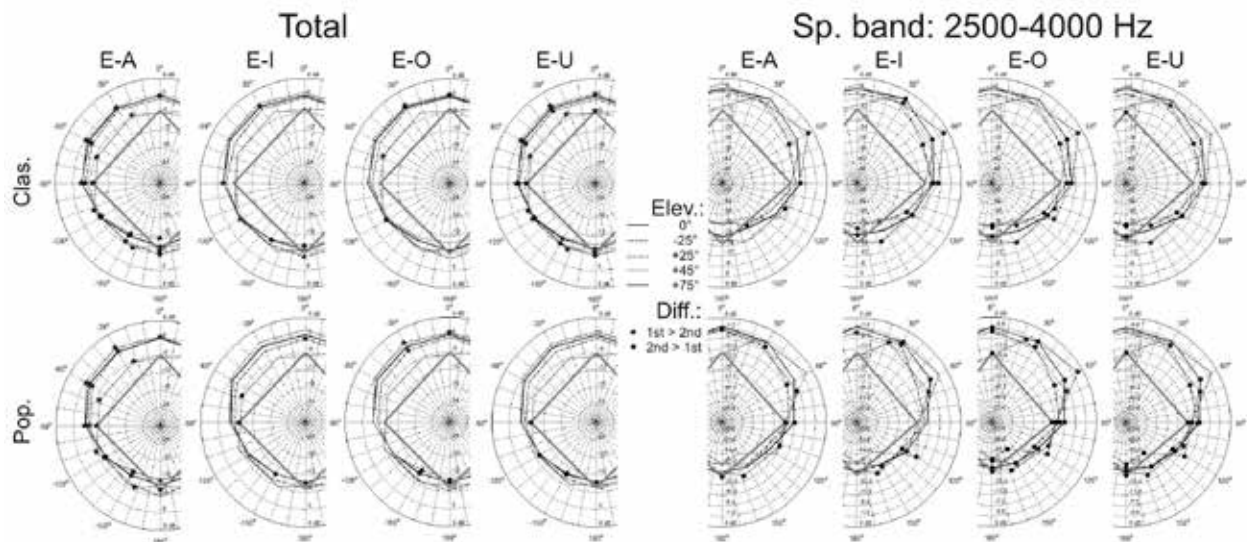


Fig. 2: Comparison of full frequency range power radiated (Total) and in the frequency band 2500–4000 Hz (right side) between vowels of classical and pop singer.

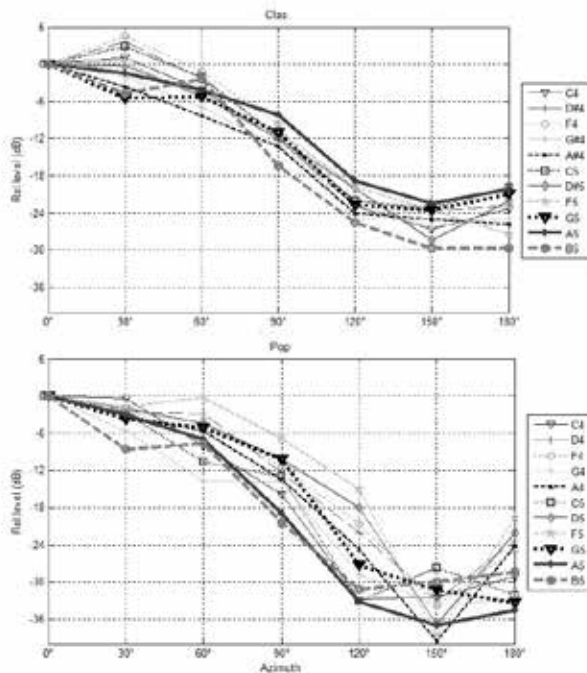


Fig. 3. Relation between full frequency range power radiated in horizontal plane and the pitch for classical (top) and pop (bottom) singer.

V. CONCLUSION

Preliminary work confirmed systematic differences in the sound radiation of two singers from different genres; the pop singer was significantly more directional, which probably was related mainly to the size of the mouth opening [6]. The study confirmed the effect of pitch on radiation and, when comparing

different vocals, suggests that the sound radiation also changes relative to the overall SPL.

REFERENCES

- [1] Dunn, H. K. and D. W. Farnsworth. Pressure field around the human head. *J. Acoust. Soc. Am.* 1939. 10, 180-98.
- [2] Flanagan, J. L. Analog measurements of sound radiation from the mouth. *J. Acoust. Soc. Am.* 1960, 32(12):1613-20.
- [3] Studebaker, G. A. Directivity of the Human Vocal Source in the Horizontal Plane. *Ear Hear.* 1985,6(6):315-19.
- [4] Chu, W. T. and A. C. C. Warnock. *Detailed directivity of sound fields around human talkers.* Institute for Research in Construction, National Research Council of Canada, Ottawa, ON, 2002, Canada, Technical Report IRC-RR-104.
- [5] Kob, M. and H. Jers. Directivity measurements of a singer. In: *Collected Papers from the Joint Meeting "Berlin 99"*. Inst. of Technical Acoustics, Technical University of Aachen., 1999. ISBN 3-9804568-5-4
- [6] Katz, B. and C. d'Alessandro. *Measurement of 3D Phoneme-Specific Radiation Patterns in Speech and Singing.* 2007. Online: https://rs2007.limsi.fr/PS_Page_14.html#LINKS
- [7] Cabrera, D. et al. Long-Term Horizontal Vocal Directivity of Opera Singers: Effects of Singing Projection and Acoustic Env. *Journal of Voice.* 2011, 25(6):e291-03.
- [8] Skarnitzl, R., Volin, J. [Reference values of vowels formants of young and adults speakers of Czech language standard]. *Akustické listy*, 2012, 18(1), s. 7-11.

DISSOCIATION OF SPOKEN AND SUNG VOCAL PRODUCTION

D. Van Lancker Sidtis^{1,2}, Y-J. Kim^{1,2}, J. J. Sidtis^{1,3}

¹ Nathan Kline Institute/Brain and Behavior Laboratory, Orangeburg NY, USA

² New York University/Department of Communicative Sciences and Disorders, NY NY, USA

³ New York University Langone Medical School/Psychiatry, NY NY, USA

Diana.Sidtis@nyu.edu, John.Sidtis@nyu.edu, yjk375@nyu.edu

Abstract. Singing and speech prosody share acoustic components (pitch, timing, and rhythm) which may be differentially impacted by brain damage. The effects of focal brain damage on pitch, timing, and rhythm in speech and singing were investigated in two persons diagnosed with dysprosodic speech; both were experienced singers and native speakers of English. Participant 1 suffered a right hemisphere infarct and Participant 2 sustained a right-sided subcortical lesion. Pitch and timing in lexical contrasts were acoustically analyzed, rhythm and pitch in spontaneous speech were quantified, and pitch and rhythm in familiar songs were measured and rated by listeners. Both participants produced lexical contrasts with non-normal pitch but normal timing. Rhythm in spontaneous speech was abnormal for P1 but not P2; rhythm in singing for both subjects was accurate. Both subjects failed in linguistic pitch but preserved linguistic timing; they differed in speech rhythm but not in sung rhythm; they differed in sung pitch ability, implying that talking and singing are modulated by disparate neurological systems.

Key words: dysprosody, singing and speech, pitch, rhythm, timing

I. INTRODUCTION

Pitch, timing, and rhythm are elements of speech prosody, or melody of speech, which have had benefit of much attention in recent years. In this study, pitch refers to measurable fundamental frequency in spoken word, phrase, or musical note; timing is described using duration measures at the level of the syllable or word; rhythm represents a regular, repeated pattern of sound in speech or song, or, a systematic arrangement of sounds with respect to stress and durations. Cerebral control of vocal fundamental frequency (F₀), or pitch and timing and rhythm in vocal production is complex and is not yet fully understood.

Brain damage affects prosodic production for linguistic, grammatical, and pragmatic contrasts in speech. For speech, some proposals have attributed all

prosodic competence to the right hemisphere (RH). However, left hemisphere damage is also associated with impaired melody of speech. Thus the role of the left and right cortical hemispheres and of subcortical nuclei in processing speech prosody remains unclear. Further, the relationship between talking and singing remains controversial. This study investigated competence for production of pitch, rhythm, and timing in spoken and sung exemplars by two experienced singers diagnosed with dysprosody of speech.

II. METHOD

Participant 1 (P1) was a 50-year-old right-handed Caucasian male with no previous neurologic or psychiatric history, who sustained a stroke to right hemisphere. Following the stroke, he was severely dysprosodic in conversation, elicitation tasks, reading, and repetition. Sentence intonation was monotone with a disturbance of rhythm. He had been an amateur singer since grade school, acting in musical theatre, a folk music group, and a barbershop quartet. After the stroke, he described himself as “tone deaf,” meaning he could no longer “carry a tune.”

Participant 2 (P2) was 36-year-old, right-handed African-American female with a basal ganglia infarct involving globus pallidus, caudate and medial putamen, more marked on the right side. She worked as a nurse on a psychiatric inpatient ward. She reported a significant lack of motivation (abulia), stated “My speech is flat.” A tape recording of a telephone conversation indicated normal to higher than normal fundamental frequency variability and range (135–420 Hz) in her pre-morbid speech. She sang in a church choir. On testing, she sang in a clear soprano voice.

For both participants in this study, speech and singing were examined in detail in terms of pitch, timing, and rhythm using two kinds of spoken samples, i.e., elicited word pairs and spontaneous discourse, and sung samples of singing familiar songs. For evaluation of pitch and timing in speech, vocal

productions of linguistic-prosodic pairs such as “greenhouse” and “green house” were elicited, and compared to exemplars produced by matched control subjects. Subjects were shown pictures or definitions (“a house that is green”) and asked to produce the appropriate term. Fundamental frequency and rhythm in spontaneous speech and singing were evaluated, with participants’ contributions being compared to healthy speakers and singers, and in the case of singing, with a piano. Listeners’ ratings of pitch and rhythm accuracy as well as “goodness” of the sample were utilized as a second measure of pitch and rhythm performance in the musical selections. Accuracies of tonal pitch and rhythm for sung samples were independently evaluated by 13 musically-trained raters recruited following IRB procedures.

III. RESULTS

Pitch relations in noun pairs differing in syllable accent (greenhouse, green house) deviated from matched HC values for both participants. Temporal relations in word-level contrasts were relatively maintained by both participants. For F0 in spontaneous speech, pitch ranges (maximum F0 – minimum F0) for both Participants 1 and 2 were lower than those of their healthy controls (Participant 1: range = 49.87 Hz vs. HC 1: 114.25 Hz; Participant 2: 28.82 vs. HC 2: 133.22) In spontaneous speech samples: These values reflect dysprosodic speech. The speech rhythm measures, Pairwise Variability Index (PVI), were 53.2 for Participant 1 (control = 64.1) and 57.2 (control = 58.5). for Participant 2. These values support the clinical observation that spontaneous speech contained distorted rhythm for P1 but not P2.

Pearson product-moment correlation coefficients were computed to assess the relationship of pitch and rhythm between study participants and HC in sung samples. For “Deck the Halls” sung by P1, there was non-significant correlations between P1’s pitches at singing and HC ($r = 0.198$; $p = 0.111$). P1’s pitches diverged from those of HC. There was a positive correlation between pitch values for singing by P2 and singing by HC ($r = 0.959$, $p = 0.0001$) and P2 and the piano ($r = 0.961$, $p = 0.0001$)

Rhythm was also evaluated acoustically for “Deck the Halls” in P1 and “Amazing Grace” in P2. Pearson Product-moment correlations were significant for rhythm measures for P1 ($r = 0.647$, $p = 0.0001$) and for P2 ($r = 0.788$, $p = .0001$). Rhythm values of P1 and P2 closely overlap with those of HC, suggesting that musical durational intervals were appropriate for both participants. When comparing Participant 1 to the healthy control (HC) singing “Deck the Halls”,

numerous pitch inaccuracies are evident. Of the 68 pitches in the song, Participant 1 sang 47 (69%) inaccurately. The greatest discrepancy can be seen in line 6, where only the first note is sung on the same pitch by both patient 1 and the HC. Portions of the song are produced on full monotone. However, the rhythmic patterned timing of the syllables was rated as mostly accurate.

For Participant 2, both pitch and rhythm elements on her familiar songs were judged correct. Productions of pitch levels of P2 and the HC were compared. There are 28 different pitches represented in the singers’ interpretation of “Amazing Grace”. Of those 28 pitches, 20 are sung the same by both singers (71.4%). Eight of them are accurate variations (28.6%). To verify our finding of relatively accurate pitch production in song for P2, a second analysis was performed on another song produced by this patient, “My country ‘tis of thee,” compared with a normal-control singer and notes from a piano. “My country ‘tis of thee” was also analyzed alongside a piano version. Healthy singers scored at around 1 on pitch and rhythm for Deck the Halls and Amazing Grace. P1 was scored as very poor on pitch (mean = 4.5) but reasonably proficient in rhythm (mean = 2.4). P2 received a greater number of good scores for both pitch (mean = 2.2) and rhythm (mean = 1.4).

III. DISCUSSION

The question underlying this research was whether characteristics of dysprosody of speech are reflected in singing, or whether they are dissociated. Two singers with dysprosodic speech following a stroke to right cortical and/or subcortical structures were studied. Spoken and sung productions analyzed in terms of pitch, timing, and rhythm. Performance on pitch, timing, and rhythm in speech as compared to singing (see Table 1) were in part dissociated. The results of acoustic analysis were consistent with listeners’ rating.

Speech	Pitch		Timing	
	P1	P2	P1	P2
Word level contrasts	✗	✗	✓	✓
Speech	Pitch		Rhythm	
	P1	P2	P1	P2
Spontaneous speech	✗	✗	✗	✓
Singing	Pitch		Rhythm	
	P1	P2	P1	P2
Familiar songs	✗	✓	✓	✓

Table 1. Overview of pitch, timing and rhythm in speech and singing for P1 and P2. (✘ = deficient and ✔ = normal performance).

For Participant 1, pitch control was impaired in spoken and sung vocalizations. Rhyme in speech was abnormal while rhythmic structure in familiar songs was normal. For Participant 2, pitch control in spontaneous speech was deficient, while timing and rhythm in speech, as well as musical pitch and rhythm in song were normal. This result implicates a role of right-sided cortical structures in control of pitch and rhythm in speech, while being dissociated in singing, where pitch was severely impaired while rhythmic relationships were intact (referencing P1). For speech, these results conform to the general model that ascribes timing features to the left hemisphere and pitch processing in speech to the right cortical and subcortical structures [1]. The surprising finding for rhythm in Participant 1, intact in singing but impaired in speech, lends support to a view of rhythm as a complex parameter that is multiply represented in the brain, providing structure to speech and singing respectively from differing systems [2].

V. CONCLUSION

These observations lend support to the notion that speech and singing are modulated by disparate cerebral systems. The dissociation in pitch and rhythmic competence for speech and singing, while timing was intact in both kinds of vocalization, implies two different control systems for production of speech and song.

[1] R. Zatorre & P. Belin, "Spectral and temporal processing in human auditory cortex," *Cereb Cortex*, 11, 946-953, 2001.

[2] S. Fujii, & C. Y. Wan, "The role of rhythm in speech and language rehabilitation: the SEP hypothesis." *Frontiers in Human Neuroscience*, 8, 1-18, 2014.

SHOULD I OPEN MY MOUTH MORE TO SING LOUDER?

A. Vurma

Estonian Academy of Music and Theatre, Tallinn, Estonia

Abstract: In literature on vocal pedagogy, we find suggestions to modify the vocal tract shape when singing louder. This should also influence the formants. Sopranos tend to tune the F1 to the spectral partial at high pitches. It is not clear whether singers tend to modify the formant frequencies for louder tones at pitches where the formant tuning may not be relevant, or what the benefit of doing so might be. Eleven male singers with a background in classical training sang a simple melody on the major scale steps I-II-III-II-I, altogether 20 times. The series included two dynamics contours, two tonalities in a comfortable voice range, and five basic vowels. Singers tended to raise the F1 slightly when they increased the loudness. The rise in F1 raised the absolute level of the singer's formant by a few decibels and the location of the centre of gravity of the voice spectrum by about 25% on average, making the sound brighter. Also, the LF1 (and hence the overall sound level) often tended to rise. It is possible that, besides reasons relating purely to energy, the importance of such changes may lie in semiotic signaling relating to the increased vocal effort.

Keywords: Singing, formants, dynamics, mouth opening, spectrum

I. INTRODUCTION

In the literature on vocal pedagogy, we may come across suggestions to open the mouth wider when singing louder [1-3]. Speech phonetics research has also revealed the tendency of covariation of the vocal tract parts' movements (and hence the shift of formants) with changes in voice intensity [4, 5]. It is well known that sopranos tend to increase their mouth opening at high pitches to tune F1 to the fundamental, and thus gain in sound level [6].

This study aims to find out whether male vocalists, when singing at moderate pitches where the formant tuning does not seem relevant, also tend to modify the F1 when singing louder, and what the benefit of doing so could be.

II. METHODS. PART I

Eleven male singers (5 tenors, three baritones, and three basses) with a background in western classical

training volunteered as the participants. Their average age was 37 y (between 24 and 63, standard deviation 13 y).

First, the vocalists were asked to read aloud three random sentences from a newspaper to determine the pitch of their speaking voice.

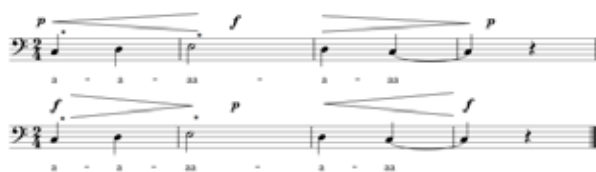


Fig. 1. The two dynamics contours used in the singing tasks. Each singer had to perform altogether 20 versions (2 dynamics contours \times 2 tonalities \times 5 basic vowels). The acoustical measurements were conducted only for the notes indicated by the asterisks.

As the next step, we asked each participant to sing 20 versions of a simple melody, consisting of the three first notes of the diatonic major scale (see Fig. 1). In the case of the first tonality, the melody started from the average speaking pitch of the person. The second tonality was a perfect fourth (P4) above the first. This made it possible (1) to use a comfortable pitch range for all participants regardless of their voice category, and (2) to keep the frequency distances between the formants and the spectral partials random. The first dynamics contour version of the task was *p-crescendo-f-diminuendo-p*, the second contour was its reverse (see Fig. 1). We conducted the recordings in a studio with low reverberation ($T_{30} = 0.2$ s). The omnidirectional microphone DPA SC4061-FM was positioned a few centimeters off-axis at the corner of the vocalist's mouth and connected to the PC via the external USB audio interface TASCAM US-122 MKII.

From the recordings, we captured the two first formant frequencies and their bandwidths (F1, F2, B1, B2) of the first note (the tonics), and the third note of the melody (the third, see Fig. 1). For this, we used manual inverse filtering (software *Sopran* 1.0.20).

III RESULTS. PART I

The average fundamental frequency (f_0) of the speaking voice of the participants was 129.8 Hz (about C3) with a standard deviation (SD) of 17.4 Hz. The lowest individual value was 103.8 Hz (about Ab2), and the highest 155.6 Hz (about Eb3).

The data for F1 and F2 as captured by the inverse filtering is presented in Fig. 2. The F1, on average, was higher in the case of all vowels when sung *forte* compared with *piano*. The F2, on average, was also slightly higher when singing *forte*.

Two overarching ANOVA-s for F1 and F2 showed that the dynamics, as well as the vowel, was a statistically significant factor affecting the values of these formants.

The results of the ANOVA for the F1: the main effect of vowel quality (/a/, /e/, /i/, /o/, /u/), $F(4, 400) = 383.2$, $p < .001$, the effect size $\eta^2_p = .79$; the main effect of dynamics (*piano*, *forte*), $F(1, 400) = 53.6$, $p < .001$, the effect size $\eta^2_p = .12$, indicating that the average value of F1 over all vowels was significantly greater for notes sung *forte* ($M = 455.7$ Hz, $SD = 104.6$ Hz) than for notes sung *piano* ($M = 421.8$ Hz, $SD = 100.9$ Hz). Although this difference of 33.9 Hz (about 8%) is quite small, it is still clearly greater than the JND of the F1, which is between 10 and 30 Hz [10].

The second multi-way ANOVA for the F2 showed the strongest main effect of the vowel quality (/a/, /e/, /i/, /o/, /u/), $F(4, 400) = 1186.3$, $p < .001$, the effect size $\eta^2_p = .92$. In the case of F2, too, the test revealed the main effect of dynamics (*piano*, *forte*), but this was only marginal: $F(1, 400) = 2.93$, $p = .088$, the effect size $\eta^2_p = .007$.

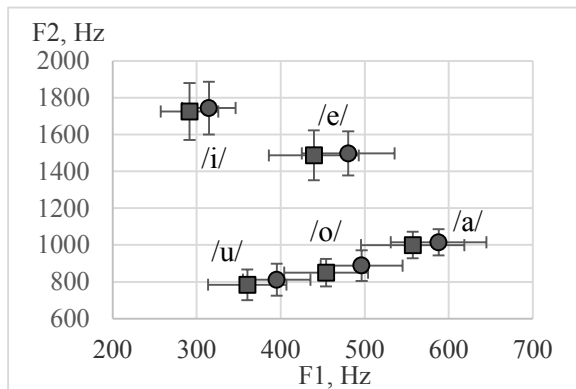


Fig. 2. The average values of F1 and F2 of five basic vowels sung *piano* (squares) and *forte* (circles) with standard deviation whiskers.

IV METHODS. PART II

We will now go on to estimate the changes in the levels of spectral partials in some strategic frequency regions that were caused by the dynamics induced shift in F1 and F2. The spectral levels on which we will focus, are: (1) the level of the strongest partial of the whole spectrum (LF1), and (2) the level of the strongest partial in the region of the singer's formant (Lsf, typically located between about 2.3 – 2.9 kHz).

The strongest partial of the spectrum is important as it is this that mainly determines the overall sound pressure level [7]. Except for very soft or very high pitches, this is the partial which is closest to the F1 [7]. A strong singer's formant gives brilliance and good carrying power to the voice [6]. It can affect the perceived loudness of the voice, as (1) in this frequency region the human hearing system is typically most sensitive [8], and (2) loudness is summed loudness in all critical bands of hearing [7]. (F1 and Fsf are typically located in different critical bands.)

Besides, the centre of gravity (CG) of the voice spectrum can be an indicative parameter for our work, as sounds with a higher centre of gravity and relatively stronger energy at higher frequency bands may be perceived as “more focused,” prominent [9], and louder [10].

According to the source-filter theory of voice production, the levels of the voice spectrum partials depend (1) on the level of the partials of the voice source spectrum, and (2) on the transfer function (TF) of the vocal tract, which reshapes the spectrum of the voice source. The TF of the vocal tract is the sum of the transfer functions of the single formants. Hence, if the distance between two formants decreases, the TF of the vocal tract both at the formants and between them increases. The increase is the biggest midway between the formants. [11]

Predicting the changes in the levels of spectral partials caused by the F1 and F2 shift is complicated by the fact that not only the vocal tract-s TF but also the distances of spectral partials from the formants (which depends on f_0) play a role. Therefore, the LF1 may be both strengthened or weakened—and hence the SPL may increase or decrease—as the F1 rises. Similarly, the Lsf can change in either direction, although its strengthening is more likely. Furthermore, the centre of gravity of the voice spectrum may be expected to increase with a rise in F1. However, it is more complicated to hypothesise about the behavior of the relative level of the singer's formant (Lsf – LF1) as this depends on both the LF1 and the Lsf.

Next, we will go on to find out how great the changes in spectral levels typically are in real singing. It is not possible directly to measure the magnitude of the changes in the spectral component levels that are caused only by the shifting F1 and F2, as other factors also play a role. For example, the most typical way to raise voice dynamics, especially at high intensity conditions, is to increase the subglottal pressure [12]. Quite typically, the glottal adduction also increases when singing louder [12]. Both these factors affect the balance in levels of the spectral partials between different frequency regions [12].

In order to obtain a rough estimate of the magnitude of the influence from purely articulatory factors (i.e.,

the F1 and F2), we used analysis by synthesis (with the help of software *Madde* 3.0.0.2).

The input parameters for the synthesis were the F1 and F2, measured from the performances of our participants. The idea was to keep all other parameters, except the F1, F2, and f_0 , always the same. (We were able to do this as the TF of the vocal tract is the sum of TF-s of single formants, and we are interested only in the magnitude of level changes at the singer's formant, not their absolute values.) In this case, the difference in the levels of the spectral partials at the strategic spectral regions of the voice spectrums of comparable sounds that were of interest to us should be caused only by the differences in their F1 and F2, and not influenced by other factors such as the subglottal pressure and glottal adduction.

We used as the input for the synthesis only about half of the whole database and included only those performance-pairs from all the performance subgroups (categorized by the vowel, tonality region and scale step position) where the difference in F1 between singing *piano* and *forte* was greatest and always bigger than the JND of the F1. The criteria for the JND of the F1 was 10 cents.

The input data for the second part of our work thus includes the information for the F1 and F2 of 200 sounds (100 *piano/forte* pairs) out of the total of 440.

For the fixed parameters, we used the default settings offered by *Madde*: the voice source spectrum tilt was -6 dB/octave, vibrato frequency 6.5 Hz, vibrato amplitude 0.3 semitones, flutter amplitude 1%. The fixed values of the higher formants were: F3 = 2.5 kHz, F4 = 2.7 kHz, F5 = 2.9 kHz. We also used the default Q values for all formants (Q = 10). All the synthesized sounds were low-pass filtered with the cutting frequency of 10000 Hz to avoid aliasing.

In the following step, we measured, from the synthesized sounds' spectrums, the LF1, Lsf, Lsf-LF1 (FFT Blackman window, bandwidth 50 Hz), and CG, all averaged over about 0.5 s of duration in the stationary part of the sound. Here, too, we used the software *Sopran* 1.0.22. Then we calculated the magnitudes of the corresponding differences between the synthesized sounds that corresponded to *forte* and *piano* executions (according to their F1 and F2).

V. RESULTS. PART II

The analysis by synthesis showed that in the case of performance-pairs where our singers shifted the F1 up when singing *forte* rather than *piano*, the average values of LF1, Lsf, Lsf-LF1, and CG also increased when singing *forte*, solely because of the change in F1 and F2 (see Tab. 1). The influence of F1 and F2 on the spectral levels was strongest for the Lsf, which was 3.5 dB greater on average in the case of those synthesized

sounds whose F1 and F2 corresponded to the *forte* dynamics. In some cases, the gain was even as big as 10 dB, while the corresponding level decreased only in one case (a decrease of only -1 dB). The tendency was similar for all vowels. Furthermore, in every case but one the centre of gravity was always higher in the case of synthesized sounds which used the F1 and F2 of *forte* dynamics. The average difference in the centre of gravity was 97 Hz (23.9 %). The direction of the change was the same for all vowels, but the difference was somewhat greater in the case of vowels with a low F1 – the /e/ and /i/. The LF1 and (Lsf – LF1) were also stronger on average in the case of synthesized sounds whose F1 and F2 corresponded to the *forte* dynamics (a difference of 0.8 and 2.7 dB respectively, with maximum gains of 12 dB and 14 dB). However, in quite a number of cases the corresponding levels were also lower at *forte*. The maximum decrease reached the values of -10 dB and -9 dB respectively.

Tab. 1. Average differences in the spectral levels and centre of gravity (CG) of synthesized sounds, which used the two lower formant values of *forte* and *piano* executions. On average, the values of all parameters in the table were greater in *forte*.

	LF1	Lsf	Lsf-LF1	CG	
	<i>f-p</i>	<i>f-p</i>	<i>f-p</i>	<i>f-p</i>	<i>f-p</i>
	dB	dB	dB	Hz	%
Mean	0.8	3.5	2.7	97	24
SD	5.9	2.2	6.4	84	19
Min	-10	-1	-9	-21	-6
Max	12	10	14	614	110

For all the parameters investigated, with the exception of LF1, the influence of the dynamics induced F1 and F2 shift was statistically significant according to the results of two-way ANOVA-s.

VI. DISCUSSION

Our results enable us to speculate about the possible reasons why singers tend to raise F1 for louder singing. In our experiment, the change in dynamics induced the changes in F1 and F2, which raised the LF1 on average when singing *forte*. It is the LF1, as the level of the strongest partial of the spectrum, that mainly determines the SPL of the whole sound. Therefore, on average we should also expect an increase in the SPL. However, the increase in LF1 was only 0.8 dB, which is smaller than the typical JND—1 dB [15]. In many cases, the LF1, on the contrary, decreased when singing *forte*. Therefore, opening the mouth wider need not necessarily increase the SPL.

Nevertheless, the benefit from raising the F1 can come from the additional gain in the subjectively perceived loudness, as the perceived loudness is the summed loudness of different critical bands. Raising the F1 when singing *forte* almost always raised the Lsf, the centre of gravity, and in the majority of cases, the Lsf-LF1. As F1 and Fsf are typically in different critical bands, the result should be an increase in perceived loudness.

We may speculate that one of the reasons why singers tend to open their mouth wider for singing louder might be related to the semiotic signaling about the changes in the vocal effort of the singer. Such changes can have meaning in the context of language prosody (e.g., for marking the lexical stress), or for expressing the musical nuances. Such signaling can encompass different simultaneous cues, which support and substitute each other if one channel fails. In phonetics, such joint acoustical phenomena can be, e.g., an increase in the overall level of the voice, the enrichment of the high part of the spectrum, a rise in the F1, or an increase in the voice pitch [13]. In music, pitch, as a rule, is dictated by the composer, and cannot be used freely in this context. Overall intensity alone (e.g., compared to the spectral emphasis) need not necessarily be the strongest and most effective cue for signaling prominence in speech prosody [14].

For some repertoire, e.g., in chamber music, or in small rooms, besides the ability to produce a loud voice, the skill of singing extremely quietly is also important. We may speculate that a semi-occluded vocal tract (which lowers the F1) allows extremely low sound levels to be produced more easily because of the favorable acoustical feedback from the vocal tract to the vocal folds' vibration [15]. According to Titze's report, a wide-narrow vocal tract configuration, where it is wide at the epilarynx tube and narrow at the mouth opening, minimizes glottal flow, and retains most of the sound inside the airways. On the other hand, the narrow-wide configuration with a bigger mouth opening (and raised F1) yields the greatest oral radiated pressure, good economy (great MDR at small maximum glottal area declination rate), and strong vocal ring [15].

VII. CONCLUSION

Singers tend to raise the F1 when singing louder, even if it is not related to formant tuning at high pitches. This adds mainly to the raising of perceived loudness as spectral balance increases towards the higher spectral partials. The increasing influence on the SPL is possible, but does not necessarily occur. The benefit of raising the F1 could also be related to signaling increased vocal effort. A small mouth orifice and lowered F1 can enhance the ability to sing piano

because of the changed feedback from the vocal tract to the vocal folds.

This study was supported by the Estonian Ministry of Education and Research (IUT12-1) and by the EU through the European Regional Development Fund (Centre of Excellence in Estonian Studies).

REFERENCES

- [1] G. Mancini, *Practical Reflections on Figured Singing* by Giambattista Mancini, Champaign, IL, Pro Musica Press, 1967.
- [2] R. Appleman, *The Science of Vocal Pedagogy: Theory and Application*, Bloomington, IN, University Press, 1986, p. 221.
- [3] O. Brown, *Discover Your Voice: How to Develop Healthy Voice Habits*, San Diego, Singular, 1996.
- [4] Schulman, R. "Articulatory dynamics of loud and normal speech", *J Acoust Soc Am*, 85(1), 295-312, 1989.
- [5] Traunmüller, H. and Eriksson, A. "Acoustic effects of variation in vocal effort by men, women, and children", *J Acoust Soc Am*, 107(6), 3438-3451, 2000.
- [6] J. Sundberg, *The Science of the Singing*, Dekalb, IL: Northern Illinois University Press, 1987.
- [7] Sundberg, J. "Perceptual aspects of singing", *Journal of voice*, 8(2) 106-122, 1994.
- [8] Hunter, E. J., and Titze I. R., "Overlap of hearing and voicing ranges in singing", *Journal of Singing*, 61(4), 387-392, 2005.
- [9] Heldner, M. "On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish", *Journal of Phonetics*, 31 39-62, 2003.
- [10] Duvvuru, S., and Erickson, M. "The effect of change in spectral slope and formant frequencies on the perception of loudness", *J Voice*, 27(6), 691-697, 2013.
- [11] Fant, G, *Acoustic Theory of Speech Production*, Mouton, The Hague, 1960.
- [12] Zhang, Z. "Mechanics of human voice production and control", *J Acoust Soc Am*, 140 (4), 2016.
- [13] Lienard, J. S., and Di Benedetto, M. G., "Effect of vocal effort on spectral properties of vowels", *J Acoust Soc Am*, 106(1), 1999.
- [14] Sluijter, A. M. C., Heuven, V. J., and Pacilly, J. J. A. "Spectral balance as a cue in the perception of linguistic stress", *J Acoust Soc Am*, 101(1), 503-513, 1997.
- [15] Titze, I. "Voice training and therapy with a semi-occluded vocal tract: Rationale and scientific underpinnings", *J Speech Lang Hear Res*, 49, 448-459, 2006.

**SESSION III - SPECIAL SESSION
EDUCATION AND REHABILITATION OF
THE ARTISTIC VOICE**

EDUCATION AND REHABILITATION OF THE ARTISTIC VOICE

SPECIAL SESSION

Coordinator: Franco Fussi

Ravenna, Italy
ffussi@libero.it

Abstract: Special Session devoted to the artistic voice. Contributions are given by: Franco Fussi, Erika Biavati, Eleonora Bruni and Elisabetta Rosa.

I. SOVTE: BETWEEN MYTH AND REALITY

Franco Fussi

Semi-occluded vocal tract exercises (SOVTEs) are widely used in the fields of voice therapy and didactics, aiming at improving vocal economy and efficiency. The rationale and theoretical underpinnings for SOVTEs have been described by Titze. SOVTEs promote an increase in vocal tract impedance, resulting in changes in the inertive reactance, with favourable effects on voice production because of a reduction of phonation threshold pressure and an increase of skewing of the glottal flow waveform. The increasing vocal tract impedance can affect the glottal function through acoustic-aerodynamic interactions and mechano-acoustic interactions.

Many different SOVTEs exist and have been described so far.

The common feature of these exercises is the reduction of the cross-sectional area of the vocal tract at or near the lips. Some of the most known SOVTEs are represented by lip and tongue trills, hummings, hand-over-mouth, resonance tubes, flow resistant straw, etc. In the use of SOVT disposals myths and legends have been created to which we will try to give an answer by providing a rational use.

II. THE SOVTE PROTOCOLS: INDICATIONS AND CRITICISM IN VARIOUS SINGING STYLES

Erika Biavati, Eleonora Bruni

Here we analyse the use of SOVT exercises in singing lessons, with the support of "tools" (bubble mask, straws, masks), but also with vocal exercises that exploit the benefits using a semi-occluded vocal tract. In particular different exercises are shown and referred to optimize performance results in various singing styles, in acting and in dubbing. Demonstrations will be offered on the application procedures and the results of some research on the effects of SOVT disposals will be exposed.

III. G.E.M.M.A. TRAINING: A LINK BETWEEN PSYCHOLOGY AND LOGOPAEDICS IN THE TREATMENT OF SINGING VOICE DISORDERS

Elisabetta Rosa

A new light to the logopaedic treatment of voice disorders in singing professionals will be cast, born from the useful collaboration with a psychotherapist/singer.

The approach to singers is not always easy for a voice therapist; to have some new tools to help compliance and efficacy can be of value. The cooperation with a Psychotherapist is paramount: G.E.M.M.A. training is primarily based on a psychometric tool (the Dispositional Flow Scale), which has been adapted and validated in Italian specifically for performance artists. So far, this training has shown interesting results in voice therapy sessions; moreover, has proven to be well accepted by singers probably because of its indirect approach, that doesn't discredit their current and actual singing technique.

SESSION IV
VOICE AND EMOTIONS

THE CORRELATION BETWEEN POETIC RHYTHM AND HEART RATE VARIABILITY IN SUBJECTS READING AND PERCEIVING RUSSIAN POETRY

K.V. Evgrafova¹, P.A. Skrelin², V.V. Evdokimova³, T.V. Chukaeva⁴

¹ Phonetics Department, Saint Petersburg University, Saint Petersburg, Russia

² Phonetics Department, Saint Petersburg University, Saint Petersburg, Russia

³ Phonetics Department, Saint Petersburg University, Saint Petersburg, Russia

⁴ Phonetics Department, Saint Petersburg University, Saint Petersburg, Russia
evgrafova@phonetics.pu.ru, skrelin@phonetics.pu.ru postmaster@phonetics.pu.ru

Abstract: There are a lot of psychological and physiological music rhythm experiments but very few related to the poetic rhythm. Beat and meter as units of rhythm are the fundamental elements of cognitive mechanisms. Poetic meter may have roots in the mechanisms controlling heart rate, breathing and locomotion which are dominated by brain stem. We designed an experiment to find out if reading verses having different metric structure can affect heart rate variability measures. The subjects were reading pieces of prose and verses in Russian while the HRV measures were being obtained.

The results demonstrated that there is relationship between a metric structure of a produced/perceived text and heart rate variability measures. They also pointed out that some acoustic features (syllable duration and accent) correlated with the maximum peaks of the heart rate. The faster rhythm enhanced the heart rate amplitudes. The main finding was that the heart rate is also strongly affected by perceiving a piece of poetry.

Keywords: Speech rhythm, heart rate variability, poetry

I. INTRODUCTION

Rhythm is a fundamental part of human life. Humans demonstrate ability to produce and perceive rhythmic behaviour in various forms. It can be applied to anatomical functions, e.g. respiratory rhythm, locomotor activity rhythm, heart (sinus, cardiac) rhythm; language (the temporal organization of speech, meter and beat in poetry) and music (its periodic composition).

Beat and meter are the fundamental elements of cognitive mechanisms. It is shown that beat perception is innate. Newborn infants expect onsets of rhythmic cycles, even unmarked by stress or the spectral features. Infants also engage in the rhythmic movement to rhythmically regular sounds and the faster movement tempo is associated with the faster auditory tempo [1].

A significant amount of research has been employed to study the neural basis of rhythm.

However, there have been a lot of psychological and physiological *music rhythm* experiments (related to music performance and perception) but very few related to the *poetic rhythm*. [3-6] In particular, there is a lack of evidence of differences in physiological responses to beat and meter as units of poetic rhythm. In phonology the term rhythm refers to the **perceived regularity** of prominent units in speech. These **regularities** may be stated in terms of patterns of stressed v. unstressed syllables, syllable length (long v. short) or pitch (high v. low), or some combination of these variables.' [2]

Normally, this research on speech rhythm (both in prose and poetry) has been focussed on timing, i.e. duration, stress, pitch and pausing. Thus existing evidence acknowledges the role of acoustic cues and explores the relative significance of these cues differs between languages.

We hypothesised that heart rate variability (HRV), one of the physiological actions, which is controlled by the autonomic nervous system, is tightly connected with speech rhythm. The goal of this research was to analyse the relationship between poetic rhythm and heart rate variability.

We designed an experiment to find out if reading/perceiving *verses* having **different metric structure** can affect **heart rate variability** measures. It was considered that rhythm perception might be influenced by native language.

Russian poetry is mainly *syllabo-tonic*, i.e. a type of accentual versification which is based on the **regular alternation of strong** (stressed) syllables and **weak** (unstressed) syllables. It employs five basic meters. The meters of syllabo-tonic versification are **binary** and **ternary**.

The binary meters are **trochaic meter**, in which an unstressed syllable occurs between two stressed ones, and **iambic meter**. The ternary meters have two unstressed syllables between the stressed ones, e.g. **dactylic meter**, the **amphibrachic** and the **anapaestic**.

II. METHODS

We recorded 4 healthy *subjects* who had reported no heart problems. During the experiment the subjects were performing several reading tasks.

For recording we used a condenser *microphone AKG HSC 271* attached to a *medical phonedoscope* (Fig. 1). The phonedoscope was placed on the subject's chest and fixed there. The subjects were asked to read pieces of prose and verses having different patterns of stressed v. unstressed syllables while the *HRV* measures were being obtained.

The recordings were made in a soundproof booth at the recording studio at the Department of Phonetics (Saint Petersburg State University).

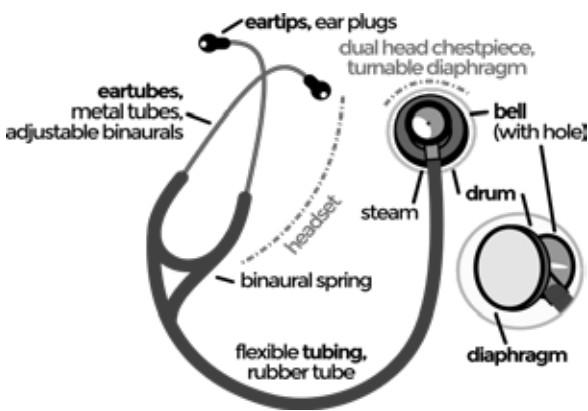


Fig. 1 A type of medical phonedoscope used in the experiment.

Thus each recording consisted of the following sessions.

Session 1

The subject was keeping *silence* while *HRV* and *breathing* were fixed.

The subject was reading a piece of *prose* while a speech signal, *HRV* and *breathing* were fixed.

Session 2

The subject was keeping *silence* while *HRV* and *breathing* were fixed.

The subject was reading a piece of *poetry* while a speech signal, *HRV* and *breathing* were fixed.

Session 3

The subject was keeping *silence* while *HRV* and *breathing* were fixed.

The subject was listening to a piece of poetry in headphones while *HRV* and *breathing* were fixed.

The *silence* recording in each session was made to obtain the reference HRV patterns for each speaker before they started to perform experiment tasks.

III. RESULTS

As the result of the experiment, we obtained

- 1) the synchronized plots of breathing and speech signal recordings (Fig. 2)

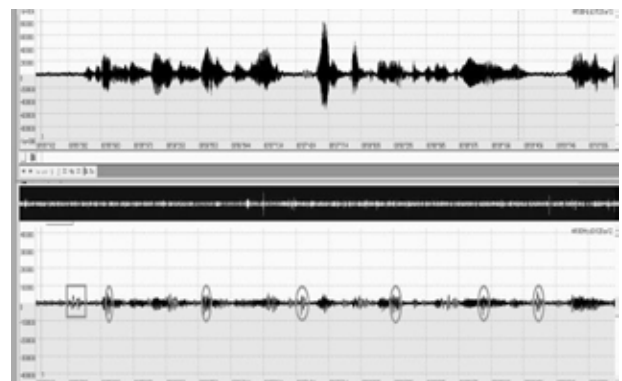


Fig. 2 Breathing and speech signal recordings synchronized.

- 2) the synchronized plots of HRV in *silence* and HRV in reading a piece of *prose* (Fig. 3-4);
- 3) the synchronized plots of HRV in *silence* and HRV in *reading* a piece of *poetry* (Fig. 5-6);
- 4) the synchronized plots of HRV in *silence* and HRV in *listening* to a piece of *poetry* (Fig. 7-8) for each subject respectively.

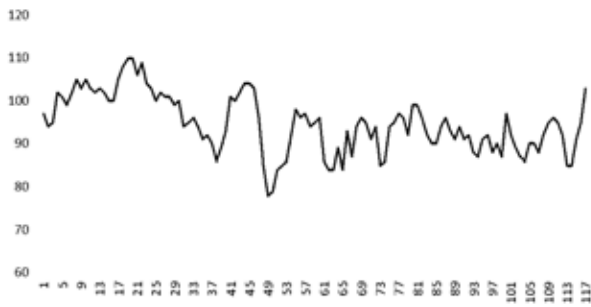


Fig. 3 HRV in *silence* (session 1, speaker 1),

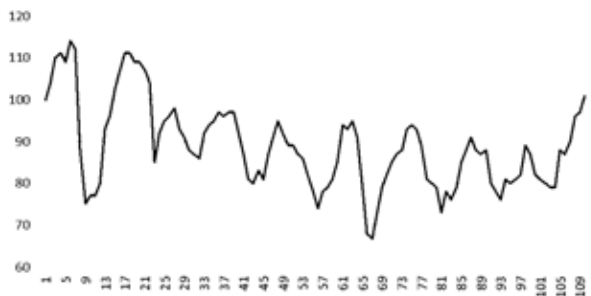


Fig. 4 HRV in *reading* a piece of *prose* (session 1, speaker 1).



Fig. 5 HRV in *silence* (session 2, speaker 1).

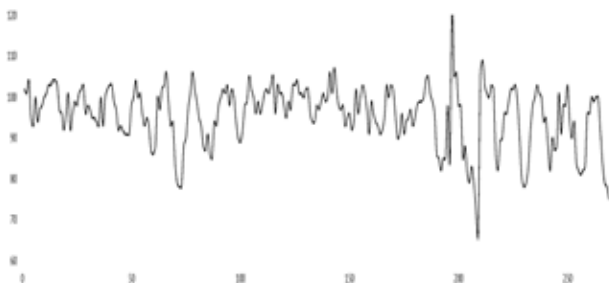


Fig. 6 HRV in *reading* a piece of *poetry* (session 2, speaker 1)

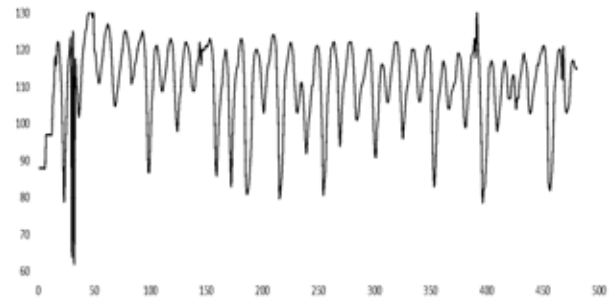


Fig. 7 HRV in *silence* (session 3, speaker 1).

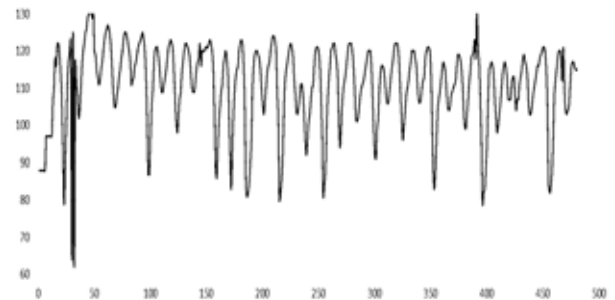


Fig. 8 HRV while *listening* to a piece of *poetry* (session 3, speaker 1).

IV. DISCUSSION

Our experiment explored how speech rhythm (in prose or in poetry) influenced the HRV. The results demonstrated that there is relationship between a metric structure of a produced/perceived text and heart rate variability measures.

The comparisons of the plots obtained showed the difference in HRV patterns during different activities.

They also pointed out that some acoustic features (syllable duration and accent) correlated with the maximum peaks of the heart rate. The faster rhythm enhanced the heart rate amplitudes. The main finding was that the heart was also strongly affected by perceiving a text having a metric structure (a piece of poetry).

V. CONCLUSION

The extended experiments with more subjects can help obtaining more evidence on the correlation between HRV patterns and certain metric types. It will be also necessary to describe the correlation between breathing/heart beat/stress patterns.

The extended research result will allow integrating the major factors of the physiological responses (respiration, skin conductance, and heart rate) to rhythmic features of speech produced/perceived into the model of embodied speech.

REFERENCES

- [1] G. Cervellin and G. Lippi, "From music-beat to heart-beat: a journey in the complex interactions between music, brain and heart," *European Journal of Internal Medicine*, vol. 22, no. 4, pp. 371–374, 2011.
- [2] D. Crystal, (Ed.) *Dictionary of Linguistics and Phonetics* (2nd ed.). Oxford: Blackwell 1985.
- S.-H. Lin, Y.-C. Huang, C.-Y. Chien, L.-C. Chou, S.-C. Huang, and M.-Y. Jan, "A study of the relationship between two musical rhythm characteristics and heart rate variability (HRV)," in *Proceedings of the IEEE International Conference on BioMedical Engineering and Informatics*, pp. 344–347, 2008.
- [4] H.-M. Wang, S.-H. Lin, Y.-C. Huang et al., "A computational model of the relationship between musical rhythm and heart rhythm," in *Proceeding of the IEEE International Symposium on Circuits and Systems (ISCAS '09)*, pp. 3102–3105, Taipei, Taiwan, May 2009
- [5] A. Zeman, F. Milton, A. Smith, R. Rylance, «By heart - An fMRI study of brain activation by poetry and prose », *Journal of Consciousness Studies*, 20, 2013, p. 132–158.
- [6] University of Exeter, «Poetry is like music to the mind, functional magnetic resonance imaging reveals», *ScienceDaily*, 2013. 9/10/2013, www.sciencedaily.com/releases/2013/10/131009125959.htm

EMOTIONALLY EXPRESSED VOICES ARE RETAINED IN MEMORY FOLLOWING A SINGLE EXPOSURE

Y-J. Kim^{1,2}, J. J. Sidtis^{2,3}, D. Van Lancker Sidtis^{1,2}

¹ Department of Communicative Sciences and Disorders, New York University, New York, NY, USA.

² The Nathan Kline Institute for Psychiatric Research at Rockland Psychiatric Center, Geriatrics Division, New York, NY, USA.

³ Department of Psychiatry, New York University Langone School of Medicine, New York, NY, USA.

Email: Y-J. Kim yjk375@nyu.edu; J. J. Sidtis john.sidtis@nyu.edu; D. Van Lancker Sidtis diana.sidtis@nyu.edu

Abstract: Previous studies in voice recognition familiarize listeners with new voices after a series of training sessions. Such approaches lack ecological validity, and thus may not fully capture our everyday experiences wherein voices are instantaneously acquired when conveyed in a meaningful context. Accumulating evidence suggests that emotionally engaging experiences are strongly encoded and strengthened in memory relative to neutral experiences. However, little is understood about how emotional context influences the acquisition of voices. In the present study, we investigate whether humans can acquire a newly familiar voice from a single, one-minute exposure to spontaneous speech when supported by a personally engaging context. Participants watched 8 one-minute videotaped narratives, half emotionally nuanced and half neutral, produced by performers. Immediately after the exposure, they heard 80 audio excerpts (half emotional, half neutral), and indicated whether or not they had heard the voice before. Participants returned one week later and completed the same recognition test. Emotional voices were more accurately recognized as having heard before than neutral voices in the delay condition, suggesting that nuanced, emotional context facilitates inducting new voices into a repertory of personally familiar voices in long-term memory. The findings further support differential cerebral processes for familiar and unfamiliar voices.

Keywords: familiar voice, emotion, voice recognition, memory, consolidation

I. INTRODUCTION

Studies in biology suggest the evolutionary significance of voice recognition ability across species [1, 2]. The evidence for instantaneous voice acquisition in species suggests that this kind of voice learning is likely a phylogenetically widespread capacity, and thus humans may share the capacity to rapidly acquire new

voices [3]. Such approaches suggest that the voice acquisition process may not be satisfactorily accounted for by frequent and long exposures. Unlike previous studies that have utilized a series of training sessions to familiarize listeners with new voices, the current study used a single, 1-minute exposure to investigate how new voices are naturalistically acquired into a familiar voice repertory in memory.

Previous scholarship in psychology has noted that memory is enhanced by emotion: people tend to remember arousing and emotionally engaging experiences better than neutral ones [4]. The memory of emotional events is subject to consolidation and persists over time [5, 6]. It is, however, unclear whether such memory advantage is applied to voices.

The current study was designed to examine the effects of emotionally expressive context on voice acquisition following a single, 1-minute exposure. We compared recognition of emotionally expressive and neutral voices immediately and one week after exposure. Three specific aims were explored: 1) whether exposure to a voice for one minute suffice for acquiring a voice as recognizable as having been heard before, 2) whether voices with emotional context are acquired and stored preferentially in memory relative to voices heard in neutral context?, and 3) whether memory for emotional voices are enhanced one week after exposure.

II. METHODS

The voice recognition task was designed to simulate natural experiences of voice acquisition by exposing participants to target voices through videotaped presentations. The task consisted of 1 exposure phase and 2 test phases (immediately after and one week later).

Participants

Thirty-five adults (26F) were recruited through advertisements to serve as listeners. All participants were native speakers of American English and reported no hearing difficulty. Age ranged from 19 to 60 years

($M=26$; $SD= 8.2$). An additional 7 participants were excluded: two did not return for the second session, four did not fit our inclusion criteria, and one encountered technical issues. The final sample included 28 participants.

Stimuli

The stimuli were spontaneous speech produced by twenty-eight female performers from a school for improvisational comedy. All performers were native English speakers, aged 20-40 years. Performers were asked to choose 2-4 topics to talk about (e.g., birthday parties, travel, family) in front of the video camera for 2-3 minutes each. They were free to select any topics of their own and use fictional material. Half of the performers were assigned to emotional condition and half were assigned to neutral condition. In the emotional condition, performers were asked to talk with heightened emotional expressiveness in an animated, nuanced way; in the neutral condition, they were asked to talk in a neutral, calm, nonexpressive way. Their speech was both audio- and video-recorded.

The recorded videotaped narratives were edited to be of approximately 1 min each. Five female adults then rated the edited videos in terms of their emotionality on a 7-point scale. This procedure was conducted to select narratives that represent extremes of emotional and neutral.

Procedure

Exposure phase During the exposure phase, each participant watched 8 one-minute videotaped narratives on the computer screen. Each videotaped narrative features one target voice. Half of the narratives were spoken with emotional expression (4 emotional target voices) and the other half were spoken with neutral expression (4 neutral target voices). The narratives were presented with a pseudo-random order with 4-s inter-stimulus intervals.

Test phase 1 Immediately following the exposure phase, each participant was presented with 80 audio excerpts one at a time. Half of the excerpts were emotional and half were neutral. At the same time, half of the test excerpts were exposed voices (voices of the same performer with different verbal content) and the other half were new voices. The duration of the videos ranged from 3 to 4 seconds. For each excerpt, participants were asked to click 'yes' button if the voice had been heard during the exposure phase and 'no' button if it was a new voice. Participants were also asked to rate their confidence in their decisions on a 5-point scale (1=not at all; 5=very).

Test phase 2 After a one-week delay, participants returned and repeated the same recognition test with no additional exposure to videos.

III. RESULTS

In this analysis, we examined the recognition accuracy (% correct) and confidence ratings as dependent measures. Recognition accuracy was defined as the percentage of correct responses (correct *yes* or correct *no*), which was calculated for each participant by dividing the number of correct responses by the total number of responses under each condition. Any participants scoring more than two standard deviations below the overall mean recognition accuracy of all participants on three or more conditions were excluded from the analyses. Two participants were removed on the basis of this criterion, and the data from the remaining 26 participants were used for all further analysis.

Recognition accuracy

We conducted a repeated measures ANOVA with mode (emotional vs. neutral), exposure (exposed vs. control), and time (immediate vs. 1-week recognition) as within-subjects factors. This ANOVA revealed a significant main effect of exposure, [$F(1, 25) = 29.24$, $p < .001$], with greater recognition accuracy for exposed voices relative to control voices. That is, across all conditions, voices that were exposed during the video session were more likely to be correctly recognized as having been seen before than voices to which participants had not been exposed before. There was also a significant effect of mode, [$F(1, 25) = 6.12$, $p = .025$], with greater recognition accuracy for emotional voices than neutral voices. This suggests the possibility that, across all conditions, participants were more likely to retain memory for emotional voices than memory for neutral voices. No significant main effect of time emerged, reflecting the fact that recognition accuracy in the immediate test did not significantly differ from that in the delayed test. There were no significant two-way interaction effects. A three-way interaction among mode, exposure, and time approached significance [$F(1, 25) = 3.76$, $p = .06$], with accuracy on emotional, exposed voices slightly higher at the delayed test and accuracy on emotional, unexposed voices slightly lower at the delayed test. Accuracy of neutral voices did not differ whether or not they were exposed for both immediate and delayed tests.

To further determine whether recognition memory for emotional voices changes over time, we conducted planned contrasts separately for each time condition (immediate and delayed), with mode (emotional vs. neutral) and exposure (exposed vs. control) as within-subject factors.

Immediate test

The results indicated a significant main effect of exposure, [$F(1, 25) = 25.61, p < .001$], such that the recognition accuracy for exposed voices was greater than for control voices. Recognition of emotional voices was higher than for neutral voice, but the difference was not significant. There was no mode by exposure interaction.

Delayed test

When tested one week later, the results revealed a significant main effect of exposure, [$F(1, 25) = 22.52, p < .001$], indicating that recognition accuracy for exposed voices was still greater in comparison to control voices after a delay. After a delay, a significant main effect of mode [$F(1, 25) = 5.07, p = .033$] emerged, which reflects the fact that recognition accuracy for emotional voices was greater than it was for neutral voices and lower for control voices. Moreover, the difference of recognition accuracy between emotional-exposed and emotional-control excerpts was 21% at immediate testing; this difference increased to 34% at the delayed testing. This finding suggests a memory enhancement effect for emotional voices after a delay. There was a significant mode \times exposure interaction, [$F(1,25) = 4.44, p = .045$]. Post hoc pairwise comparisons revealed that recognition accuracy was significantly greater for emotional voices than for neutral voices only when the participants had been previously exposed to the voices, $p = .004$ (Fig. 1).

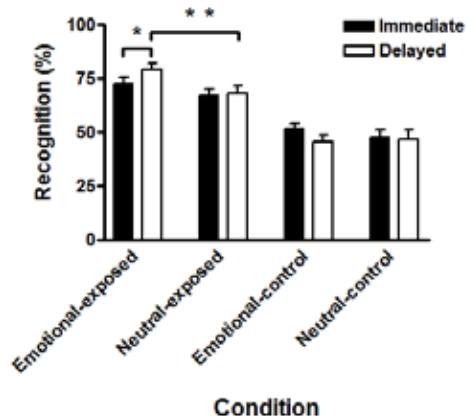


Fig. 1. Recognition memory performance at immediate and delayed tests. * $p < 0.05$, ** $p < 0.01$

The role of emotion in memory was further confirmed in paired samples (two-tailed) t-tests contrasting emotional-exposed voices and neutral-exposed voices at immediate versus delayed tests. For emotional-exposed voices, recognition accuracy was significantly better on the delayed test than on the immediate test [$t(25) = -2.23, p = .035$, two-tailed, See Fig. 1]. This observation is consistent with the findings

from ANOVA analyses ($p = .033$), suggesting that prioritized processing for emotional voices benefits the transfer of these voices into long-term storage. For the neutral-exposed voices, recognition accuracy did not significantly differ between the immediate and delayed tests.

In line with these findings, recognition accuracy for emotional-exposed voices was higher at both testing times, and, in this analysis, it was significantly higher than that of neutral-exposed voices on the delayed test [$t(25) = 3.12, p = .004$, two-tailed].

Confidence ratings

Confidence ratings were analyzed using the same method as described above for recognition accuracy. A repeated measures ANOVA with mode (emotional vs. neutral), exposure (exposed vs. control), and time (immediate vs. delayed) as within-subjects factors revealed a significant main effect of mode [$F(1, 25) = 45.07, p < .001$], such that confidence ratings were significantly higher for emotional voices than for neutral voices across all conditions. There was also a main effect of exposure [$F(1, 25) = 34.56, p < .001$], which reflected higher confidence for exposed voices compared to control voices. The results also showed a significant main effect of time, [$F(1, 25) = 4.30, p = .049$], suggesting that confidence ratings significantly decreased on the delayed test administered 1 week after encoding. Mean overall accuracy in the delayed test did not change, as high as 60% correct, but confidence ratings were overall lower.

As with the analysis used for recognition accuracy, confidence ratings were examined separately for immediate and delayed time condition. Thus, a 2 \times 2 repeated-measures ANOVA on the mean confidence ratings, separately for each time condition, with mode (emotional vs. neutral) and exposure (exposed vs. control) as within-subjects factors was conducted.

Immediate test

For the immediate test, there was a significant main effect of mode, [$F(1, 25) = 24.27, p < .001$], with higher confidence for emotional voices in comparison to neutral voices. There was also a significant main effect of exposure, [$F(1, 25) = 17.09, p < .001$], with higher confidence for exposed voices relative to control voices. There was no significant interaction between mode and exposure. Post hoc pairwise comparisons on each mode revealed that when the voices were emotional, participants' confidence ratings were significantly higher for recognizing exposed than control voices, $p < .001$ (Fig. 2). In contrast, when the voices were neutral, participants' confidence ratings did not significantly differ between recognition of exposed and control voices.

Delayed test

For the delayed test, both the main effects of mode [$F(1, 25) = 39.54, p < .001$] and exposure [$F(1, 25) = 31.26, p < .001$] were significant. As was the case in the immediate test, participants' confidence was higher in their memory for emotional vs. neutral voices and also higher for exposed vs. control voices. However, the interaction effect was no longer significant.

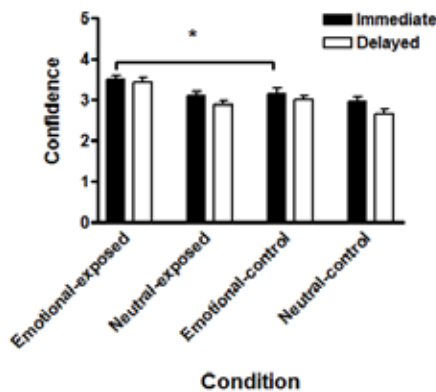


Fig. 2. Confidence ratings for immediate and delayed tests. * $p < 0.001$

IV. DISCUSSION

The current study explored the differential effect of emotionally expressive context on voice acquisition by comparing differences in strength of memory for emotionally nuanced voices relative to neutral voices. Based on the recognition accuracy data, we demonstrated that participants are able to acquire a voice as familiar following a single one-minute exposure. These findings suggest the possibility that humans possess remarkable capability to recognize and store a great deal of voice information in memory, which may have arisen early in human evolution [3].

In addition, our data showed that memory was selectively enhanced for emotionally expressive voices relative to neutral voices over a 1-week retention interval. This result for emotional voices suggests that emotionally nuanced context, mimicking naturalistic conditions, serves as a facilitator by inducting new voices into a repertory of personally familiar voices in long-term memory.

Analysis of confidence ratings further supported the preferential enhancement of memory for emotional voices. Specifically, participants were generally more confident when they recognized emotionally expressed voices than when they heard neutrally produced ones. Participants were also more confident in their recognition of exposed voices than they were in control voices.

The present findings are consistent with a wealth of studies demonstrating that emotional experiences hold a privileged access to memory [5, 6]. Furthermore, the findings are compatible with the neuropsychological data, suggesting differential brain mechanisms subserving processing of familiar and unfamiliar voices [7-9].

V. CONCLUSION

In summary, this study demonstrates for the first time that human voice patterns can be acquired following a single, 1-minute exposure when heard in a nuanced, emotionally arousing context. This opens up new approaches for understanding psychological processes underlying voice recognition.

REFERENCES

- [1] A. Searby, P. Jouventin, and T. Aubin, "Acoustic recognition in macaroni penguins: An original signature system," *Animal Behaviour*, vol. 67, pp. 615–625, 2004.
- [2] I. Charrier, N. Mathevon, and P. Jouventin. "Mother's voice recognition by seal pups," *Nature*, vol. 412, pp. 873, 2001.
- [3] D. Van Lancker Sidtis. "Ancient of Days: The vocal pattern as primordial big bang of communication," In P. Belin, S. Frühholz Eds. *The Oxford Handbook of Voice Perception*. Oxford University Press. 2018.
- [4] L. Cahill, R.J. Haier, J. Fallon, M.T. Alkire, C. Tang, D. Keator et al. "Amygdala activity at encoding correlated with long-term, free recall of emotional information," *Proceedings of the National Academy of Sciences USA*, vol. 93, pp. 8016–8021, 1996.
- [5] F. Heuer, and D. Reisberg. "Vivid memories of emotional events: The accuracy of remembered minutiae," *Memory & Cognition*, vol. 18, pp. 496-506. 1990.
- [6] K. S. LaBar and R. Cabeza. "Cognitive neuroscience of emotional memory," *Nat Rev Neurosci*, vol. 7, pp. 54-64, 2006.
- [7] D. Van Lancker and J. Kreiman. "Unfamiliar voice discrimination and familiar voice recognition are independent and unordered abilities," *Neuropsychologia*. vol. 25, pp. 829-834, 1987.
- [8] S. Lattner, M.E. Meyer and A.D. Friederici. "Voice perception: Sex, pitch, and the right hemisphere," *Human Brain Mapping*. vol. 24, pp. 11-20, 2005.
- [9] S.R. Schweinberger. "Human brain potential correlates of voice priming and voice recognition," *Neuropsychologia*, vol. 39, pp. 921–936, 2001.

ELECTRODERMAL ACTIVITY AND SPEECH FEATURES AS PREDICTORS FOR AROUSAL LEVEL CHANGES AFTER AFFECTIVE WORD PRONUNCIATION

C. Marzi¹, A. Greco^{2,3}, E. P. Scilingo^{2,3}, N. Vanello^{2,3}

¹ Institute for Computational Linguistics, National Research Council of Italy, Pisa, Italy

² Dipartimento di Ingegneria dell'informazione, University of Pisa, Pisa, Italy

³ Research Center "E. Piaggio", University of Pisa, Pisa, Italy

Email: claudia.marzi@ilc.cnr.it, alberto.greco@unipi.it, enzo.scilingo@unipi.it, nicola.vanello@unipi.it,

Abstract: This work explores the possibility of estimating subject arousal through the analysis of speech and electrodermal activity (EDA). One critical issue to be clarified is the reliability of EDA signal during speech production. To accomplish this task, a relation among EDA, speech activity and subject arousal during isolated affective word pronunciation task, will be investigated.

The results show that significant information on subject arousal can be still obtained by analyzing EDA during speech. In fact, a significant relationship between EDA features and self-reported arousal can be observed. In addition, a quantitative linear model relating EDA- and speech-related features could be identified.

These preliminary results indicate how the analysis of concurrent acquisition of EDA and speech deserves further attention and could offer a valid approach for the prediction of subject arousal during speech production, as a method for validating self-assessment ratings.

Keywords: speech, electrodermal activity, regression model, word pronunciation, arousal.

I. INTRODUCTION

Emotion recognition from speech is a complex task and still different issues remain to be solved. One source of complexity is the interplay between, internal, push factors, and external pull factors [1]. While the former are related to the effect of speaker emotional state, the latter are related to environmental factors or social rules that might influence speech production and hide the speaker real emotional status. Other approaches for emotion recognition try to exploit signals related to both central and autonomic nervous system (ANS) functions, such as heart rate variability, respiration and electrodermal activity (EDA) [2]. EDA refers to alterations in the conductance of the skin, due to changes in the sweat gland activity that are psychologically induced [3]. Since the sweat glands are directly

controlled by the sympathetic branch of the ANS, the EDA is considered as an effective correlate of the sympathetic nervous system activity [4].

Despite several attempts have been made to create a robust emotion recognition system using ANS correlates, this task suffers from specificity issues and is limited when applied to real world scenarios. For instance, EDA might be influenced by both respiration and speech, limiting its use and interpretation when speech activity is present [4]. Particularly, speech activity induces physiological irregular respiration that activates the sympathetic reflex and consequently affect the sweat gland dynamics. Accordingly, during experimental recordings, especially in the case of low-intensity stimulation, subjects are usually asked to avoid body movements, irregular respiration and speech activity [4, 5].

In this work, we aim at presenting preliminary findings about the concurrent use of EDA and speech features for the description of subject arousal level changes after emotional word pronunciation. More specifically, we propose a model able to predict the subject arousal related to the pronunciation of specific words, including both EDA and speech related features. Our goal is to improve current knowledge about the relation between speech and EDA, thus resulting in both a better comprehension of the influence of speech production on EDA, and in the development of emotion recognition systems merging both information channels.

II. METHODS

A. Experimental protocol.

Sixty Italian words have been selected from a set of 1121 words, which contains a translation from English into Italian of 1034 English words of the ANEW database, and 87 Italian words taken from a database of semantic norms [6]. Each word in the database is characterized by a distribution of arousal and valence ratings that were self-assessed by 1084 subjects participating in the study (on a scale ranging from 1 to

9) [7]. In the present study, we selected the sixty Italian words by controlling for length and elicited arousal level in the original large populations. Specifically, only single nouns were selected with a homogenous word length (6-7 characters). In addition, to obtain two homogenous groups characterized by low and high arousal level elicited in the speaker, we selected only words with ranked as low arousal (mean < 4.0 , standard deviation (SD) < 2.9) and high arousal (mean > 6.3 , SD < 2.2). Thirty numbers were also selected and considered as neutral.

Eighteen healthy volunteers, Italian native speakers (12 females, 6 males), were enrolled to take part in the study. Each participant was instructed to read aloud the words that appeared on the screen of a PC. Speech signal was recorded in a quiet room with low reverberation by means of a high-quality microphone, sampled at 48 kHz with a resolution of 32 bits (AKG Perception P220 Condenser Microphone, M-Audio Fast-Track). After an initial resting session of 3 mins, each word was shown for 2s and interspaced by the next one by 12s. Words belonging to the same arousal group (i.e., numbers, low arousal nouns, or high arousal nouns) were shown in succession, whereas the group order was randomized among subjects.

After the recording session, each subject was asked to score the arousal level elicited by the pronunciation of each word. The obtained arousal score was adopted for each word in the successive analysis, instead of the mean arousal level that was indicated in the database. A correlational analysis was preliminary performed between the arousal levels indicated by the database, and the scores assessed by the participants, revealing a significant statistical correlation ($r=0.67$, $p<0.001$).

B. EDA processing.

EDA was analyzed using the cvxEDA algorithm [8], which allows the decomposition of the skin conductance signal into its two basic components: a slow-varying tonic component, which represents the sympathetic tone, and an event-related fast-varying phasic component, which reflect the sympathetic arousal. After the decomposition process, we computed the mean of the tonic signal (MeanTonic) in the time-windows correspondent to each word reading. Moreover, we performed a frequency analysis in order to calculate the EDAsymp index, defined as the spectral power of EDA signal on the frequency band between 0.045 to 0.25Hz. This frequency band has been proved to be a reliable index of the sympathetic nervous system activity.

C. Speech processing.

For each spoken word, a first set of features comprising prosodic information as described by speech fundamental frequency (F_0) were estimated as well as

features related to the analysis of Mel-frequency cepstral coefficients (MFCCs). Specifically, the SWIPE' algorithm was used to obtain F_0 -derived features. These comprise measure of F_0 values within each word, as median and median absolute deviation (MAD), as well as geometric features describing F_0 profile. These include features borrowed from Taylor's tilt intonational model [10], and estimated for all voiced events [11]. The F_0 values belonging to each subject were normalized with respect to the values estimated with the number speech task, to account for subject specific mean value differences. In the present work we also included features parsimoniously describing the MFCC time dynamics within each word: specifically, the median and median absolute deviation (MAD) of the temporal changes of indexes of the largest MFCC at each time point, MedianMFCC and MAD_MFCC respectively, were taken into account.

D. Statistical Analysis Experimental protocol.

We applied different statistical models to highlight possible general relationship among the elicited arousal levels and the speech and ANS-related features.

Firstly, a linear regression approach was adopted to explore possible general relationship between the acquired features and arousal scores.

Linear fixed effects (LM) have been preliminary modelled by considering the arousal levels as dependent variable, with EDA features or speech related features as independent ones.

Then Linear mixed-effects (LME) have been modelled by adding both participants and words as random effects.

Finally, we evaluated a hierarchical linear model (HLM) to better explain each word arousal level, using EDA and speech related features as predictors. Specifically, a two-level model was estimated, with the first level describing within subject variability across different words, and with the second level representing the relationship between predictors and predicted values at group level. This model gives the opportunity of estimating a linear relation both at group and at single subject level. Both the model parameters and noise covariances are jointly estimated at both levels. At group level, the linear model parameters are obtained by weighting each subject contribution according to her/his reliability, that is related to the associated variance. An Expectation Maximization approach was used to jointly estimate the model parameters and their covariance at both levels under Gaussian assumptions. The statistical significance of regression parameter at group level was assessed using a t -test statistics related to the conditional mean of the parameter given the observations.

III. RESULTS

A. Linear model with fixed effects

A highly significant relation has been found between MeanTonic and the self-assessed arousal scores ($p < 0.001$). On the contrary, no significant relation has been found by modelling F_0 as predictor for arousal scores.

An interesting result was given by the relation between EDAsymp and arousal: EDAsymp is a statistically significant positive predictor for self-assessed arousal scores ($p = 0.003$). An even more robust model was obtained by adding, as categorical predictor, the a priori arousal classes of words (i.e. either low or high arousal), as confirmed by an ANOVA model comparison ($p < 0.001$).

When considering participants as independent categorical variable together with EDAsymp we obtained an explained variance (conditional R^2) of 0.37 and a statistical significance ($p < 0.001$) for the arousal class and the interaction with EDAsymp, but not for the predictor EDAsymp.

B. Linear model with fixed and random effects

By adding participants and words as random effects, we achieved a slightly better explained variance (conditional $R^2 = 0.40$), and a statistical significance for both EDAsymp ($p < 0.01$) and arousal classes ($p < 0.001$). Explained variance increases further ($R^2 = 0.42$) and predictors are thoroughly significant ($p < 0.001$) when independent variables are EDAsymp and participants, with words as random effects.

However, when we added to the mixed models some more predictors, e.g. the mean value of EDA tonic, the mean F_0 , MedianMFCC, MAD_MFCC, we did not derive robust significant predicting models.

All these models fitted only random intercepts, thus estimating common slopes for fitted covariates. A hierarchical approach is a principled solution to approach intercepts and slopes that may vary independently by a grouping variable, such participants are in our study.

C. Hierarchical Linear model

The HLM model was found to converge using the mean value of EDA tonic component, the mean word F_0 and MAD_MFCC. The regression coefficient related to EDA component was found to be statistically significant ($p = 0.02$), with a conditional $R^2 = 0.4$. Instead, despite the model convergence, word mean F_0 and MAD_MFCC related coefficients were found to be not significantly different from zero ($p = 0.07$ and $p = 0.4$, respectively).

Given the highly significant relation that the linear regression analysis revealed between EDAsymp and the

arousal level, indicated by each participant, an HLM model was adopted to predict this feature, instead of self-assessed arousal scores. In this case, the model converged to a solution ($R^2 = 0.5$) using Word_ F_0 , MAD_MFCC and MeanTonic also highlighting significant linear relation between Word_ F_0 and MeanTonic with this EDA-related features. Specifically, a statistically significant ($p = 0.013$) positive linear coefficient was found relating Word_ F_0 and EDAsymp. A positive linear relation ($p = 0.007$) was also found between MeanTonic and EDAsymp.

IV. DISCUSSION

This preliminary study highlights the reliability of EDA-related features during speech tasks. In fact, the linear regression analysis, with either fixed or fixed and random effects, could reveal a significant relation between arousal level and EDAsymp. However, the analysis performed with these models highlighted a high subject by subject variability, as confirmed by considering participants as categorical variable. This was found for instance when studying the relation between meanTonic, EDAsymp and arousal scores. These non-homogenous results could be related to different factors, such as the individual variability in the changes of speech-related parameters and emotions, or possible confounding factors as anxiety levels [11]. Moreover, a further source of variability can be related to the different relevance of each word for each subject. In addition, each word was shown in isolation, namely out of a specific context, thus allowing possible different interpretations by each participant.

The proposed hierarchical linear model specifies the group-level relation taking into account the variability of each subject. The hierarchical linear model weights the contribution of each subject, at group level, according to the inverse of the associated variance. Such a variability could have different sources, such as a difficulty in self reporting the elicited arousal level, or some other cognitive, behavioral factors affecting the task execution. As a matter of fact, the HLM could highlight a positive correlation between word self-reported arousal with tonic component of EDA. If on one side, this appears to be quite trivial, due to the link between EDA and arousal, on the other it reinforces the hypothesis that speech activity does not totally hide this information. Moreover, a sign of a complementary information achievable by analyzing speech and EDA was highlighted by the close-to-significance value obtained by Word_ F_0 .

The highly significant relation between arousal and EDAsymp, which is demonstrated to reflect the sympathetic nervous system activity, prompted us to use EDAsymp as a dependent variable in the HLM, instead of the self-assessed arousal scores. Indeed, the level of

arousal perception in humans is strongly related to the level of activation of the sympathetic nervous system. Interestingly, in this case, both the regression coefficients related to meanTonic and Word_F₀ were found to be significantly different from zero. Accordingly, this result suggests that a model taking into account both EDA- and speech- derived features can predict the arousal level perception elicited by word pronunciation.

This latter result, when compared with the one obtained using the self-assessed arousal level as dependent variable, leads to further considerations. The first one pertains to the reliability of self-assessed arousal in this word pronunciation task, that should be further investigated. Moreover, we cannot exclude that the high correlation between EDAsymp and Word_F₀ could be partially explained by speech-related artifacts in the EDA. However, the close-to-significance result pertaining to the Word_F₀ coefficient in the HLM, when predicting self-assessed arousal, seem to mitigate this scenario.

The improvement of model significance using EDAsymp with respect to the arousal score might be also related to its quantitative nature and not a discrete ordinal one as for the arousal scores. In fact, the use of the regression models under Gaussian assumptions to predict ordinal values, as those obtained from psychological scales is not optimal. In this case, other approaches can be considered. However, the proposed model has the advantage of easily accounting for replicates at single subject level and allows to easily model variable intercepts and slopes for each subject. Moreover, the HLM approach could also be extended to model noise covariance according to experimenter hypotheses, as for instance dependency among successive measures.

V. CONCLUSION

This preliminary work indicates that the informative content of EDA signal about subject arousal, is not hidden by concurrent speech activity. Moreover, both the mean value of EDA tonic component and speech fundamental frequency, were found to increase along with increasing arousal level, during word pronunciation.

This work points favorably towards the possibility of building a quantitative model of arousal jointly exploiting speech and EDA. This would represent a step forward with respect to a previous study [5] where an arousal classification results were improved by jointly considering speech and EDA features and a starting approach to clarify and disentangle the relative contribution of speech to EDA signal change.

REFERENCES

- [1] K. Scherer, "Vocal Affect Signaling: A Comparative Approach," *Advances in the Study of Behavior*, vol. 15, pp. 189–244, 1985.
- [2] S. Jerritta, M. Murugappan, R. Nagarajan, K. Wan, "Physiological signals based human emotion recognition: a review". In *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*, IEEE, pp. 410–415, 2011.
- [3] A. Greco, A. Lanatà, L. Citi, N. Vanello, G. Valenza, E.P. Scilingo. "Skin admittance measurement for emotion recognition: A study over frequency sweep". *Electronics*, vol. 5, n. 3, p. 46, 2016.
- [4] W. Boucsein. *Electrodermal Activity*, Springer Science & Business Media, 2012.
- [5] A. Greco, C. Marzi, A. Lanatà, E. Scilingo, and N. Vanello "Combining Electrodermal Activity and Speech Analysis Towards a more Accurate Emotion Recognition System", in *IEEE EMBC Conference*, 2019.
- [6] M. Montefinese, E. Ambrosini, B. Fairfield, and N. Mammarella, "The adaptation of the affective norms for english words (anew) for italian," *Behavior Research Methods*, vol. 46, no. 3, pp. 887–903, 2014.
- [7] M. Montefinese, E. Ambrosini, B. Fairfield, and N. Mammarella, "Semantic memory: A feature-based analysis and new norms for italian," *Behavior research methods*, vol. 45, no. 2, pp. 440–446, 2013.
- [8] A. Greco, G. Valenza, A. Lanata, E.P. Scilingo, "cvxEDA: A Convex Optimization Approach to Electrodermal Activity Processing", *IEEE Trans Biomed Eng.* vol. 63, n. 4, pp. 797–804, 2016.
- [9] A. Camacho, J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, 2008.
- [10] P. Taylor, "Analysis and synthesis of intonation using the tilt model," *The Journal of the acoustical society of America*, vol. 107, p. 1697, 2000.
- [11] A. Guidi, N. Vanello, G. Bertschy, C. Gentili, L. Landini, and E. Scilingo, "Automatic analysis of speech f₀ contour for the characterization of mood changes in bipolar patients," *Biomedical Signal Processing and Control*, vol. 17, pp. 29–37, 2015.

A COMMONLY SHARED CODE FOR SADNESS IN HUMAN VOCALIZATIONS AND MUSIC REVEALED BY HUMAN INFANT CRIES

G. Zeloni¹, F. Pavani²

¹ Società Psicoanalitica Italiana/International Psychoanalytical Association/Azienda USL Toscana Centro
²Center for Mind/Brain Sciences, CIMEC, University of Trento, Italy/Dep. of Psychology and Cognitive Sciences, University of Trento, Italy/Centre de Recherche en Neurosciences Cognitive, Lyon, France

Abstract: In Western music, minor chords, modes and intervals evoke sadness. This may reflect an emotional interpretation of intervals common to music and vocal expressions. Here we studied vocal expressions in pre-verbal infants (aged 6 to 8 months) to test whether intervals that typically evoke sadness in music (i.e., minor 2nd and minor 3rd) are more represented in cry compared to neutral utterances. Results showed that the unison, major 2nd, minor 2nd, major 3rd, minor 3rd, perfect 4th and perfect 5th are all represented in infant vocalizations. However, minor 2nd outnumbered other intervals in cry vocalizations, being present in 17 out of 20 of these audio files (52.2% of all measured intervals), but only in 4 out of 23 neutral audio files (13.7%). These novel findings suggest that the association between minor intervals and sadness may develop in human adults because a critically relevant social cue (infant cry) is concurrently characterized by the prevalence of minor 2nd and negative emotional valence.

Keywords: Sadness, minor second interval, music

I. INTRODUCTION

The link between language, music and emotions has attracted attention of philosophers and scientists for centuries. Yet, it remains unclear why minor chords, modes and intervals used in western music evoke sadness more than major ones [1]. It has been proposed that human vocal expressions and music communicate sadness through a commonly shared code, based on the perceived relation between pitches – i.e., musical intervals [2,3]. Curtis and Bharucha [3] recorded human actresses uttering bisyllabic speech samples to convey four different emotions (anger, happiness, pleasantness and sadness), and found that the interval between the two pitches of sad utterances approximated a minor 3rd. In the same study, they also found that participants listening to different musical intervals associated sadness with minor 3rd and with descending minor 2nd.

While Curtis and Bharucha [3] provide initial evidence in support of the hypothesis of a shared code

between human vocal expression and music based on musical intervals, it is possible that intentional vocal utterances of human actresses are contaminated by cultural influences, including the experience of musical intervals mostly associated with sadness. A stronger test of the shared code hypothesis would be to assess the proportion of the different musical intervals in non-intentional human utterances produced by human infants. In this work, we analyzed musical intervals in two types of vocalizations, cry and neutral utterances, of pre-verbal infants between 6 and 8 months of age. The hypothesis of a shared code for conveying sadness predicts a higher proportion of minor thirds and minor seconds in infant cry compared to infant neutral vocalization.

II. METHODS

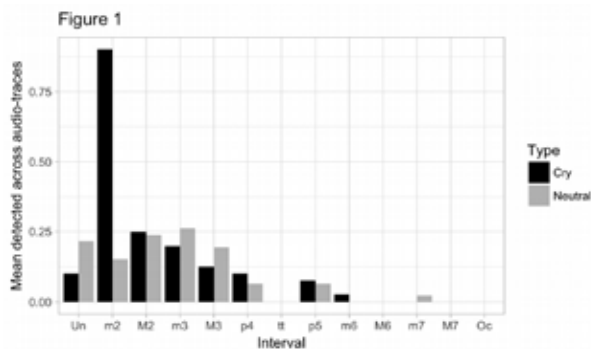
Infant cry and infant neutral vocalizations were selected from the Oxford Vocal (OxVoc) Sounds database [4], based on infant vocalizations free from background noise, and obtained from video recordings of infants filmed in their own homes during daily interactions with their primary caregiver. All infants were full-term, healthy, and aged between 6 and 8 months ($M = 6.7$ months, $SD = 0.9$; 9 males). Cry vocalizations ($n = 21$) occurred primarily when infants were separated from their caregivers, average $F0$ was 445.54 Hz ($SD = 84.81$) and average number of vocal bursts 1.90 ($SD = 1.09$). Neutral vocalizations ($n = 25$) occurred when infants interacted calmly with their caregiver, average $F0$ was 347.34 Hz ($SD = 122.34$) and average number of vocal bursts 1.88 ($SD = 0.97$). Cry vocalizations were rated as significantly more arousing and having negative valence compared to neutral vocalizations ($p < 0.0001$ on paired t-test [5]).

The procedure to identify musical intervals in each vocalization occurred in two steps. First, each vocalization was examined using an automated procedure for note identification developed at the University of Florence [5], that also provided a visual description of the overall melodic contour of each vocalization. Second, each vocalization together with

the results of the automated analysis procedure were provided to three expert judges, professional musicians and professors at the Florence Music Conservatory (Conservatorio Cherubini). All judges were naïve as to the purpose of the study, and were only instructed to identify audible notes and musical intervals in each vocalization using the provided materials and a piano. Each vocalization was discussed until collegial unanimous categorization was obtained; vocalizations for which such an agreement was not possible were excluded from further analyses ($n = **$ cry vocalizations; $n = **$ neutral vocalizations).

III. RESULTS

Musical intervals detected in both cry and neutral vocalizations were the unison (un), major 2nd (M2), minor 2nd (m2), major 3rd (M3), minor 3rd (m3), perfect 4th (p4), perfect 5th (p5). The other musical intervals [Tritone (tt), major 6th (M6), major 7th (M7) and octave (Oct)] were found neither in the crying nor in the neutral vocalizations (fig.1). To compare the presence of each of these intervals between the two groups of vocalizations (cry vs. neutral) we used the non-parametric Wilcoxon rank sum test in R studio. A significant difference was found only for the number of minor 2nd intervals ($W = 1372$, $p\text{-value} < 0.0001$). In cry vocalizations minor 2nd intervals were present in 17 out of 20 audio files (on average 0.90 times; total detected = 36; 52.2% of all measured intervals), whereas in neutral vocalizations they were present in 4 out of 23 audio files (on average 0.15 times; total detected = 7; 13.7% of all measured intervals). The number of ascending and descending 2nd minor intervals (14 vs. 22, respectively) in cry vocalizations was statistically comparable ($W = 172$, $p\text{-value} = 0.42$). For all other intervals no significant difference in number of occurrences between cry and neutral vocalizations emerged (all $p\text{-values} > 0.3$).



IV. DISCUSSION

The key finding of the present study is that more than half of the intervals detected in cry vocalizations in pre-verbal infants were minor 2nd, whereas this interval was minimally represented in neutral vocalizations. For long time the minor 2nd have been described as denoting melancholy and anguish [1,6]. The systematic association between the minor 2nd and negativity in adults have been interpreted as the consequence of automatic visceral affective response to dissonance, i.e., the way pairs of sounds physically interact in the auditory system [3], and/or cultural influences. The present findings offer an alternative interpretation, showing that human adults may develop this association because a critically relevant social cue (infant cry) is concurrently characterized by the prevalence of minor 2nd and negative emotional valence.

V. ACKNOWLEDGMENTS

Special thanks to Prof. De Lisi Leonardo, Prof. Fabbrini Gianni and Prof. Tomasello Santa of the Luigi Cherubini Conservatory of Music in Florence for their fundamental contribution to the analysis of audio files

REFERENCES

- [1] Maher, T. F., & Berlyne, D. E. (1982). Verbal and exploratory responses to melodic musical intervals. *Psychology of Music*, 10, 11–27.
- [2] Brown, S. (2000). The “musilanguage” model of musical evolution. In N. L. Wallin, B. Merker, & S. Brown (Eds.), *The origins of music* (pp. 271–300). Cambridge, MA: The MIT Press.
- [3] Curtis, M. E., & Bharucha, J. J. (2010). The minor third communicates sadness in speech, mirroring its use in music. *Emotion (Washington, D.C.)*, 10(3), 335–348. <http://doi.org/10.1037/a0017928>
- [4] Parsons, C. E., Young, K. S., Craske, M. G., Stein, A. L., & Kringelbach, M. L. (2014). Introducing the Oxford Vocal (OxVoc) Sounds database: a validated set of non-acted affective sounds from human infants, adults, and domestic animals. *Frontiers in Psychology*, 5(180), 562. <http://doi.org/10.3389/fpsyg.2014.00562>
- [5] Manfredi, Claudia*; Bandini, Andrea; Melino, Donatella; Viellevoye, Renaud; Kalenga, Masendu; Orlandi, Silvia (2018). Automated detection and classification of basic shapes of newborn cry melody. *BIOMEDICAL SIGNAL PROCESSING AND*

CONTROL, vol. 45, pp. 174-181, ISSN:1746-8094
DOI [Accesso ONLINE all'editore](#)

[6] Cooke, D. (1959). *The language of music*. London:
Oxford University Press.

SESSION V
VOICE QUALITY

TOWARDS ROBUST FEATURES IN VOICE DISEASE DIAGNOSIS: MFCCS VS. PNCCS

M. Madruga¹, Y. Campos-Roca², C. J. Pérez¹

¹Departamento de Matemáticas, Universidad de Extremadura (Cáceres, Spain)

²Departamento de Tecnología de los Computadores y las Comunicaciones, Universidad de Extremadura (Spain)
mariome@unex.es; ycampos@unex.es; carper@unex.es

Abstract: Voice impairment is a common condition for many people, and affects their daily lives and, in some cases, their professional lives too. There are many diseases that lead to that impairment, from physiological to neurological, but all of them require of a specialized doctor and complex techniques for an accurate diagnosis.

For that reason, research of computer aided diagnosis systems is of great interest. However, most studies about this subject isolate the voice signal with technical setups that can not be matched in real scenarios, like a physician's or a hospital diagnosis room.

Noise is a variable that should be considered if we intend to build cheap, portable systems that would help in early diagnosis. MFCCs are a set of features that has been widely used in this field for many years and has also been proposed as a discriminating feature set in other fields such voice recognition. Recent research in that last field suggest that there is another feature set, PNCCs, which offers an increased robustness against noise in voice recognition.

In this paper we address the robustness of these two feature sets comparing their ability to recognize healthy from pathological voice in the case of polyps and Reinke's Edema diseases.

Keywords: MFCC, Noise Robustness, PNCC, Polyps, Reinke's Edema.

I. INTRODUCTION

Voice is the main communication tool for humans and, therefore, its health is an important factor for their quality of life. For certain professions, such as teachers, singers or actors, voice is also the main working tool and they have a high risk of developing voice disorders for its excessive use, affecting not only their professional activities, but their everyday lives.

Voice disorders may have a wide variety of origins, from common neurological disorders such Parkinson's disease, to purely physiological reasons such vocal folds injuries [1]. The main methods to diagnose laryngeal diseases are direct inspection of the larynx

using laryngoscopy and videostroboscopy, complex techniques which are invasive, uncomfortable, require well trained specialists and dedicated tools and should be performed only in hospital environment; or evaluation of voice quality by direct auditions, strongly dependent on the physician's experience. Either way, currently diagnosis is a complex and expensive activity.

In recent years, computer-aided detection of diseases has attracted considerable scientific interest in search for an effective screening method for pathologies in an early stage, and voice related disorders are increasingly getting attention [1], [2]. However, such techniques are usually developed and tested in strictly controlled environments, ensuring optimal acoustical conditions, far from the real scenario of a physician's diagnosis room. It is important, therefore, to find an acoustical feature set which is robust against noisy conditions and could be useful in a wider range of clinical situations.

One widely used feature set is Mel Frequency Cepstral Coefficients (MFCCs) [3] traditionally used in speaker identification and voice recognition, which have been adopted in voice quality assessment [4], [5], [6]. An alternative to MFCC has been proposed in voice recognition: Power Normalized Cepstral Coefficients (PNCCs) [7], which show a better behavior under noisy conditions. The goal of this paper is to perform a comparison between the performance that both sets offer for voice disease diagnosis under realistic noisy conditions, including environmental and electrical noise.

II. METHODS

In order to perform a robustness comparison, we considered the Massachusetts Eye and Ear Infirmary (MEEI) Voice Disorders Database. This database features a wide range of voice diseases, including speech samples of 25 patients with Reinke's Edema and 20 patients with polyps, recorded at a sampling frequency of 25 KHz, and samples of 53 healthy voices, recorded at 50 KHz. Sustained phonations of the vowel /a/ were considered, which were recorded under very strict technical and acoustical conditions (sound proofed rooms and studio quality equipment).

The differences between healthy and pathological voices were addressed as follows: Regarding sampling frequency, all voice recordings were resampled at 25 KHz. Recording lengths vary among recordings, so the first second of each recording was used, trimming those whose length is greater.

Database recordings were performed under a strictly controlled acoustical environment, including the use of sound proofed rooms, which yields samples with very high signal to noise ratios (SNRs). As we stated before, those are not realistic conditions, so we simulated different noisy conditions making use of additive noise. We generated white random noise, using the random normal generator in Python Numpy library, creating a synthetic acoustical environment. Additionally, to create a more realistic acoustical environment, noise present in a diagnosis room located in Hospital San Pedro de Alcántara, Cáceres (Spain) was recorded. This room is partially acoustically isolated from surrounding aisles and waiting halls. The recording equipment was an AKG 520 head-worn condenser cardioid microphone attached to a TASCAM US322 interface, using Audacity 2.0.5 software. Recording length is 11 minutes 53 seconds and the sample rate 44.1 KHz.

Noise was added to the voice database using different SNRs, ranging from 30 dB to -10 dB in 2dB steps. Four types of noise were added: synthetic and actual noise, and within each type, two ways to select the noise sample were considered:

1. The same noise sample is added to all the recordings in the database.
2. A different noise sample is randomly selected for each of the database recordings, increasing variability.

In the case of actual noise, it was downsampled to 25 KHz to match the sample rate of the speech recordings.

Feature extraction and data analysis were performed using homebrew software. We chose Python 3.5 as a development environment for three main reasons: it is freely available, its quick prototyping characteristics, and the fact that there is huge number of free useful and reliable libraries available.

Acoustic features were extracted using freely available Python libraries: MFCCs were obtained from https://github.com/jameslyons/python_speech_features and PNCCs were taken from <https://github.com/supikiti/PNCC>. In both cases we used a window of 0.02 seconds with a stride of 0.01 seconds, FFT size of 512 samples and mel filterbank of size 60, which are the shared parameters, leaving the remaining ones with their default values.

The classification was performed using Support Vector Machine provided by Python SciKit library. A stratified random cross-validation scheme was applied.

For each disease considered we randomly generated 1000 sets of train-test indexes in a proportion of 2/3 to 1/3, which later would be used in every classification task for all SNRs. The reason for this is to avoid the variability that would result from generating a new train-test set for each classification task, and as a byproduct we reduced computational load. The selection method was designed to yield random stratified sets, so the healthy-diseased proportion remains constant and equal to the ratio shown in the original database, 25 Reinke's edema recordings out of 78 samples and 20 polyps patients out of 73 individuals. These sets were used to perform 1000 classification tasks for each SNR and we finally averaged their results.

III. RESULTS

We have classified the voice recordings using SVM within each SNR including the clean case (no noise added). The evolution of classification accuracy in relation to SNR is shown in the following plots. Each one depicts accuracy behavior in each of the 8 considered combinations of disease (polyps and Reinke's edema) and added noise (actual noise and synthetic noise, randomly chosen and fixed noise sample).

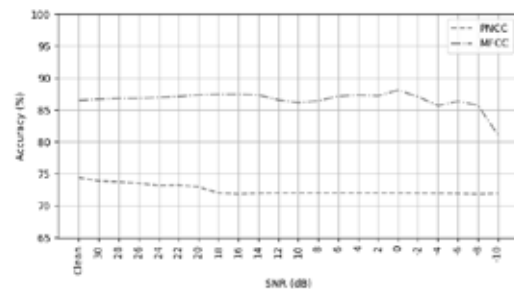


Figure 1 Classification accuracy versus SNR for polyps, actual noise fixed sample.

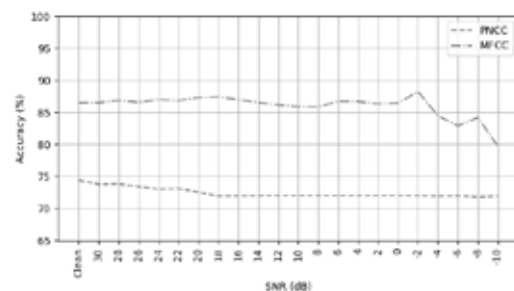


Figure 2 Classification accuracy versus SNR for polyps, actual noise random sample.

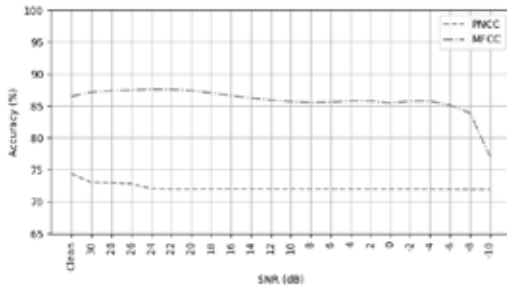


Figure 3 Classification accuracy versus SNR for polyps, white noise fixed sample.

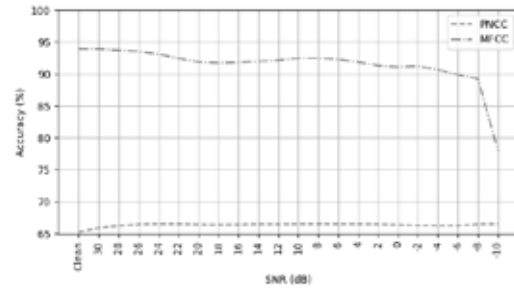


Figure 7 Classification accuracy versus SNR for Reinke's Edema, white noise fixed sample.

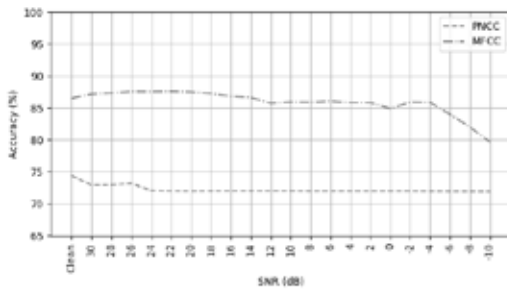


Figure 4 Classification accuracy versus SNR for polyps, white noise random sample.

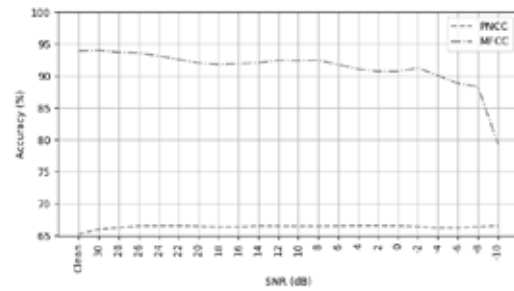


Figure 8 Classification accuracy versus SNR for Reinke's Edema, white noise random sample.

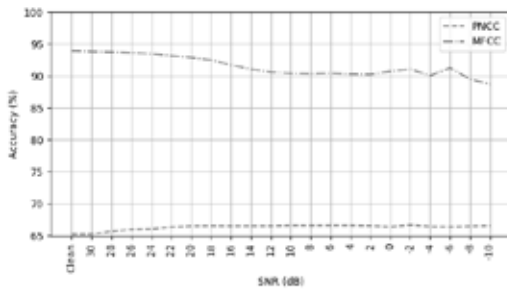


Figure 5 Classification accuracy versus SNR for Reinke's Edema, actual noise fixed sample.

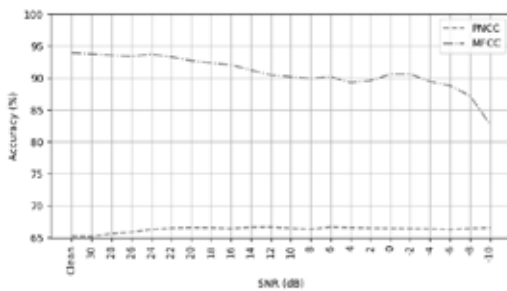


Figure 6 Classification accuracy versus SNR for Reinke's Edema, actual noise random sample.

IV. DISCUSSION

In the case of polyps, depicted in figures 1 through 4, we can see a consistent accuracy for MFCCs, close to 87.5%, in the case of clean recordings and high SNR down to 12dB, where we can see slight discrepancies in each of the noise conditions, as we can see if we compare figs. 1, 2 and figs 3, 4. Once passed the threshold of a SNR of 0dB, accuracy abruptly decays. PNCCs, on their side, show a steady behavior of an accuracy around 72,5% in the whole SNR range, no matter the noise level or noise conditions considered, except for the cases with very high SNR where their performance is slightly better, reaching in the clean case a 74% accuracy.

Reinke's Edema, on its side, shows an even higher accuracy in the case of MFCCs, almost reaching an accuracy of 95% in the case of clean recordings, and showing again a steady evolution as we increase noise and decrease SNR, slightly decaying until we, again, reach the threshold around -10dB, where drops to 65%, as we can see in figures 5 through 8. Figures 7, 8 show a slightly higher accuracy for SNR near the threshold in the case of white noise with respect to actual noise, as we see in figures 5, 6. PNCCs show again an almost constant accuracy, no matter the amount of noise added, but the level is very low, around 65%.

PNCCs show a somehow robust behavior, as their performance is almost constant in the whole SNR range, but with an accuracy of 65% for Reinke's

Edema, or 73% for Polyps, their ability to diagnose is clearly outperformed by MFCCs by a huge margin of 13% in the case of Polyps and an even greater one of 30% for Reinke's Edema.

This difference in accuracy obtained holds for a huge range of SNR, even in situations that we would never find in a hospital or diagnosis room (unrealistic low SNR).

V. CONCLUSION

We have shown that PNCCs cannot be considered as a substitute for MFCCs in computer-aided diagnosis of voice disorders, unlike in the case of voice recognition, as the performance obtained is considerably worse. MFCCs show a robust performance even in noisy conditions even in non-realistic situations. It is worth mentioning that, while overall results are quite similar for the four types of noise, white noise accuracy seems to be slightly higher than accuracy obtained for actual noise, which shows the possibility of actual noise being more intrusive than synthetic noise.

REFERENCES

- [1] Ladan Baghai-Ravary and Steve W Beet. Automatic speech signal analysis for clinical diagnosis and assessment of speech disorders. Springer Science & Business Media, 2012.
- [2] S. Hegde, S. Shetty, S. Rai, T. Dodderi, A survey on machine learning approaches for automatic detection of voice disorders, *Journal of Voice*, 2018.
- [3] K.S.R. Murty and B. Yegnanarayana. Combining evidence from residual phase and MFCC features for

speaker recognition, *IEEE Signal Processing Letters*, Vol. 13, No. 1, pp. 52-55, 2006

[4] J.I. Godino-Llorente, P. Gomez-Vilda, M. Blanco-Velasco. Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters, *IEEE Transactions on Biomedical Engineering*, Vol. 53, pp. 1943-1953, 2006

[5] R. Fraile, N. Saenz-Lechon, J.I. Godino-Llorente, V. Osma-Ruiz, C. Fredouille. Automatic detection of laryngeal pathologies in records of sustained vowels by means of mel-frequency cepstral coefficient parameters and differentiation of patients by sex, *Folia Phoniatica et Logopaedica*, Vol. 61, pp. 146-152, 2009

[6] A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig. Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity, *Journal of the Royal Society Interface*, Vol. 8, pp. 842-855, 2011a

[7] Chanwoo Kim and Richard M Stern. Power-normalized cepstral coefficients (PNCC) for robust speech recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(7):1315-1329, 2016.

ACKNOWLEDGMENTS

This research has been supported by project MTM2017-86875-C3-2-R (MINECO), and projects IB16054, GR18108 and GR18055 (Junta de Extremadura/European Regional Development Funds, EU).

ELECTROGLOTTOGRAPHIC VOICE MAPS OF UNTRAINED VOCALLY HEALTHY ADULTS WITH GENDER DIFFERENCES AND GRADIENTS

Rita R. Patel¹ and Sten Ternström²

¹ Dept. of Speech & Hearing Sciences, Indiana University, Bloomington, USA

² Division of Speech, Music, & Hearing, KTH Royal Institute of Technology, Stockholm, Sweden
patelrir@indiana.edu and stern@kth.se

Abstract: Baseline data from adult speakers are presented for time-domain parameters of electroglottographic waveforms. 26 vocally healthy adults (13 males and 13 females) were recruited for the study. Four dependent variables were computed: mean contact quotient, mean peak-rate-of-change in the contact area, index of contacting, and the audio crest factor. Small regions around the speech range distribution modes on the f_0 /SPL plane were used to define means and gradients. Males and females differed considerably in the audio crest factor of their speaking voice, and somewhat in their EGG contact quotient when measured at the mode point of the individual speech range profile. In males, contacting tended to increase somewhat with f_0 and SPL around the mode point, while in females it tended to decrease.

Keywords: Electroglottography, voice maps, FonaDyn, normal voice

I. INTRODUCTION

Non-invasive electroglottography (EGG) could be useful for pediatric voice assessment, but not enough is known about how child EGG signals compare to those of adults. As a preliminary to comparing adult voice production to that of children, a set of baseline adult data characterizing the EGG has been compiled.

Practically all voice parameters are sensitive to sound pressure level (SPL) and fundamental frequency (f_0). In conventional study paradigms, the location in the f_0 /SPL coordinate system (the *voice field*) [1] is typically selected by asking the subject to phonate “at a comfortable pitch and loudness” (or perhaps at a few different levels of vocal sound level and frequency). However, it is necessary to assess also the sensitivity to the location of sampling. A *voice map* is a representation of the 2D-distribution of some dependent variable over the voice field. On the voice map, the *gradient* of a given variable will quantify the sought sensitivity. Voice maps were acquired over both a speech range (SRP) and a voice range on sustained vowel (VRP). Over these ranges, three EGG wave shape parameters [5] and the audio crest factor were computed from the

recorded waveforms. The crest factor of a waveform is defined as the ratio of its peak amplitude to its RMS amplitude. In audio waveforms of open vowels, it is related to the maximum flow declination rate (MFDR), but it is much easier to acquire. Small regions around the speech range distribution *modes* were used to compute means and gradients of the four variables.

II. METHOD

Participants and setup. A total of 26 vocally healthy adults in the age range of 22-45 years ($M=28$ years) without any acute respiratory illnesses were recruited for the study. All subjects were non-smokers and non-classically trained singers, except two adults (1 male and 1 female) had approximately 10 years of vocal training. Participants received a cinema voucher for their efforts. Simultaneous EGG (Glottal Enterprises EG2, www.glottal.com) and acoustic recordings (head-mounted cardioid condenser microphone (AKG model C520, www.akg.com) at 6 cm mouth-to-microphone distance were made on 13 males and 13 females in a small room treated with absorptive acoustic materials, with a reverberation time $T_{60} < 0.1$ s. The microphone gain was calibrated for SPL for each participant by matching the level of a sustained vowel on the voice map display to that of the C-weighting sound level meter, held at 30 cm from the subjects’ mouth [2]. The EGG and acoustic signals were recorded in two separate channels at a sampling rate of 44.1 kHz and 16 bit quantization, using a digital audio interface (RME Fireface UCX, www.rme-audio.com) and a custom-developed public-domain real-time analysis software called FonaDyn v2.0.1 [3]. Voice map data from FonaDyn was then read into Matlab (vR2017b, www.mathworks.com) for making further statistics and graphs.

Acquisition. Each participant first produced 3 trials of reading a short phonetically balanced passage called the Rainbow Passage [4] at habitual pitch and loudness levels, to elicit spontaneous speech production for the speech range profile (SRP), for a total duration of about 6 minutes per subject. Then, as large a voice range as possible was elicited involving phonatory

productions of low to high pitch in soft and loud phonation, on the vowel /a/ [6].

Defining ‘mode regions’. Each cell in the voice map is one semitone wide and one decibel high. The distribution mode corresponds to the cell that contains the largest number of cycles, i.e. the most commonly occurring location on the f_0 /SPL plane. To make a robust sampling of values that would be representative of each subject’s voice during the reading task, the distribution mode of the voice map was broadened to a region of maximum occurrence, as follows. The cycle counts of all cells were sorted in descending order. The cells with the largest cycle counts were collected until 50% of the total number of EGG cycles seen during the entire task were accounted for. Even though this operation does not require the cells to be adjacent on the f_0 /SPL plane, it resulted in one small connected 2-D region of cells around the mode cell, in all subjects (Figure 1). This region will be denoted Γ (gamma). Typically, the male Γ contained 3000-6000 cycles and the female Γ about twice that, due to the higher f_0 .

For every individual EGG cycle, the following time-domain EGG shape parameters are computed and averaged per cell: (1) the quotient of contact by integration (Q_{ci}), (2) the normalized maximum rate-of-change in the contact area (Q_{Δ} a.k.a. $dEGG_{maxN}$), and (3) the index of contacting (I_c). Additionally, (4) the audio crest factor is similarly computed and averaged. Here, the values of the per-cell means were taken as dependent variables, with f_0 and SPL being the independent variables.

Statistics. For each dependent variable, the mean and standard deviation over Γ was computed. Also, the slope over Γ was computed in the directions of f_0 (horizontal) and SPL (vertical) by performing linear regression against f_0 and SPL, respectively. This gives estimates of how sensitive each variable is to the location of sampling. Note that unless one or both slopes are zero, the direction of the maximum gradient will be oblique with respect to the axes, and the combined slope will be somewhat greater than either of these two component slopes. However, the horizontal \times vertical slopes are easier to interpret and to report.

Average EGG wave shapes at the mode cell were computed for each subject, for visual assessment (Figure 2).

III. RESULTS

The results for the reading task are given in Table 1. The extents of the Γ regions were much the same for males and females, with standard deviations of about ± 1.3 semitones and ± 2 dB. The results for the full voice range task were highly variable, mostly because the task of controlling f_0 and SPL separately is an unusual one for untrained subjects, and the recommended protocol [6] could not be strictly followed.

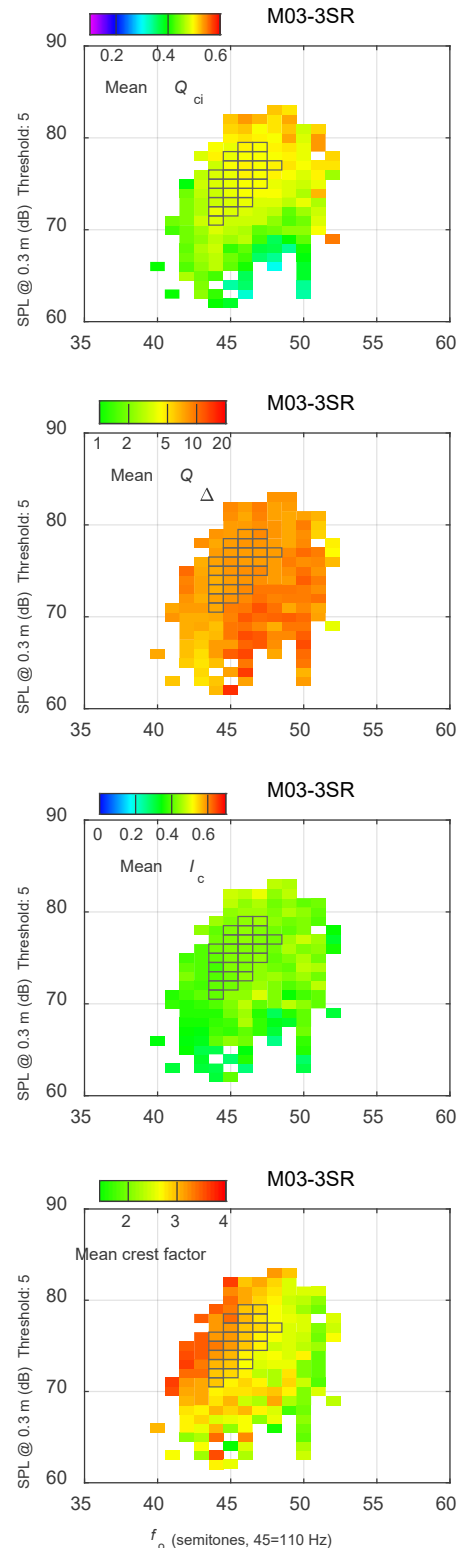


Figure 1. Example voice maps from a male subject, for three takes of the ‘Rainbow Passage’. The black grids show the region Γ for this subject. Only cells with 5 cycles or more are shown.

Table 1. Statistics over Γ regions, across subjects. The t -test (2-tailed, unequal variances) p -values refer to the females-males comparison; $p < 0.003$ with Bonferroni correction for 16 hypotheses corresponds to $p < 0.05$; significant p -values in **bold**.

Over Γ	Unit	FEMALES ($N=13$)			MALES ($N=13$)			t-test $p < 0.003$	
		means	\pm std	slope	means	\pm std	slope		
f_o	mean	Hz	190.2	16.9	116.0	12.0		1.2E-11	
	\pm std	ST	1.34	0.64	1.28	0.50		0.811	
SPL	mean	dB	71.66	1.61	73.58	3.21		0.071	
	\pm std	dB	1.88	0.39	2.12	0.45		0.167	
Q_{ci}	mean	-	0.42	0.03	0.46	0.04		0.032	
	slope vs f_o	1/ST	-0.0038	0.0074	-0.90%	0.0022	0.0053	0.48%	0.027
	slope vs SPL	1/dB	-0.0014	0.0036	-0.33%	0.0024	0.0031	0.53%	0.008
Q_{Δ}	mean	-	8.6300	2.2177	10.2231	2.8808		0.128	
	slope vs f_o	1/ST	-0.3641	0.3582	-4.22%	0.1512	0.3987	1.48%	0.002
	slope vs SPL	1/dB	-0.1789	0.1849	-2.07%	0.0119	0.1951	0.12%	0.017
I_c	mean	-	0.3854	0.0617	0.4463	0.0820		0.043	
	slope vs f_o	1/ST	-0.0129	0.0100	-3.34%	0.0073	0.0118	1.63%	9.8E-05
	slope vs SPL	1/dB	-0.0059	0.0058	-1.53%	0.0037	0.0050	0.82%	1.5E-04
Crest	mean	-	2.1408	0.1367	2.8518	0.2493		3.2E-08	
	slope vs f_o	1/ST	-0.0365	0.0319	-1.70%	-0.0215	0.0576	-0.75%	0.422
	slope vs SPL	1/dB	-0.0124	0.0231	-0.58%	0.0223	0.0265	0.78%	1.7E-03

Therefore, only an example of one subject is given here (Figure 3), with comments in the figure caption.

IV. DISCUSSION

Apart from the expected difference in mean f_o between sexes, the only mean difference that reached significance was the crest factor ($p < 10^{-7}$), while the contact quotient Q_{ci} came close ($p < 0.032$). Interestingly, when averaged across subjects, the relative gradients (% change per ST or dB) of all three EGG parameters were slightly positive for the males, on the order of +1%, and negative for the females, on the order of -2% (gray columns in Table 1). This difference was statistically significant for four out six slopes. Overall, about one third of the individuals did *not* follow this trend, or had a negligible gradient.

V. CONCLUSION

In vowels when reading a text, males and females differ considerably in the audio crest factor, and somewhat in the EGG contact quotient, when these are measured at the mode point of the individual SRP. Even within the confines of the small range of habitual speech, the amount of VF contacting in males increases somewhat with increasing f_o and SPL, while in females it decreases. When widening the inquiry from the speech range to the full voice range, such gradients are invariably large and individual, and need to be accounted for as the research question dictates. In running speech averaged over a few minutes, the sampling location still matters, if less critically, but the outcome

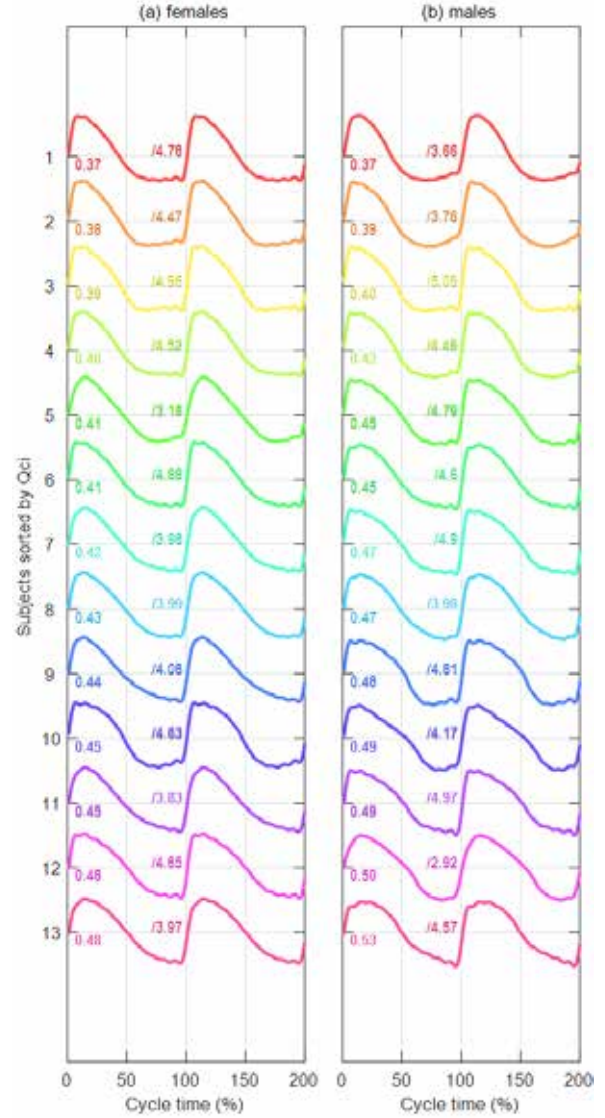
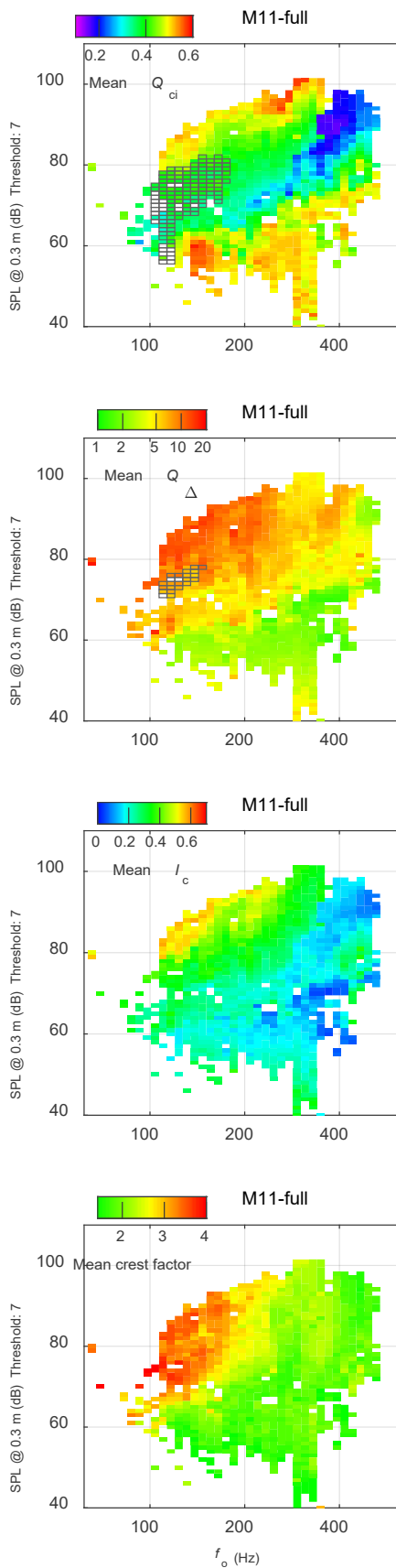


Figure 2. Most common EGG wave shapes of all subjects at their f_o /SPL modes, rebuilt from their 10 first Fourier components [3], and sorted by increasing contact quotient Q_{ci} (numbers at left); normalized in amplitude and time. The Q_{Δ} values in Table 1 are averages from the original waveforms, while those shown here (with /slash) are lower, since they were computed from these band-limited reconstructions.

is also less informative regarding the individual's phonatory function.

ACKNOWLEDGEMENTS

We are grateful for a guest researcher grant from the Wenner-Gren Foundations, supporting Rita Patel's stay at KTH, Stockholm, in the spring of 2019. Thanks also to Maria Södersten and Anita McAllister for helpful discussions and materials.



← Figure 3. Example of a FonaDyn full-range voice map, of male adult subject 11. Such maps are based on data from around 100,000 EGG cycles (5-8 minutes of vocalizing) on /a/. Only cells visited for 7 cycles or more are shown here. Each dependent variable is mapped into a separate layer of the map. (a) The quotient of contact by integration Q_{ci} (0.1 ... 0.6). Note how in this subject it becomes very small (deep blue) in the high falsetto range. Higher Q_{ci} values at low SPL are caused by incomplete VF closure and EGG noise. Below 50 dB SPL, the EGG noise makes the present data unreliable. The overlaid **black grid** shows the speech range of this subject. (b) The normalized peak dEGG Q_{Δ} (1...20) goes into the green (<2) when VF contacting is incomplete or absent. The overlaid black grid shows the Γ range of this subject. (c) The index of contacting I_c is defined as $Q_{ci} \times \log(Q_{\Delta})$ and thus combines the information in (a) and (b) (0 ... ≈ 1). (d) The audio crest factor gives an indication of the relative amount of high frequencies in the acoustic spectrum ($\sqrt{2}$... 4).

REFERENCES

- [1] Pabon, P. (2018). *Mapping Individual Voice Quality over the Voice Range, The Measurement Paradigm of the Voice Range Profile*, doctoral dissertation, KTH Royal Institute of Technology, Stockholm, Sweden. ISBN 978-91-7729-958-5.
- [2] Patel RR, Awan SN, Barkmeier-Kraemer J, Courey M, Deliyski D, et al. (2018) Recommended Protocols for Instrumental Assessment of Voice: American Speech-Language-Hearing Association Expert Panel to Develop a Protocol for Instrumental Assessment of Vocal Function. *Am J Speech Lang Pathol* June: 1-19.
- [3] Ternström S, Johansson D, Selamtzis A (2018) FonaDyn — A system for real-time analysis of the electroglottogram, over the voice range. *SoftwareX* 7: 74-80. doi:10.1016/j.softx.2018.03.002 . (OA)
- [4] Fairbanks G (1960) *Voice and articulation drillbook*. New York: Harper & Row, pp. 124-139.
- [5] Ternström, S. (2019). Normalized time-domain parameters for electroglottographic waveforms. *J Acoust Soc Am.*, 146, EL65–EL70, July 2019. (OA)
- [6] Pabon, P. (2012). *Standard Protocol for VRP Recording*, Voice Profiler Users Group. Online at kc.koncon.nl/staff/pabon/VRP/VPusersGroupStandardProtocolApril2012PPabon.pdf.

ACOUSTIC AND ELECTROGLOTTOGRAPHIC PARAMETRISATION OF PHONATORY QUALITY PROVIDE VOICE PROFILES OF PATHOLOGICAL SPEAKERS

Manfred Pützer¹, Wolfgang Wokurek²

¹ Language Science and Technology, Neurophonetics and Clinical Phonetics, Saarland University, Saarbrücken, Germany

² Institute for Natural Language Processing, University of Stuttgart, Stuttgart, Germany
puetzer@coli.uni-saarland.de, wokurek@ims.uni-stuttgart.de

Abstract: The present study firstly tries to find subgroups of pathological male and female phonation using data from a number of 534 pathological speakers. Secondly, this subgroup classification provides a basis for achieving voice profiles of pathological phonatory quality. Using complementarily orientated electroglottographic and acoustic parametrisation of phonatory quality, sustained vowel productions of 267 male and 267 female speakers were considered. In a first step, a clustering technique differentiates three subgroups within each gender on the basis of the EGG- and three subgroups on the basis of the acoustic parameters. In a second step, this subgroup definition allows one to present voice profiles of pathological speakers by combining the parameter means of the electroglottographically determined subgroups with those of the acoustically determined subgroups. The presented voice profiles provide a finer reference basis for the classification of different pathological phonation types as well as for the evaluation of shifts in individual phonation behaviour due to therapy or spontaneous recovery.

Keywords: Electroglottographic voice analysis — Acoustic voice analysis — Subgroup classification of pathological phonation — Voice profiles

I. INTRODUCTION

The evaluation of normal and/or pathological phonatory quality is often motivated by the demand to provide characteristic data of different voice profiles of normal and/or pathological speakers [1-4]. In so doing, the aim is to distinguish them from one another. A mean voice profile for individual speakers as well as a profile for a group of normal and/or pathological speakers can be defined by a number of voice-quality parameters, e.g. acoustic or electroglottographic parameters [5].

On the one hand, a profile of a group of pathological voices can be used as a reference to determine changes in individual phonation behaviour caused by any number of typical day-to-day factors affecting the speaker's phonation. On the other hand, it can also be used to determine the pathological status of

an individual speaker or to evaluate changes in individual phonation behaviour during and after therapeutic measures [6-8].

An important rationale to consider different voice profiles in a group of normal and/or pathological speakers as a potential reference basis for the evaluation of phonatory quality is based on the knowledge of a wide range of normal and pathological voices. Therefore, a clear definition of overall "normal and pathological phonatory quality" using different methods and different phonatory quality parameters cannot be given. Furthermore, it is not possible instrumentally to categorically delimit normal from pathological phonation behaviour. The latter point of view is also supported by studies which have not only documented a wide range of normal voices, but also of pathological voices [e.g., 4, 9]. Additionally, an overlap between normal and pathological voices has also been reported in these studies. The reasons for this lies for one thing in the fact that different (patho)physiological processes can belong to the same phoniatric classification and therefore deliver for example a comparable acoustic output and secondly, different sound pressure signals can arise from the same (patho)physiologic condition.

The goal of the presented study is therefore at first to show more adequately the above-mentioned and well known wide range for pathological voices by a definition of subgroups using data from a large number of pathological speakers. Secondly, the study is aimed at presenting voice profiles of pathological speakers on the basis of this subgroup definition. The subgroup definition and the presentation of voice profiles can be achieved on the basis of electroglottographic and acoustic measurements.

II. METHODS

A. Subjects

We studied a group of 534 pathological speakers (267 males and 267 females). The speakers ranged in age from 18 to 65 years. Mean age in the group was 38 years.

B. Voice Data

All signals are taken from the "Saarbruecken Voice Database" (www.stimmdatenbank.coli.uni-saarland.de) [10], which contains voice recordings from more than 2200 people. The speakers were selected randomly in equal numbers from the male and female "pathological" sections of the database. The continuous vowels [i:], [a:], and [u:] produced at the subjects' normal pitch were used for this study. Recordings had been acquired in the following way: Electroglottogram (EGG) and microphone signals were recorded simultaneously in a sound-treated room. The microphone signal was recorded using a headset condenser microphone (NEM 192.15, Beyerdynamic, Heilbronn, Germany) which fits comfortably over the ears and behind the head and allows the distance to the lips to be kept constant during speech, independent of head movements. The EGG-signal was acquired with a Portable Laryngograph from Laryngograph LTD. Both signals were fed directly into a Computerised Speech Lab (CSL) station (model 4300B) at a sampling rate of 50 kHz and with a 16-bit amplitude resolution. For each signal, a portion between positive zero-crossings of more than half a second was selected, starting 0.5 seconds after the beginning of phonation. The selected portion of the signal contains considerably more than the 20 to 30 pitch periods which, according to Klingholz [11] are needed to draw the conclusions about phonatory quality.

C. Data Analysis

The phonation behaviour was determined by combining the analysis of the EGG and the sound pressure-signal. The EGG-signal was analysed using the EGG program developed by Marasek [12].

The sound pressure signals were submitted to a method of analysis adapted by Wokurek and Pützer [13-15] which is based on voice parameters proposed by Stevens and Hanson [16]. This method uses the fact that the coarse spectral shape of the acoustic excitation is mainly defined by pitch cycle duration, duration of the open phase and airflow dynamics during closing and opening of the vocal folds. This excitation spectrum is modified by articulation via the vocal tract transfer function. Hence, to estimate parameters of the excitation spectrum the vocal tract resonances (formants) are removed from the recorded microphone signal. The parametrisation method relies on a sufficiently regular voice and uses the harmonic amplitudes to express the amount of spectral decay in various frequency regions as spectral decay gradients expressed in decibels per bark.

D. Statistical Analysis

The data were analysed using SPSS version 22. A clustering technique (k-means clustering) was executed using all seven EGG-parameters and all five acoustic parameters to differentiate subgroups of the pathological male and female speaker groups. The number of clusters was determined using a hierarchical cluster analysis. This method is used to find relatively homogeneous clusters of cases based on a set of measured characteristics. The subgroup classification (k-means clustering) provides a basis for achieving voice profiles of pathological phonatory quality. It was verified by means of a discriminant analysis (stepwise method) [4].

III. RESULTS

First, clustering results reveal three subgroups within each gender on the basis of the EGG, and three subgroups on the basis of the acoustic parameters. In Table 1 and 2 an example for subgroup definition on the basis of the EGG (Table 1) and acoustic parameters (Table 2), respectively, is given. Mean values of the parameters defining the three subgroups (with standard deviation and Wilk's lambda) are shown in table 1 for pathological female speakers and in table 2 for pathological male speakers. Statistically significant subgroup differences can be pointed out.

Table 1: Subgroups for pathological female speakers on the basis of the EGG parameters ($p < 0.001$).

Analysis approach	Parameters	Wilks-lambda	1 (n=147)	2 (n= 99)	3 (n= 61)
Phases of closure	SCV	0.248	1085.59 (502.64)	1736.29 (338.49)	781.87 (334.49)
	SCA	0.256	427.85 (260.83)	578.23 (310.48)	1449.82 (496.62)
	ECV	0.178	894.64 (190.29)	1280.75 (293.63)	1213.76 (285.06)
	CV	0.185	970.82 (195.50)	1434.68 (293.79)	1414.77 (328.24)

Table 2: Subgroups for pathological male speakers on the basis of the acoustic parameters ($p < 0.001$).

Analysis approach	Parameters	Wilks-Lambda	1 (n=56)	2 (n=63)	3 (n=32)
Open quotient	OQG	0.352	5.95 (1.81)	1.51 (1.73)	4.72 (1.88)
Glottal opening	GOG	0.435	0.16 (1.19)	0.38 (1.49)	-4.43 (1.66)

Second, this subgroup definitions allows one to present nine voice profiles of pathological speakers per gender by combining the parameter means of the electroglottographically determined subgroups with those of the acoustically determined subgroups. The

statistically different parameter values of the subgroups further reflect - for both genders - group specific differences of adduction and abduction behaviour. In addition, the achieved subgroups call attention to a *continuum* within pathological phonation, which is characterised by individually varying phonatory quality [4].

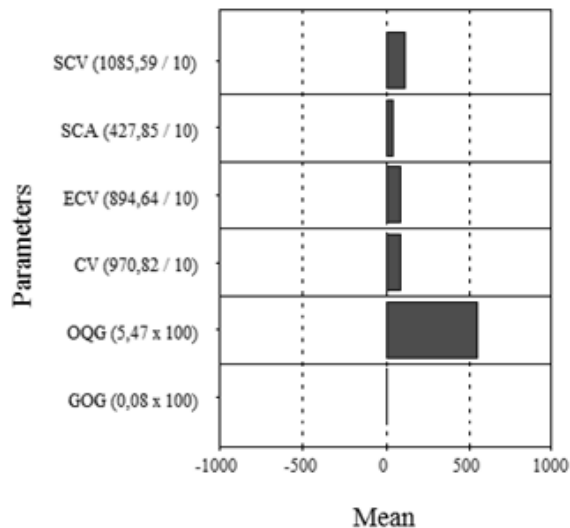


Fig. 1: Voice profile for pathological female phonation (means of electroglottographic and acoustic parameters).

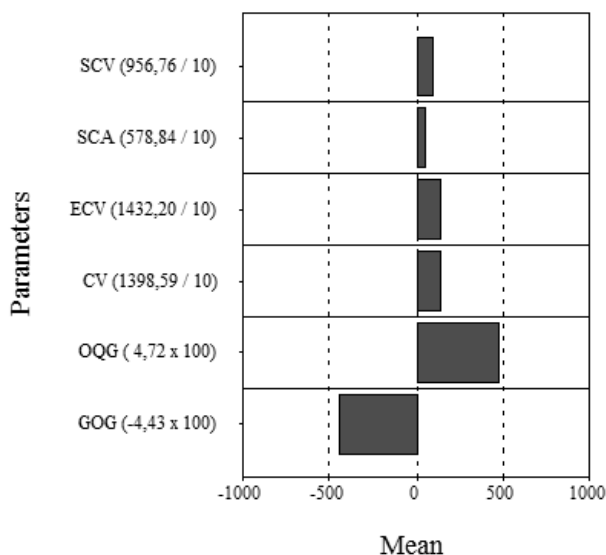


Fig. 2: Voice profile for pathological male phonation (means of electroglottographic and acoustic parameters).

Furthermore, the zero correlation scores suggest that the applied parameterisation of the two analysis approaches characterises, in a complementary manner, different parts of the phonatory cycle.

Thus, the complementary orientation of the two analysis approaches points to a potential for a finer differentiation of phonatory quality [5].

In Fig. 1 and 2 an example of a voice profile is given: Fig. 1 deals with pathological female phonation and Fig. 2 deals with pathological male phonation.

IV. DISCUSSION

In the present study, the EGG- and the sound pressure signal were submitted to two methods of analysis for defining phonatory quality of pathological male and female speakers. Significantly different parameter means underline the two methods for dimensioning pathological phonation. A finer-grained phonation quality definition is demonstrated by sub-dividing the two speaker groups (male and female speakers) into significantly different subgroups (three subgroups per gender). This calls attention to a *continuum* within pathological phonation, which is characterised by individually varying phonatory quality [4]. In particular, each individual phonatory quality may be localised within the parameter region of one of the subgroups of pathological male and female voices. Furthermore, it is shown in the study that the applied parameterisation of the two analysis approaches characterise in a complementary manner different parts of the phonatory cycle.

While the electroglottographic parameters analytically register certain phases of vocal fold vibrations and contain certain approaches of their quantification (variability of the start of the closing phase: SCV, steepness of the start of the closing phase: SCA, variability of the end of the closing phase: ECV, variability of the contact phase: CV) the acoustic analysis focuses on the open quotient (OQG) and the degree of glottal opening (GOG). In this manner, it is shown that the subgroup definition methodologically goes back to parameters which represent different phases and different approaches of quantification. A complementary orientation of the two analysis approaches is demonstrated. This complementary orientation points to a potential for a finer differentiation of phonatory quality.

A further main result of the study is that different voice profiles have been achieved on the basis of these above-mentioned subgroups. The voice profiles resulting from these dimensions in their individually different interactions ultimately demonstrate phonatory quality. The above-mentioned complementary orientation of the two analysis methods can also be shown by the different parameters constituting particular voice

profiles. A comparison of the relevant parameters of the two analysis methods used to characterise pathological male and female voice profiles complement each other in the representation of the physiology of the pathological glottal cycle. Considering these profiles, firstly, a finer basis of reference for the classification of individual changes in pathological phonation is provided. Secondly, this basis can also serve to determine the potential pathological status of an individual speaker and can be important for the evaluation of shifts in individual phonation behaviour due to therapy or spontaneous recovery. In this way, evidence for the possibility of controlling for individual phonation variation is more precisely given. The "voice" as a complex, multi-dimensional phenomenon is once again apparent.

V. CONCLUSION

The presented study demonstrates the need for a differentiated classification of pathological phonation into subgroups. Furthermore, on the basis of these subgroups voice profiles can be defined. They are intended to serve as a basis of reference in prospective studies to evaluate phonatory quality for different normal and pathological phonation types as well as to verify shifts in individual phonation behaviour due to therapy or spontaneous recovery. In view of the different subgroups it is essential that a combined auditory and instrumental study should be carried out to ascertain whether the subgroups are auditorily indistinguishable, or whether there are auditory differences that are, at present, not captured by the differentiating electroglottographic and acoustic parameters.

REFERENCES

- [1] J. Muñoz, E. Mendoza, M.D. Fresneda, G. Carballo, and P. López, "Acoustic and Perceptual Indicators of Normal and Pathological Voice," *Folia Phoniatr Logop* 55, pp. 102-114, 2003.
- [2] T. Bhuta, L. Patrick, and J.D. Garnett, "Perceptual evaluation of voice quality and its correlation with acoustic measurements," *J Voice* 18, pp. 299-304, 2004.
- [3] J. Koreman, M. Pützer, and M. Just, "Correlates of varying vocal fold adduction deficiencies in perception and production: Methodological and practical considerations," *Folia Phoniatr Logop* 56, pp. 305-320, 2004.
- [4] M. Pützer and W.J. Barry, "Instrumental dimensioning of normal and pathological phonation using acoustic measurements", *Clin Linguist Phon* 22 (6), pp. 407-420, 2008.
- [5] M. Pützer and W. Wokurek, "Stimmprofile zur Normalstimme auf der Grundlage akustischer und elektroglottographischer Analysen", *Laryngo Rhino Otol* 94(5), pp. 303-310, 2015.
- [6] P. Zwirner and E. Kruse, "Wertigkeit der akustischen Stimmanalyse für die Qualitätssicherung in der Behandlung von Stimmstörungen", in *Aktuelle phoniatisch-pädaudiologische Aspekte*, M. Groß Ed. Berlin, Renate Gross Verlag, 1995, pp 29-32.
- [7] P. Zwirner, D. Michaelis, and E. Kruse, "Akustische Stimmanalysen zur Dokumentation der Stimmrehabilitation nach laserchirurgischer Larynxkarzinomresektion," *HNO* 44, pp. 514-520, 1996.
- [8] M. Pützer, "Stimmqualität und Artikulation bei Dysarthrophonien in der individuellen, tendenziellen und referentiellen Bewertung. Ein instrumenteller Beitrag zu Phonations- und Artikulationsvariationen", *PHONUS* 13, Institut für Phonetik, Universität des Saarlandes, Habilitationsschrift, Saarbrücken 2008.
- [9] V. Parsa and D.G. Jamieson, "Acoustic discrimination of pathological voice: sustained vowels versus continuous speech," *J Speech Hear Res* 44, pp. 327-339, 2001.
- [10] M. Pützer and W.J. Barry, "Saarbrücken Voice Database", Institute of Phonetics, Saarland University, available at: <http://www.stimmdatenbank.coli.uni-saarland.de>
- [11] F. Klingholz, "Jitter," *Sprache Stimme Gehör* 15, pp. 79-85, 1991.
- [12] K. Marasek, "Electroglottographic Description of Voice Quality", *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung* Vol. 3 (2), Stuttgart 1997.
- [13] A. Schweitzer, W. Wokurek, and M. Pützer, "Convergence of Harmonic Voice Quality Parameters in Spontaneous Dialogues", in *Proc. 19th ICPhS 2019*, Melbourne, pp. 363-367, 2019.
- [14] M. Pützer and W. Wokurek, "Multiparametrische Stimmprofil-differenzierung zu männlichen und weiblichen Normalstimmen auf der Grundlage akustischer Analysen," *Laryngo Rhino Otol* 94(5), pp. 303-310, 2015.
- [15] 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, 2006.
- [16] 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, pp. 147-170, 1998.

VOCAL QUALITIES OF SARCASTIC UTTERANCES: CROSS-LINGUISTIC STUDY OF ENGLISH AND KOREAN

Seung-yun Yang

Department of Communication, Arts, Sciences, and Disorders, Brooklyn College/CUNY, New York, NY, USA
Brain and Behavior Laboratory, Nathan Kline Institute, Orangeburg, NY, USA
Seung-yun.yang@brooklyn.cuny.edu

Abstract: Sarcasm, a form of nonliteral language, refers to the use of words that mean the opposite of what is intended. Sarcasm is prevalent in everyday conversation and the appropriate understanding of the speaker's intent is crucial for successful communication.

Vocal qualities associated with sarcastic and literal utterances were investigated using listeners' ratings and acoustic analyses. Listeners' ratings revealed that nasal, breathy, and pharyngealized vocal quality is associated with sarcastic utterances. Acoustic analyses showed that jitter and harmonics-to-noise ratio were significantly different between sarcastic and literal utterances. Correlation between listeners' ratings and acoustic analyses were seen.

The results support listeners' ability to discriminate sarcastic from literal utterances. The results also support that specific vocal qualities were associated with sarcastic utterances.

Keywords: sarcasm, vocal quality, acoustic analyses, listeners' rating

I. INTRODUCTION

Sarcasm, frequently used in everyday communication, refers to an utterance whose intended meaning is opposite to the meanings of words comprising that utterance. Previous listening studies and acoustic analyses have revealed that difference in prosody between sarcastic and literal utterances are salient enough for listeners to identify sarcastic utterances without any information of the discourse context. It was also revealed that specific prosodic cues (pitch/fundamental frequency (F0) and speech rate) were suggested to be associated with sarcastic utterances [1], [2], [4 – 8].

Acoustic analysis of prosody has advantages in being quantitative and reliable. Standard acoustic cues of prosody such as duration and fundamental frequency

(F0) can be captured by acoustic analyses. However, acoustic measurement does not always match perceptual judgement. Perceptual assessment can provide a more comprehensive evaluation of prosody. Vocal quality of prosody requires multidimensional assessment and may not be properly captured by acoustic analysis. Previous research includes acoustic analyses to examine acoustic cues associated with sarcastic utterances; however, few systematic listening studies have examined the perceptual vocal qualities of sarcastic utterances in the context of acoustic analysis.

The purpose of this study was to systematically investigate what voice qualities optimally categorize utterances into different types of sentences (sarcastic and literal sentences) utilizing perceptual vocal quality rating scales and acoustic analyses. This study also examined the relationship between vocal quality rating and acoustic cues.

II. METHODS

Stimuli acquisition: Five native English speakers and 5 native Korean speakers participated to produce 6 sentence pairs (sarcastic and literal utterances). The speakers produced short target utterances in response to disambiguating linguistic-situational contexts provided by the experimenter.

Vocal quality rating: Ten native English speakers and 10 native Korean speakers participated as raters for vocal quality. The raters were students in Master's program in Communication Disorders and had had knowledge of pathological and non-pathological voices. All English participants were born, raised, and educated in the States and all Korean participants were born, raised, and educated in Korea. Raters rated the utterances produced by 5 English and 5 Korean native speakers utilizing 6 voice quality features: breathy, whispery, creaky, harsh, pharyngealized, nasal. For each of the 6 voice quality characteristics, a 5 point equal-appearing interval scale was utilized. The scale ranged from 1 to 5, with 1 referring to the absence of the feature in question.

Acoustic analyses: All the utterances produced by native speakers of English and Korean were analyzed with the Praat Software (V.6.0.36). Acoustic parameters related to vocal quality were selected: jitter, shimmer, and harmonic-to-noise ratio (HNR).

Statistical analyses: Pair-wise-t-tests were conducted to compare between sarcastic and literal utterances with respect to each vocal quality feature and acoustic cues. Correlation coefficients were calculated to examine the relationship between the listeners' ratings and acoustic cues.

III. RESULTS

Paired-samples t-test were performed to examine performance of vocal quality ratings by each language group on the two sentences types (sarcastic or literal). Of the 6 parameters presented to listeners for ratings (breathy, whispery, creaky, harsh, pharyngealized, nasal), only one emerged as significant in both English and Korean. Listeners' vocal quality ratings revealed that English and Korean sarcastic utterances were rated to be nasal compared to literal utterances ($t(29) = 5.324$; $p < 0.001$] for English; $t(17) = 2.546$; $p < 0.001$] for Korean).

When language groups were individually assessed, breathy and pharyngealized vocal qualities were also associated with sarcastic utterances. It was revealed that Korean sarcastic utterances were produced with breathy voice quality compared to literal utterances in Korean ($t(29) = 2.346$; $p = 0.009$] and English sarcastic utterances were produced with pharyngealized vocal quality ($t(29) = 2.546$; $p = 0.007$]).

Acoustic analyses revealed that there were significant differences in jitter (local) between English sarcastic and literal utterances. Higher jitter (local) was associated with English sarcastic utterances ($t(29) = 1.252$; $p = 0.003$]). Significant differences in harmonics-to-noise (HNR) emerged between Korean sarcastic and literal utterances. Sarcastic utterances were produced with lower HNR compared to literal utterances in Korean ($t(29) = -2.382$; $p = 0.001$]).

The relationship between vocal quality ratings and acoustic analyses was examined. Correlation analyses between vocal quality ratings and acoustic analyses showed that HNR moderately correlated with breathiness ratings [$r = 0.31$, $p < 0.001$].

IV. DISCUSSION

The present study investigated vocal qualities that possibly differentiate sarcastic and literal utterances in English and Korean by utilizing subjective (vocal quality ratings) and objective (acoustic analyses) measures. Results from listeners' vocal quality ratings revealed that certain vocal qualities are associated with sarcastic utterances. English sarcastic utterances were judged to be produced with nasal and pharyngealized vocal qualities and Korean sarcastic utterances were rated to produce with nasal and breathy vocal qualities. Acoustic analyses also confirmed that sarcastic utterances were produced with different vocal qualities from literal utterances. Higher jitter (local) was associated with English sarcastic utterances and lower HNR was associated with Korean sarcastic utterances. The correlation between HNR and breathiness vocal quality is consistent with previous findings that HNR is sensitive to the presence of breathiness [3].

The results were in accordance with previous findings that specific acoustic cues and vocal qualities were associated with sarcastic utterances.

V. CONCLUSION

The results of this study contribute to our understanding of production differences between sarcastic and literal utterances. Sarcastic utterances were produced with specific vocal qualities compared to literal utterances. The results also supports native listeners' ability to identify vocal qualities associated with sarcastic utterances.

REFERENCES

- [1] L. Anolli, R. Ciceri, and M.G. Infantino, "Irony as a game of implicitness: Acoustic profiles of ironic communication," *J Psycholinguist Res*, vol. 29, pp. 275-311, 2000.
- [2] S. Attardo, J. Eisterhold, J. Hay and I. Poggi, "Multimodal markers of irony and sarcasm," *Humor*, vol. 16, pp. 243-260, 2003.
- [3] G. de Krom, "Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments," *J Speech Lang Hear R*, vol. 38, pp. 794-811, 1995.
- [4] L. Milosky and C. A. Wroblewski, "The Prosody of irony," Paper presented at the International Society for Humor Studies Conference, Ithaca, NY 1994.
- [5] P. Rockwell, "Lower, slower, louder: Vocal cues of sarcasm," *J Psycholinguist Res*, vol. 29, pp. 483-495, 2000.
- [6] P. Rockwell, P. "Sarcasm on television talk shows: Determining speaker intent through verbal and

nonverbal cues,” in *Psychology of Moods*, A. Clark Eds. Nova Science, New York, 2005, pp. 109-140.

[7] P. Rockwell, “Vocal features of conversational sarcasm: A comparison of methods,” *J Psycholinguist Res*, vol. 36, pp. 361-369, 2007.

[8] M. Shapely, M. “Prosodic variation and audience response,” *Papers in Pragmatics*, vol.1, pp. 66-79, 1987

SESSION VI
VOCAL FOLDS DYNAMICS

PERTURBATION OF TIMES AND MAGNITUDES OF CYCLE MAXIMA OBSERVED IN DIPLOPHONIC VOICES

P. Aichinger

Medical University of Vienna, Department of Otorhinolaryngology,
Division of Phoniatics-Logopedics, Vienna, Austria
philipp.aichinger@meduniwien.ac.at

Abstract: A model of diplophonic waveforms capable of estimating modulation noise from observed waveforms is presented. Twenty-nine glottal area waveforms (GAWs) and audio waveforms obtained simultaneously are analysed via modelling. Analysis involves (i) fitting of a quasi-static waveform model, which includes extraction of up to two simultaneous fundamental frequencies (f_o s), and (ii) improvement of the fitting via estimating times and magnitudes of cycle maxima for each cycle and f_o individually. The modulating model is shown to outperform the quasi-static model quantitatively in the least-squares sense. Also, qualitative improvement is observed exemplarily via visually comparing recorded and modelled waveforms, as well as magnitude spectrograms thereof.

Keywords: Diplophonia, glottal area waveforms, audio waveforms, perturbation.

I. INTRODUCTION

Diplophonia is a voice phenomenon characterized on the perceptual level by two simultaneous pitches, which may be a sign of a voice disorder. In current clinical practice, auditory detection of diplophonia is performed by medical professionals to support the indication, selection, evaluation, and optimization of clinical treatment techniques. A point of criticism may be the subjective nature of auditory detection, which results in significant divergences between ratings obtained repeatedly (cf. intra-rater variability), at different medical centres (cf. inter-centre variability), or by different professionals (cf. inter-rater variability). Also, diverging definitions of diplophonia exist, which further impedes clinical and scientific communication.

A means for improving our understanding of diplophonia is the modelling of diplophonic waveforms. Diplophonic waveforms were recently modelled as the sum of two cyclic waveforms with different fundamental frequencies [1]. Other approaches include mechanical models [2,3]. A frequent assumption is that model parameters do not vary within analysis frames. In particular, features that

were neglected in the past include random perturbations, i.e., jitter and shimmer, as well as deterministic co-modulation of the coupled glottal oscillators. This presentation seeks to relax these limitations. The presented model is purely kinematic, and model parameters are extracted from recorded signals automatically. No assumptions regarding the cyclicity of modulation noise are made.

II. METHODS

A database of laryngeal high-speed videos and simultaneous audio recordings is used [4]. Twenty-nine voice samples are selected for analysis as follows. Via auditory and visual inspection, thirty-eight samples that only contain diplophonia and monophonia during phonation are selected. Five voice samples are discarded because visibility of vocal fold vibration is impeded. Additional four voice samples are discarded, because the examiner's voice is audible during the patient's phonation. GAWs and simultaneous audio recordings are analysed. GAWs are scaled and detrended before analysis.

A modulating model is compared to a quasi-static model. The modulating model is capable of varying the time and magnitude of each cycle maximum. First, f_o candidates are extracted via spectral peak picking and repetitive execution of the Viterbi algorithm. Second, a candidate waveform is obtained for each f_o candidate. A unit-pulse train is created for each f_o candidate, and cross-correlated with the recorded signal, yielding a pulse shape. The pulse shape is transformed to the frequency domain and input to a Fourier synthesizer driven by f_o . The Fourier synthesizer puts out candidate waveforms.

Third, f_o s are selected to minimize the root mean square (RMS) difference between the model waveform and the recorded waveform $d(n)$, i.e., the model error waveform $e(n)$. The model waveform is the sum of the selected candidate waveforms. No more than two f_o s may be active simultaneously.

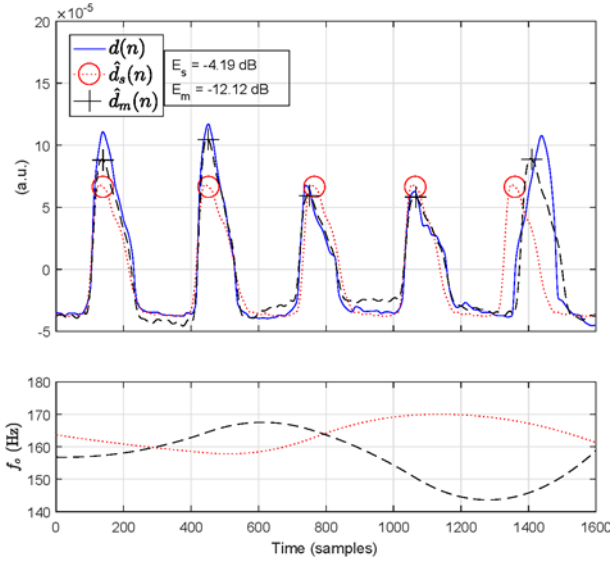


Figure 1: Example of waveforms involved in the modelling of a monophonic snippet of a clinically observed glottal area waveform (GAW) $d(n)$. The top plot shows the GAWs, and the bottom plot the instantaneous frequencies.

For the modulating model only, the times and magnitudes of the cycle maxima are optimized to further minimize the model error. Each change of times and magnitudes of maxima involves resynthesis of the model waveform, i.e., reestimation of the pulse shapes, Fourier transformation thereof, and Fourier synthesis of the waveform model's summands, i.e., the candidate waveforms. The Fourier synthesizer is driven by instantaneous f_o s, and its output is multiplied by an amplitude modulation function. Instantaneous f_o s and the amplitude modulation function are estimated from the modulated quasi-unit pulse trains.

In more detail, a candidate's pulse train estimate

$$\hat{d}(n) = \hat{A}(n) \cdot \sum_{p=1}^{10} \left[\hat{a}_p \cdot \cos(p \cdot \hat{\theta}(n)) + \hat{b}_p \cdot \sin(p \cdot \hat{\theta}(n)) \right] \quad (1)$$

is obtained for each selected f_o candidate, where n is the discrete time index, $\hat{A}(n)$ is the estimate of the amplitude modulation function, p is the partial index going from 1 to 10, \hat{a}_p and \hat{b}_p are the Fourier coefficients of the pulse shape estimate $\hat{r}(l)$, and $\hat{\theta}(n)$ is the estimate of the instantaneous phase. The candidate index γ is left out for convenience.

The pulse shape estimate $\hat{r}(l)$, the amplitude modulation function estimate $\hat{A}(n)$, and the instantaneous phase estimate $\hat{\theta}(n)$ are obtained as follows.

First, the quasi-unit pulse train

$$u(n) = \sum_{\mu} \hat{s}(\mu) \cdot \delta[n - \mu \cdot \hat{N}_0 - \hat{j}(\mu) - \Delta_{\Phi}], \quad (2)$$

where $\mu \in \mathbb{Z}$ is the pulse index, the amplitude perturbation $\hat{s}(\mu)$ is vector initialized as $\hat{s}(\mu) = 1 \forall \mu$, the delta function $\delta(x=0) = 1$ and $\delta(x \neq 0) = 0$, the timing perturbation vector $\hat{j}(\mu)$ is initialized as a zero vector, and a phase shift Δ_{Φ} is initialized as 0. The cycle length $\hat{N}_0 = f_s/f_o$ is obtained for f_o candidates, which are upsampled from a hop size of 16 ms to a sampling frequency f_s of 50 kHz using shape-preserving piecewise cubic interpolation.

The pulse shape estimate

$$\hat{r}(l) = \frac{1}{\sum_n u_{lin}(n)} \cdot \sum_n u_{lin}(n) \cdot d_{lin}(n-l), \quad (3)$$

which is the normalized cross-correlation of the quasi-unit pulse train $u_{lin}(n)$ and the recorded signal $d_{lin}(n)$, the instantaneous phases of which are linearized using the linear regression of $\hat{\theta}(n)$ with regard to n and resampling. Linear interpolation is used for resampling signal $d_{lin}(n)$ and nearest neighbour interpolation is used for resampling pulses $u_{lin}(n)$. The lag index l is the time index with respect to the centre of the pulse shape and goes from $-\hat{N}_0/2 + 1$ to $\hat{N}_0/2 - 1$.

The quasi-unit pulse train is shifted in time by Δ_{Φ} , such that the maximum of the pulse shape estimate $\hat{r}(l)$ is centred, i.e., $\text{argmax}[\hat{r}(l)] = 0$. As a result, the pulses of $u(n)$ align with maxima of the recorded pulse train $\hat{d}(n)$.

The instantaneous phase estimate $\hat{\theta}(n) = \pi \cdot \sum_{\mu} [2 \cdot \mu + 1]$ at pulse locations, i.e., at $n = \mu \cdot \hat{N}_0 + \hat{j}(\mu) + \Delta_{\Phi}$, and is obtained via cubic spline interpolation in between. The amplitude modulation function $\hat{A}(n) = \hat{s}(\mu)$ at the pulse locations, and is obtained via shape-preserving piecewise cubic spline interpolation in between.

The timing perturbation $\hat{j}(\mu)$ and the amplitude perturbation $\hat{s}(\mu)$ are obtained as $[\hat{j}(\mu), \hat{s}(\mu)] = \text{argmin}_{j,s} \{E[j(\mu), s(\mu)]\}$, with $E = 20 \cdot \log_{10} \left\{ \sqrt{e^2(n)} / \sqrt{d^2(n)} \right\}$, where the error of the modulating model $e(n) = d(n) - \hat{d}_m(n)$, and the model waveform $\hat{d}_m(n) = \sum \hat{d}(n)$ is the sum of the selected candidate waveforms.

The times and heights of the cycle maxima are estimated using the interior-point algorithm [5,6]. In monophonic frames, i.e., frames with one f_o , the times and heights are jointly optimized for each individual pulse, as proposed in [7]. In diplophonic frames, i.e.,

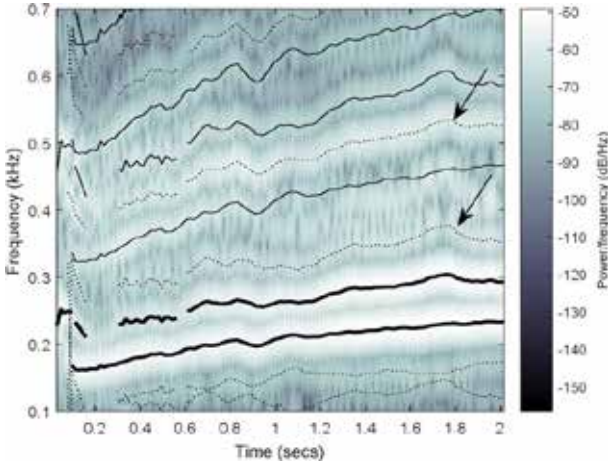


Figure 2: Magnitude spectrogram of a diplophonic voice sample $d(n)$, including f_0 s (thick solid lines), partials (thin solid lines), and combination tones (dotted lines). Arrows point at examples of frequencies at which increased power spectral densities are observed in between fundamentals and partials.

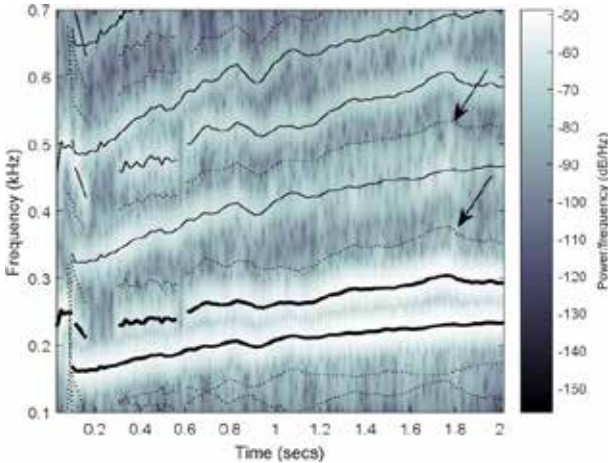


Figure 3: Magnitude spectrogram of a waveform $\hat{d}_s(n)$ output by the quasi-static model. A lack of power spectral density is observed in between fundamentals and partials, e.g., at frequencies indicated by arrows.

frames with two f_0 s, each pulse of the faster pulse train is optimized together with the closest pulse of the slower pulse train. Thus, in monophonic frames two parameters are jointly estimated, i.e., $\hat{j}(\mu)$ and $\hat{s}(\mu)$ of the μ^{th} pulse, and in diplophonic frames four parameters are jointly estimated, i.e., $\hat{j}(\mu)$ and $\hat{s}(\mu)$ of the μ^{th} pulse of the faster pulse train together with $\hat{j}(\mu)$ and $\hat{s}(\mu)$ of the closest pulse of the slower pulse train.

The quasi-static model waveform is denoted as $\hat{d}_s(n) = \sum_{\gamma} \hat{d}(n)$, with $\hat{d}(n)$ obtained with

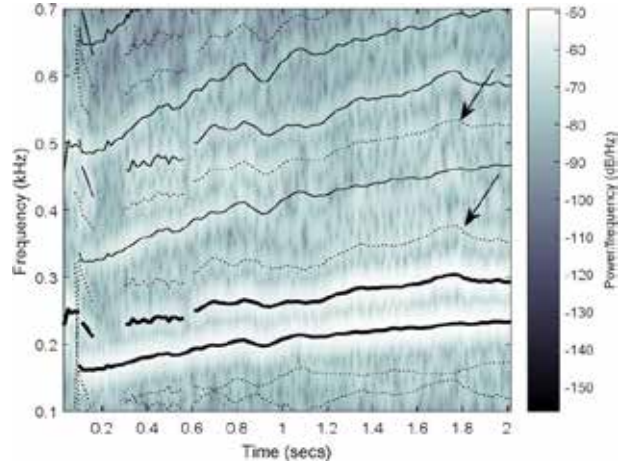


Figure 4: Magnitude spectrogram of a waveform $\hat{d}_m(n)$ output by the modulating model. The power spectral density in between the fundamentals and partials is increased, e.g., at frequencies indicated by arrows (cf. Fig. 3).

$\hat{j}(\mu) = 0 \forall \mu$, and $\hat{s}(\mu) = 1 \forall \mu$. The waveform outputted by the modulating model is denoted as $\hat{d}_m(n) = \sum_{\gamma} \hat{d}(n)$, with $\hat{d}(n)$ obtained with $-\hat{N}_0/4 < \hat{j}(\mu) < \hat{N}_0/4$, and $0.1 < \hat{s}(\mu) < 3$.

III. RESULTS AND DISCUSSION

Fig. 1 shows waveforms involved in the modelling of an example of a monophonic GAW. Five pulses are shown in the top plot. The times and magnitudes of the cycle maxima of the recorded waveform are perturbed ($d(n)$, solid line). The dotted line reports the quasi-static model waveform $\hat{d}_s(n)$. Its pulses are unperturbed mainly. The circles report times and heights of the pulses of an amplitude normalized version of $u(n)$ used for obtaining $\hat{d}_s(n)$. The largest errors are observed for the first, second, and last pulse. The dashed line reports the waveform $\hat{d}_m(n)$ output by the modulating model. Its pulses are perturbed in a way that aims at mimicking the perturbation observed in $d(n)$. The crosses report times and heights of the pulses of an amplitude normalized version of $u(n)$ used for obtaining $\hat{d}_m(n)$. The bottom plot reports the instantaneous frequencies of the quasi-static and the modulating model (dotted and dashed lines respectively). The delay of the last pulse causes a decrease of the instantaneous frequency of the waveform output by the modulating model between 1000 and 1400 samples approximately. The weak fluctuations of the instantaneous frequency of the waveform output by the quasi-static model are caused by the upsampling of \hat{f}_0 involved in the creation of $u(n)$. The model error level E_m of the modulating model is smaller than the error level E_s of the quasi-static model.

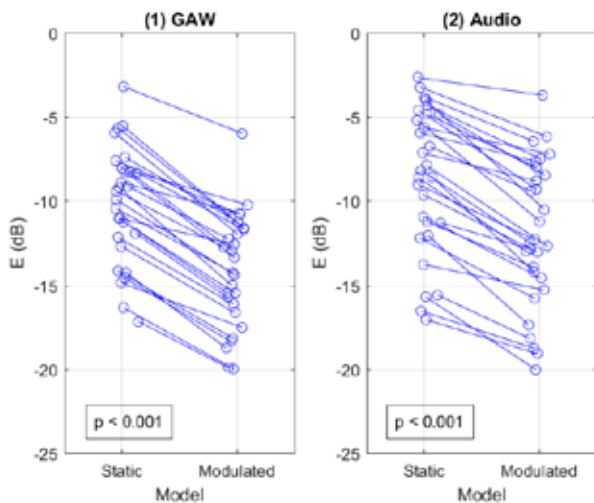


Figure 5: Model errors reported for (1) the glottal area waveforms (GAW), and (2) the audio waveforms.

Figs. 2 to 4 show magnitude spectrograms of a recorded waveform, its quasi-static model waveform, and the waveform outputted by the modulating model, including f_o estimates (thick solid lines), partials (thin solid lines), and combination tones (dotted lines). The power spectral density is increased at many of these frequencies. In particular, an increase of power spectral density is observed in the recorded waveform and the waveform outputted by the modulating model between fundamentals and partials, e.g., at frequencies indicated by arrows, whereas a lack of power spectral density is observed in the waveform outputted by the quasi-static model.

Fig. 5 reports model errors averaged for each voice sample, including lines for paired comparisons of the quasi-static and the modulating model. The error levels for the audio signals are higher than the error levels for the GAWs. Error reduction owing to the modelling of modulations is smaller for audio signals than for GAWs.

IV. CONCLUSION

A waveform model for diplophonic voices is proposed. The model is able to modulate times and magnitudes of cycle maxima for each glottal oscillator. As a benchmark, the model is used with its pulse-to-pulse modulator being bypassed, i.e., as a quasi-static model. This benchmark is outperformed by the proposed model in terms of modelling accuracy, which indicates that estimation of modulation noise is important for the modelling of diplophonic voices. This constitutes to our understanding of irregularities observed in pathological voices. The difference between the error of the quasi-static and the error of the modulating model is a possible feature of the amount of

modulation noise present in a signal. The model error of the modulating model is a possible feature of the amount of additive noise present in a signal.

Limitations of the approach and suggestions for future work include the following. First, the estimation of the modulation noise is computationally expensive. Second, no distinction is made between cyclic and acyclic modulation. Third, the GAW and the audio signal may be used jointly to estimate the quasi-unit pulse train, which is a vector of vocal excitation shared by the two. Finally, a compensation of cross-talk between the estimates $\hat{d}(n)$ of the individual simultaneous f_o s may be added based on the cross-correlation of the estimates $\hat{d}(n)$.

ACKNOWLEDGEMENT

This work was supported by the Austrian Science Fund (FWF): KLI 722-B30.

REFERENCES

- [1] P. Aichinger, M. Hagmüller, B. Schneider-Stickler, J. Schoentgen, F. Pernkopf, Tracking of Multiple Fundamental Frequencies in Diplophonic Voices, *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 26 (2018) 330–341. doi:10.1109/TASLP.2017.2761233.
- [2] M. Tigges, P. Mergell, H. Herzel, T. Wittenberg, U. Eysholdt, Observation and modelling of glottal biphonation, *Acta Acust. United with Acust.* 83 (1997) 707–714.
- [3] J. Lucero, J. Schoentgen, J. Haas, P. Luizard, X. Pelorson, Self-entrainment of the right and left vocal fold oscillators, *J. Acoust. Soc. Am.* 137 (2015) 2036–2046. doi:10.1121/1.4916601.
- [4] P. Aichinger, I. Roesner, M. Leonhard, D. Denk-Linnert, W. Bigenzahn, B. Schneider-Stickler, A database of laryngeal high-speed videos with simultaneous high-quality audio recordings of pathological and non-pathological voices, in: *Proc. Int. Conf. Lang. Resour. Eval.*, 2016: pp. 767–770. doi:10.13140/RG.2.2.15467.34088.
- [5] R.H. Byrd, J.C. Gilbert, J. Nocedal, A trust region method based on interior point techniques for nonlinear programming, *Math. Program. Ser. B.* 89 (2000) 149–185. doi:10.1007/s101070000189.
- [6] R.A. Waltz, J.L. Morales, J. Nocedal, D. Orban, An interior algorithm for nonlinear optimization that combines line search and trust region steps, *Math. Program.* 107 (2006) 391–408. doi:10.1007/s10107-004-0560-5.
- [7] P. Aichinger, F. Pernkopf, J. Schoentgen, Detection of extra pulses in synthesized glottal area waveforms of dysphonic voices, *Biomed. Signal Proces.* 50 (2019) 158–167. doi:10.1016/j.bspc.2019.01.007.

TRACKING OF MULTIPLE FUNDAMENTAL FREQUENCIES IN STANDARD TEXT READINGS OF DIPLOPHONIC SPEAKERS

P. Aichinger

Medical University of Vienna, Department of Otorhinolaryngology,
Division of Phoniatics-Logopedics, Vienna, Austria
philipp.aichinger@meduniwien.ac.at

Abstract: A need to establish objective criteria for improving clinical voice assessment is observed. Diplophonia is a symptom of pathological voice that is characterized by the auditory perception of two simultaneous pitches. Two simultaneous f_o s are extracted, which correspond to the perceived pitches. Audio waveform modelling (AWM) based f_o -tracking is trained and tested with reference to tracks obtained via manual annotation of magnitude spectrograms. The corpus contains audio recordings of the German standard text read by a total of forty subjects who were reported by medical doctors specialized in voice disorders to sound diplophonic. AWM- f_o -tracking includes (i) candidate f_o -tracking using spectral peak picking and repetitive execution of the Viterbi algorithm, (ii) Fourier synthesis of candidate waveforms for comparison with the recordings, and (iii) a candidate selection heuristic. With previously available parameter settings, a mean error rate of 24.83 % is observed. By training, it improves to 21.37 %, which is a top-notch result with regard to the properties of this class of signals.

Keywords: Voice quality assessment, diplophonia, multiple fundamental frequencies, audio waveform modelling

I. INTRODUCTION

Diplophonia is auditorily characterized by the perception of two simultaneous pitches in the voice sound, which may be caused by a voice disorder [1,2]. Objective characterization of diplophonic voice is necessary for several reasons. First, the use of an objective means for characterizing diplophonia would improve (global) comparability of patient data between clinical/research centers. Second, it would improve comparability of patient data obtained before and after clinical treatment, e.g., logopedic therapy, or surgery. In particular, the quality of indication, selection, evaluation, and optimization of treatment may improve when objective means of voice quality characterization are used. Third, the presented approach may provide a means of reassurance to clinicians, if the presence or absence of diplophonia is auditorily doubtful in a

patient. Finally, a means of clinician training may be established via displaying objective analysis results for the anchoring of subjective ratings.

In this paper, objective characterization of diplophonic voice is attempted via extracting up to two simultaneous f_o s from audio recordings of forty subjects, who were clinically rated in the past as diplophonic. This approach is chosen on account of the strong relationship between f_o s and perceived pitches. The proposed method for f_o -tracking is an advancement of a previously published method [3]. The original contribution of this paper is a structural update in the f_o -candidate selection, and the training and evaluation with reference to f_o -tracks manually mined from audio spectrograms of text readings.

II. METHODS

A. Speech recordings

Audio recordings of forty subjects reading the German standard text "Der Nordwind und die Sonne" are analyzed. The subjects were recruited in the past from the outpatients of the Medical University of Vienna, Department of Otorhinolaryngology, Division of Phoniatics-Logopedics [4]. All subjects suffered from voice disorders and were reported by medical doctors specialized in voice disorders to sound diplophonic during anamnesis, i.e., two simultaneous pitches were auditorily perceived in the subjects' voices. The length of the audio recordings varies between 38 and 119 seconds.

B. Reference f_o -tracks

Signals from electroglottography (EGG) are frequently used as a reference in f_o -tracking problems [5]. However, this approach is not suitable for the analysis of diplophonic voices, because sufficient glottal closure is required for EGG signals to contain all needed information, which is mostly not observed in diplophonia. For diplophonic sustained vowels, laryngeal high-speed videoendoscopy was used as a reference in the past [3]. Such videos are not available for read speech, because the epiglottis

frequently hides the vocal folds from endoscopic view, which motivates the use of a spectrogram annotation based approach.

Log-magnitude spectrograms of the speech recordings are annotated manually with respect to f_o -tracks using a custom GUI. Fig. 1 shows an example of an annotation including a log-magnitude spectrogram, candidate f_o s, and selected reference f_o -tracks, as well as their partials and combination tones. In addition, audio waveforms are available to the annotator for auditory judgement via playback, and displayed for visual judgement with regard to characteristic diplophonic beating observed in diplophonic waveforms due to the superposition of two cyclic signals with different f_o s.

C. Error measures

Agreement of reference f_o -tracks with estimated AWM- f_o -tracks is evaluated by means of error rates E_{Total} (%), and relative frequency errors E_{Fine} (%) [6]. E_{Total} denotes the proportion of time frames in which either a wrong number of f_o s is estimated, or the relative frequency error E_{Fine} is larger than 20 %. In frames with two f_o s, the relative frequency errors are added.

A modified error measure $E'_{Total} = E_{Total} + N_c/T$ is used for the training of candidate f_o -tracking. N_c is the number of candidate AWM- f_o -tracks, and T is the length of the audio recording in seconds. The modification of E_{Total} by N_c/T penalizes interruption of candidate AWM- f_o -tracks. In the previous publication, N_c was not divided by T , because all voice sample were equally long there (2.048 seconds) [3].

D. AWM- f_o -tracking

Candidate f_o -tracking: First, the audio signal is preprocessed and a magnitude spectrogram is obtained. The audio recording is normalized to a maximum of +/- 0.99, resampled at 48 kHz, and FIR filtered with coefficients [1,-1] for pre-emphasis of high frequencies. The magnitude spectrogram is obtained at frequencies from 63 to 600 Hz using 32 ms Chebyshev windows having a sidelobe attenuation of 50 dB with a hop size of 16 ms. The resolution is 0.5 Hz. The Viterbi algorithm is executed five times to extract multiple candidates as described in [3].

Three adjustable HMM parameters α , σ_A , and σ_B are used in candidate f_o -tracking. α is the probability of a candidate f_o -track remaining inactive from one frame to next, or conversely, α controls the probability of birth events. σ_A is the deviation of Gaussians modelling transitions between active states, i.e., σ_A controls the mobility of f_o -tracks. The higher σ_A , the further f_o s tend to travel from one frame to the next.

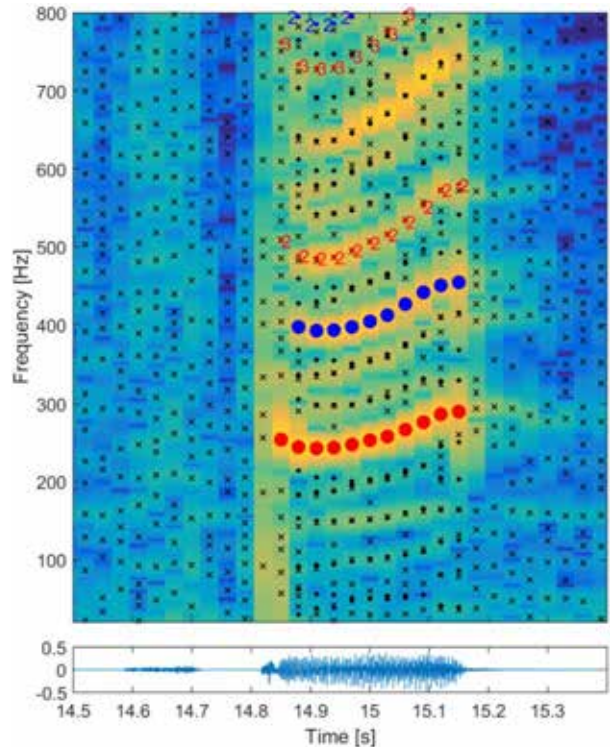


Figure 1: Example of a reference f_o -track annotation showing a log-magnitude spectrogram, candidate f_o s (crosses), and the selected reference f_o -tracks (large filled circles), as well as their partials (numbered dots), and combination tones (unnumbered dots).

Finally, σ_B is a compressive parameter used in the conversion of peak heights to the probabilistic domain. The smaller σ_B , the more suppressed get small spectral peaks.

Candidate waveform synthesis: For each candidate f_o -track, a candidate waveform is synthesized as described in [3]. Here, a frame length of 32 ms and a hop size of 16 ms are used. First, for each candidate, a unit pulse train with f_o equal to the candidate's frequency is generated. Second, an average pulse shape is obtained by cross-correlating the unit pulse train with the audio signal and normalizing. The candidate waveforms are obtained as

$$d_\gamma(n) = \sum_{p=1}^{10} \begin{bmatrix} a_p^\gamma \cdot \cos(\omega_\gamma \cdot p \cdot n) + \\ b_p^\gamma \cdot \sin(\omega_\gamma \cdot p \cdot n) \end{bmatrix}, \quad (1)$$

where the Fourier coefficients a_p^γ and b_p^γ are obtained via Fourier transformation of the time domain pulse shape, γ is the candidate index, p is the partial index, $\omega_\gamma = 2 \cdot \pi \cdot f_o^\gamma$, and n is the discrete time index.

Candidate selection: In the past, majority voting based candidate preselection and brute force final selection were used for short recordings. This approach

is not feasible for long recordings due to large computation times [3]. A faster candidate selection proposed in [7] is used here and generalized for diplophonic voices. In particular, candidate combinations are considered, which include up to two temporally overlapping candidate f_o -tracks.

First, audio recordings are split into segments at frames without candidates. Second, for each segment separately, candidates are selected as described hereafter.

The candidate waveforms d_γ are sorted with respect to signal energy $\sum d_\gamma^2$. A binary candidate selection vector $\kappa_\gamma \in \{0,1\}$ is initialized as a zero vector, the length of which equals the number of candidates. An error waveform $e_\kappa(n) = d(n) - d_\kappa(n)$, where $d(n)$ is the audio recording, and the sum of the selected candidate waveforms $d_\kappa(n) = \sum[\kappa_\gamma \cdot d_\gamma(n)]$. The error measure

$$\Delta(\kappa) = 20 \cdot \log_{10} \left(\frac{\sqrt{\sum d^2(n)}}{\sqrt{\sum e_\kappa^2(n)}} + \lambda \cdot \frac{N_{v,\kappa}}{N_f} \right), \quad (2)$$

where λ (dB) is a regularization parameter that penalizes the addition of candidate waveforms, $N_{v,\kappa}$ is the number of voiced frames summed over all candidates involved in a particular candidate combination κ , and N_f is the length of the segment in frames.

The optimal candidate selection vector minimizes the error measure, i.e., $\kappa_{opt} = \text{argmin}[\Delta(\kappa)]$. For each γ individually, the state of the γ^{th} element of κ is switched. If more than two candidates overlap temporally after a switch, or if Δ is not decreased by the switch, the switch is reverted. Switching goes from the first to the last γ , and is repeated in a loop until no decrease of Δ is observed anymore. The final estimates of f_o are the candidates selected by the optimal candidate selection vector κ_{opt} .

Parameter optimization: Four parameters are optimized. Regarding candidate f_o -tracking, the HMM parameters α , σ_A , and σ_B are optimized in order to minimize the mean of the modified error measure E'_{Total} . Regarding candidate selection, the regularization parameter λ is optimized in order to minimize the mean of E_{Total} .

Three-fold cross-validation is applied, taking into account the frequency of occurrence of diplophonia in individual subjects. In particular, the corpus is randomly split into three subsets, approximately equal in size, and balanced for the frequency of occurrence of diplophonia. In each of the three cross-validation iterations ($k = 1, 2$, and 3), two subsets serve as training data, and the remaining one serves as testing data.

For the training of the HMM parameters, golden section search is used with parabolic interpolation. The parameters are initialized with values that were used in the past for the analysis of sustained vowels [3], i.e., $\alpha = 0.8235$, $\sigma_A = 3.748$, and $\sigma_B = 2.1765$. They are optimized in the order α , σ_A , and σ_B , using bounds of $[.1, .9]$, $[2, 6]$, and $[1, 3]$. The termination tolerances for the optimization of the individual parameters are set to .1, 1, and 1. The three parameters are optimized in a loop, which is stopped as soon as improvement stops.

For the optimization of λ , E_{Total} is obtained for λ going from 0 to 8 dB, using a stepsize of 0.2 dB in the interval from 3 to 6 dB, and a stepsize of 1 dB elsewhere. The λ with the lowest mean E_{Total} is selected as the optimum. Optimal values of α , σ_A , and σ_B are used when training λ .

III. RESULTS AND DISCUSSION

The optimal parameters are favorably similar for each training corpus. Optimal α is found to be 0.5944 for $k = 1$ and 2 , and 0.5611 for $k = 3$. Optimal σ_A and σ_B are for all k 3.5279 Hz and 2.1765 Hz respectively.

The following differences between parameters optimized for sustained vowels and parameters optimized for read texts are observed [3]. First, optimal α is lower for f_o -tracking in standard text readings than for sustained vowels (0.8235 for "fast" candidate tracking, and 0.8389 for "accurate" candidate tracking). This decrease reflects a larger frame based probability of voice onset in standard text readings than in sustained vowels. Second, σ_A is slightly lower than for sustained vowels (3.74 Hz for "fast" candidate tracking, and 4.255 Hz for "accurate" candidate tracking). This decrease is counter intuitive because it is suggestive of a smaller mobility of f_o -tracks in standard text readings than in sustained vowels. Third, λ for standard text readings is about 4-5 times larger than it is for sustained vowels (1.011 dB). This increase is due to the structural change of the candidate selection heuristic.

Fig. 2 shows boxplots of the performance measure E_{Total} obtained for the training of the penalty parameter λ (dB). The subplots (1), (2), and (3) report results obtained for the three training corpora. Bathtub-like shapes are observed, i.e., E_{Total} is larger at extreme values of λ than elsewhere, reflecting a favorably wide operating interval going from approximately 2 to 6 dB. The minimal means of E_{Total} are observed at 4.0, 4.8, and 4.2 dB.

Fig. 3 reports E_{Total} and E_{Fine} for parameters α , σ_A , σ_B , and λ . In 'Vowel settings' previously parameters available are used [3]. In 'Speech settings' parameters optimized for text readings are used. Performance

improvement owing to the parameter optimization is observed.

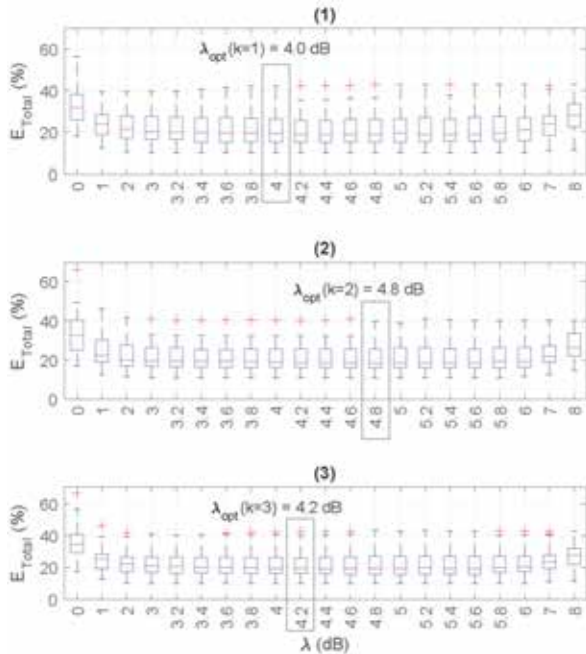


Figure 2: Error measure E_{Total} with respect to penalty parameter λ (dB). Training corpora (1), (2), and (3).

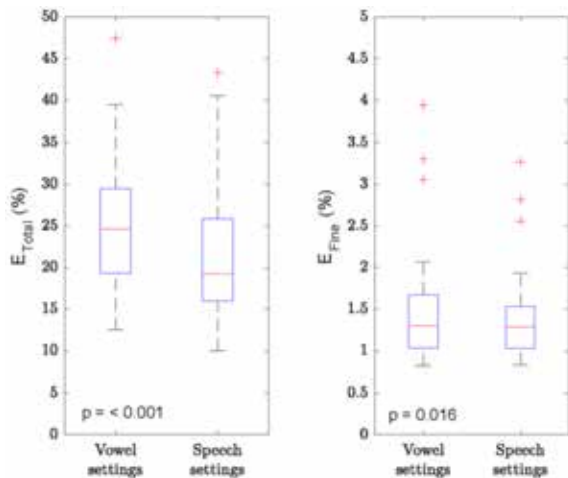


Figure 3: Error measures E_{Total} and E_{Fine} with respect to the settings of HMM parameters and the regularization parameter λ . For 'Vowel settings', parameters reported in [3] are used, for 'Speech settings', parameters optimized here for analyzing text readings are used.

IV. CONCLUSION

Training and testing of the tracking of f_o s in audio recordings of German standard text readings is reported. AWM- f_o -tracking is evaluated with reference

to f_o -tracks obtained via manual annotation of log-magnitude audio spectrograms. With previously available parameters, a mean error rate E_{Total} of 24.83 % and a mean relative frequency error E_{Fine} of 1.47 % are observed. The parameter training improves these error measures to means of 21.37 % and 1.39 % respectively.

The large operating interval of the regularization parameter λ , as well as the similarity of the HMM parameters α , σ_A , and σ_B across cross-validation subsets are favorable. The high levels of noise and disturbances challenge the f_o -tracking, the performance of which is quantitatively lower than in less difficult f_o -tracking tasks. Nevertheless, top-notch performance is observed with regard to the properties of the class of signals used here.

ACKNOWLEDGEMENT

This work was supported by the Austrian Science Fund (FWF): KLI 722-B30.

REFERENCES

- [1] P.H. Dejonckere, J. Lebacqz, An analysis of the diplophonia phenomenon, *Speech Commun.* 2 (1983) 47–56.
- [2] P. Aichinger, Diplophonic Voice - Definitions, models, and detection, PhD dissertation, Graz University of Technology, Austria, 2015. doi:10.13140/RG.2.1.4273.5763.
- [3] P. Aichinger, M. Hagmuller, B. Schneider-Stickler, J. Schoentgen, F. Pernkopf, Tracking of Multiple Fundamental Frequencies in Diplophonic Voices, *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 26 (2018) 330–341. doi:10.1109/TASLP.2017.2761233.
- [4] P. Aichinger, I. Roesner, M. Leonhard, D. Denk-Linnert, W. Bigenzahn, B. Schneider-Stickler, A database of laryngeal high-speed videos with simultaneous high-quality audio recordings of pathological and non-pathological voices, in: *Proc. Int. Conf. Lang. Resour. Eval.*, 2016: pp. 767–770. doi:10.13140/RG.2.2.15467.34088.
- [5] A. Tsanas, M. Zaňartu, M.A. Little, C. Fox, L.O. Ramig, G.D. Clifford, Robust fundamental frequency estimation in sustained vowels: detailed algorithmic comparisons and information fusion with adaptive Kalman filtering., *J. Acoust. Soc. Am.* 135 (2014) 2885–2901. doi:10.1121/1.4870484.
- [6] M. Wu, D. Wang, G. Brown, A multipitch tracking algorithm for noisy speech, *IEEE Trans. Speech Audio Process.* 11 (2003) 229–241. doi:10.1109/TSA.2003.811539.
- [7] P. Aichinger, F. Pernkopf, J. Schoentgen, Detection of extra pulses in synthesized glottal area waveforms of dysphonic voices, *Biomed. Signal Process.* 50 (2019) 158–167. doi:10.1016/j.bspc.2019.01.007.

AERODYNAMICS OF GLOTTAL VIBRATION ONSET

P. H. DeJonckere¹, Lebacqz²

¹Federal Agency for Occupational Risks, Brussels, Belgium

²Neurosciences Institute, University of Louvain, Brussels, Belgium
ph.dejonckere@outlook.com ; jean.lebacqz@uclouvain.be

Abstract: A physiological voice onset starts from an immobile narrow glottal slit crossed by a continuous airflow, and a few oscillations (a single one in some cases) precede the first glottal closure. Combined measurements of flow, area and pressure provide a detailed qualitative and quantitative analysis of the intraglottal mechanical events at the precise moment of starting oscillation in a physiological onset. Our *in vivo* measurements of airflow and glottal area show that the very first oscillation occurs exactly at the time when turbulence appears at the level of the glottal narrowing, i.e when the Reynolds number reaches its critical value. Turbulence can act here as an aspecific flick, triggering the oscillator, the frequency of oscillation being determined by its mechanical properties. Furthermore, the first noticeable glottal oscillations are sinusoidal: the VFs are neither steeply sucked together by a negative Bernoulli pressure, nor burst apart by the lung pressure. Our measurements show that, at the critical time, the rising positive lung pressure is balanced by the rising negative Bernoulli pressure generated by the transglottal flow.

Keywords: Vocal onset, intraglottal pressure, Bernoulli, turbulence, Reynolds number.

I. INTRODUCTION

The onset of vocal fold (VF) vibration is a transient phenomenon, in which the forces in play progressively adjust until a steady state is reached [1]. The acting forces are lung pressure, intraglottal pressure, myoelastic tension of the VF oscillator and inertance of the supraglottal vocal tract [2,3]. In normal subjects, the most frequently observed type of voice onset in spontaneous speech is the soft onset, and it may be considered as the 'physiological' onset. It starts from an immobile narrow glottal chink crossed by a continuous airflow, with a few oscillations (possibly a single one) preceding the first glottal closure (Fig. 1) The present study, using accurate flow, area and pressure measurements *in vivo* investigates the

mechanism that accounts for the initiation of the oscillation i.e. the trigger of the transition from an immobile, spindle-shaped glottis (Fig. 2), crossed by a continuous airflow, to a weakly damped oscillating system [1].

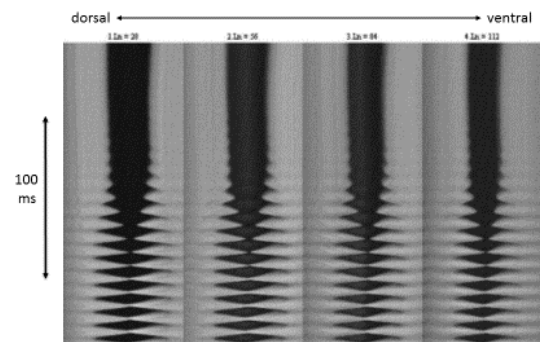


Fig. 1 : Videokymogram at four equidistant levels of the vibrating glottis, obtained by highspeed video. Soft, somewhat breathy onset. /a:/; healthy male subject (~125 Hz; 65 dB at 10 cm). Vertical axis : time. Progress from top to bottom.

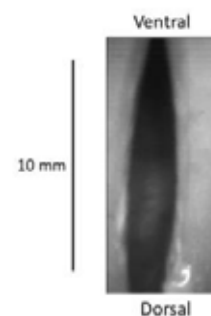


Fig. 2: Spindle-shaped glottis immediately before a soft onset (snapshot from high speed video).

A key-element is the Reynolds number [4]. Small values of this dimensionless number indicate that

viscous forces are predominant, and that the flow is laminar (sheet-like), characterized by a smooth, constant fluid motion. Large values indicate that viscous forces are small and that the flow is essentially inviscid, dominated by inertial forces, which tend to produce chaotic eddies, vortices and other instabilities. The Reynolds number is used to predict the onset of turbulence, and the transition from laminar to turbulent flow.

In the case of a circular pipe, the Reynolds number is defined as:

$$Re = \text{velocity (m/s)} \times \text{inside diameter of the pipe (m)} / \text{kinematic viscosity of the fluid (m}^2\text{/s)}. \quad (1)$$

As a general rule, for values of $Re < 2000$, the flow is stationary, thus laminar. For values between 2000 and 4000, the system is in transition, characterized by growing coherent bidimensional vortices, joining of vortices, separation of vortices and turbulent state in 3 dimensions. For values of $Re > 4000$ the fluid is non-stationary and tri-dimensional [5]. For straight pipes, the transition is sudden: once triggered, the transition to turbulence can be described as 'catastrophic', i.e. it is abrupt and distinct concerning flow properties such as the pressure drop.

Turbulence of the fluid has important consequences for the pipe. It is well known that pipe vibrations result from the inherent spatially and temporally varying pressure at the pipe wall.

Hence, if an exact concomitance of the occurrence of airflow turbulence, revealed by a critical value of Reynolds number, and the onset of glottal oscillations can be shown, it strongly pleads in favour of a causal relationship between them: the sudden occurrence of flow instability can be supposed to act as the flick that triggers the oscillator.

During a physiological voice onset, the lung pressure progressively increases. The lung pressure tends to burst apart the VF from their spindle-shaped pre-phonatory configuration (Fig. 2). Simultaneously, the transglottal flow generates a negative pressure at the level of the glottal narrowing that can be calculated using Bernoulli's equation. Our experimental setup allows verifying quantitatively whether the two forces balance each other.

II. MATERIAL & METHODS

Glottal area (light flow)

The glottal area was derived from a photometric record obtained by transilluminating the trachea. The light flux was detected by a photovoltaic transducer positioned as dorsally as possible in the pharynx (photoglottography) [2]. 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's lead - through the hermetically sealed hole

normally intended for the handpiece of a Rothenberg mask [2]. The current produced by the photodiode was preamplified by a current-to-voltage converter with a linear response up to 2 kHz. The calibration procedure has been described previously [2,3]. The measured glottal area at maximal glottal opening can be related to the peak of the photodiode current. Since the precise position of the photodiode cannot be reproduced from record to record, in each record, the light signal was expressed as a fraction of the maximal amplitude at full opening.

For calibration of the photoglottographic signal a rigid 90° Wolf laryngeal telescope and an ATMOS Strobo 21 LED stroboscope (Atmos Medizin Technik, Lenzkirch, Germany) were used to obtain still images of the entire glottis at the time of maximal opening, as early as possible after onset. The telescope has a magnifying facility, with narrow depth of field and critical sharpness adjustment; scale paper was filmed at the same focal length, critical care being given to maximal sharpness. Stroboscopic images yielded a value of maximal area of 30.6 mm².

Airflow

The glottal flow waveform (flowglottogram) was recorded using a Rothenberg mask and an inverse filtering system (model MSIF2, Glottal Enterprises, Syracuse, NY). The calibration procedure and the correction for time delay were described earlier [2,3].

Pressure

The intra-oral pressure was measured by means of a Millar Mikro-Tip catheter (Model SPC-751, Millar Instruments, Inc. Houston, USA). The short flow interruption method was used for indirect measures of subglottic pressure, particularly the phonation threshold pressure (PTP).

The EGG-signal, used as reference for monitoring the contact surface changes, was detected using a portable electroglottograph (Laryngograph Ltd, London, UK) Model EG90. Sounds were detected by a Sennheiser MD 421 U microphone.

All signals were recorded by means of a 4-channels Pico Scope 3403D module (Pico Technology Ltd, St Neots, England, UK) and stored in a computer.

Vocal emissions

Out of a corpus of about 100 recordings of short vocal emissions on /ə/ into the Rothenberg mask with the light sensor in situ, 72 soft / slightly breathy onsets were selected, all of them at frequencies of 105 – 130 Hz, corresponding to the average speaking frequency of the subject, and at comfortable loudness, i.e. 62 – 73 dB_A at 10 cm of the lips. Criteria for selection were (1) full display of all traces at the time of onset and until occurrence of the first closed plateau on

photoglottographic / flowglottographic traces, and (2) limited drift of the flow and light traces in order to allow valid measurements of airflow and area. For PTP-measurements, 15 records (repetitions of the syllable /pi/) were selected out of a corpus of 47, at a F_0 corresponding to the average speaking frequency of the subject, around 115 Hz. Criteria for selection were again full display and horizontal stability of both traces.

The subject was a healthy trained vocalist, experienced in controlling voicing parameters [2,3].

Calculation methods

Reynolds number at the time of starting oscillation

The shape of the glottis is not cylindrical; thus, an expression must be found to calculate an equivalent diameter of the glottis to be introduced in Eq. (1) [6]. After calibration, the ventro-dorsal length of the glottis was found to be 13 mm and the maximal width was 3 mm. The contour of the glottal image could be well fitted with an ellipse, the major and the minor axes of which were respectively the ventro-dorsal length and the maximal width of the glottis picture. The difference between the calculated area of the ellipse and the measured area of the glottis was less than 1 %. The equivalent diameter of an ellipse is given by

$$ed = 1.55 A^{0.625} / P^{0.25} \quad (2)$$

where A and P are the area and the perimeter of the ellipse respectively.

Intraglottal pressure at the time of starting oscillation

The intraglottal pressure at the time of starting oscillation is the result of the combination of the positive lung pressure and the negative pressure generated by the Bernoulli-effect of the air flow in the glottal narrowing. The PTP represents the minimum subglottal pressure needed to initiate oscillation of the vocal folds. Hence, it may be assumed that the lung pressure at the time when oscillation starts is very close to the PTP, which can be accurately estimated by measuring the intraoral pressure. The negative pressure generated by the air flow is calculated by Bernoulli's equation:

$$P = -k \cdot \rho v^2 / 2 \quad (3)$$

where ρ is air density, v is air speed and k is a pressure loss coefficient for glottal entry and viscous drag. The value of k was set at 1.37 [1,2].

III. RESULTS & DISCUSSION

Occurrence of intraglottal airflow turbulence

Fig. 3 shows an example of a recording. Vertical arrows on the area trace indicate the level at which oscillation starts and the maximal amplitude (100%).

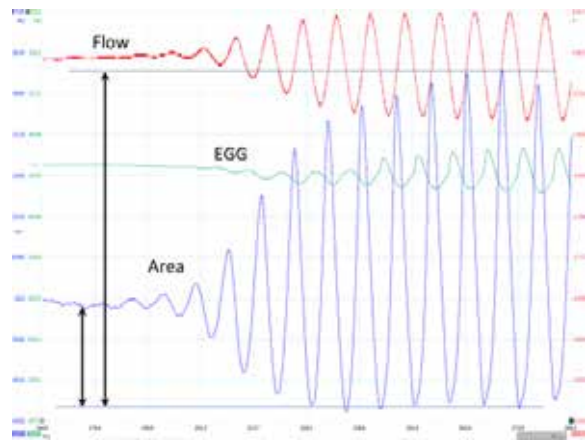


Fig. 3: From top to bottom: airflow (Rothenberg mask), electroglottogram and glottal area (photoglottogram) during a soft onset (area increases upwards). Total duration of the recording: 124 ms.

Fig. 4 shows the same recording. In this case, the flow level at which oscillation starts is indicated by the vertical arrow on the airflow trace with respect to the baseline reached when complete glottal closure is observed.

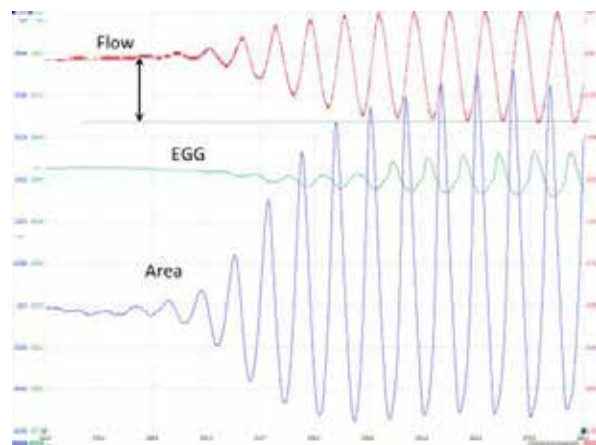


Fig. 4 : idem as Fig. 3.

The mean calculated Reynolds number for the 72 recordings is 3073 ± 479 (mean \pm SD). In Fig. 5, the airflow value (L/s) measured at the start of oscillation is plotted against the equivalent diameter (mm) calculated from the measured glottal area at the same time point.

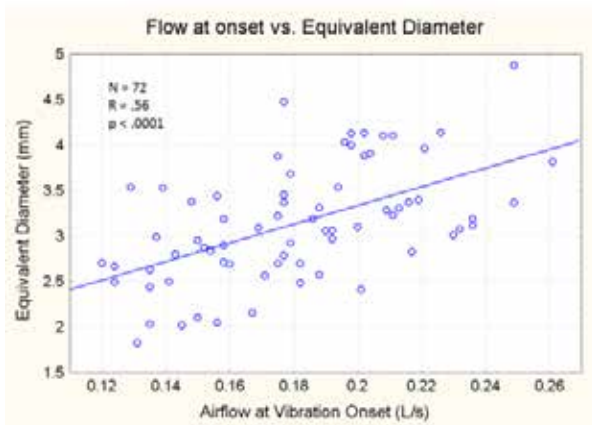


Fig. 5: Airflow (L/s) measured at the start of oscillation as a function of the equivalent diameter (mm) calculated from the measured glottal area at the same time point.

The positive correlation ($R = 0.56$; $p < .0001$) indicates that when the glottal area at which oscillation starts increases, the value of airflow also increases. Consequently, the velocity of air particles remains approximately constant; thus, the Reynolds number, which is directly proportional to the air particle velocity (1), is also approximately constant at the time of oscillation initiation.

The neutral (atmospheric) pressure at intraglottal level when the oscillation starts.

Fig. 6 shows an example of the method of estimation of subglottal lung pressure. The upper trace is the sound recorded by a microphone and the lower trace is the intraoral pressure. One single vocalization - out of a series of ten - is shown (repetitions of the syllable /pi/). At the moment of lip opening, the intraoral pressure suddenly drops to approximately the atmospheric pressure and increases as soon as the lips close again. A few oscillation cycles persist - even with a larger amplitude - when the lips are already closed.

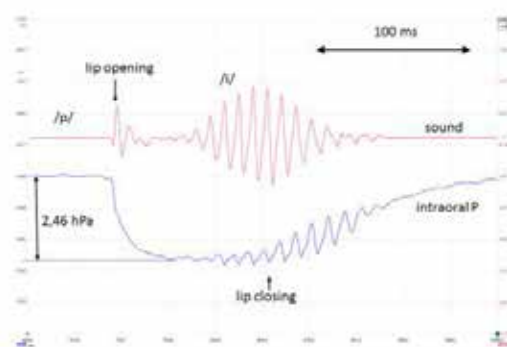


Fig. 6 : Evaluation of phonation threshold pressure by the short flow interruption method. Upper trace is

sound, lower trace is intraoral pressure. The PTP measured in this case is 2.46 hPa.

The average value of PTP for 15 soft onsets is 2.52 (SD 1.78) hPa. The estimated lung pressure at the start of the oscillation may thus reasonably be assumed to be close to the phonation threshold pressure. The average value of air velocity at the start of the oscillation is 16.74 (SD 1.81) m/s. The resulting pressure drop is:

$$\Delta P = - (16.74 \text{ m/s}^2 \cdot 1.14 \text{ kg/m}^3) / 2 = -219 \text{ Pa or } -2.19 \text{ hPa (SD 1.26)}$$

which is nearly equal to the PTP value.

V. CONCLUSION

In vivo measurements of airflow and glottal area show that the oscillation starts exactly at the time when turbulence occurs at the level of the glottal narrowing. The sudden occurrence of turbulence can be supposed to act as the flick that triggers the oscillator consisting in the ensemble of the VFs and the air of the vocal tract, and this oscillator is known to be weakly damped. Its mechanical properties determine the frequency of the oscillation. Further, at the critical time, the rising positive lung pressure is exactly balanced by the negative Bernoulli pressure generated by the increasing transglottal flow. This explains that, at the start of a soft onset, the oscillation axes of the VF remain approximately equidistant and are neither grossly drawn medially toward each other, nor pushed apart laterally. The oscillatory motion of each VF is symmetrical (adduction / abduction) with respect to a neutral position.

REFERENCES

- [1] J. Lebacqz and P.H. DeJonckere,, "The dynamics of vocal onset." *Biomedical Signal Processing and Control*. Vol. 49: pp. 528-539, 2019.
- [2] P.H. DeJonckere, J. Lebacqz and Titze I.R., "Dynamics of the driving force during the normal vocal fold vibration cycle." *J Voice*. Vol. 31, pp. 649–661, 2017.
- [3] P.H. DeJonckere and J. "In Vivo Quantification of the Intraglottal Pressure: Modal Phonation and Voice Onset." 2019 *J Voice* 2019 in press. <https://doi.org/10.1016/j.jvoice.2019.01.001>
- [4] <https://www.grc.nasa.gov/WWW/BGH/reynolds.html>
- [5] C. Gavilan Moreno, "Turbulence, vibrations, noise and fluid instabilities. Practical approach." In: *Computational Fluid Dynamics*, O.H. Hyoungh Woo Ed., DOI: 10.5772/71100, London: IntechOpen Ltd, 2010.
- [6] https://www.engineeringtoolbox.com/elementary-surfaces-d_1473.html

A GLOTTAL AREA WAVEFORM MODEL FOR MULTI-PULSED VOCAL FRY

V. Devaraj¹, P. Aichinger¹

¹ Department of Otorhinolaryngology, Division of Phoniatics-Logopedics, Medical University of Vienna, Austria
vinod.devaraj@meduniwien.ac.at, philipp.aichinger@meduniwien.ac.at

Abstract: The purpose of this study is to model vocal fry glottal area waveforms (GAWs) and to distinguish different types of vocal fry based on their vibratory patterns. Euphonic GAWs are also modelled. GAWs are obtained from laryngeal high-speed videos with 4000 frames per second. The GAWs are modelled using a sinusoidal carrier with an estimated frequency and phase, and a modulator which is obtained as a superposition of the lower and upper envelopes of the observed GAW. The lower and upper envelopes are estimated by interpolating the lower and upper extrema of the observed GAW respectively. This model is termed as target model. A control model waveform is obtained using constant lower and upper envelopes. The modelling error is determined to evaluate the fit for each of the models. The difference between the two models errors is found to be larger for vocal fry subjects than for euphonic subjects, which makes these groups distinguishable automatically. To further distinguish different vibratory patterns of vocal fry, the entropy of the modulator's power spectral density and the modulator-to-carrier frequency ratio are obtained.
Keywords: Vocal fry, euphonia, glottal area waveform, laryngeal high-speed videos

I. INTRODUCTION

Voice quality plays an important role in assessment of dysphonia [1]. Typical voice qualities are breathy, hoarse, diplophonic, creaky, and fluttering. Vocal fry is an additional voice quality. It does not have any standard definition and is commonly also referred to as glottal fry, pulse registers, creak, glottalization or creaky voice [2, 3]. Contradictorily, vocal fry was also denoted as one of the types of creaky voice. The other classifications of creaky voice are multi-pulsed voice, aperiodic voice, non-constricted creak and tense/pressed voice [4]. Vocal fry is characterized by low F_0 , shimmer, jitter, and damping of pulses [2, 3, 4, 5]. Vocal fry was characterised by larger shimmer and jitter than modal phonation [2, 5]. Sub-glottal air pressure and airflow were found to be less in vocal fry than in the modal registers [2].

Vocal fry has been identified in normal as well as in dysphonic speakers. The use of vocal fry was investigated in 104 first year graduate students [6]. 86% of the students had normal voice quality and the remaining 14% of these students were reported to have abnormal voices along with vocal fry. However, 16% of the students with normal voice quality also used vocal fry. In addition, vocal fry was reported to be one of the salient voice characteristics of male patients with contact granuloma [7]. Hence, vocal fry may be an indicator for the presence of a pathology, which makes the study of characteristics of vocal fry important.

Important physiological features of vocal fry include the number of openings and closings of the vocal folds in a single meta cycle, and the existence of a long closed phase. Vocal fry associated with single pulsing with a long closed phase was obtained through laryngeal high-speed videos [8]. Authors of [9] investigated the patterns of vocal fold vibration in vocal fry through ultra-slow motion pictures and found that the cords open and close twice per meta cycle in a rapid succession and afterwards remain in contact for a relatively longer period. Four voice types (modal, vocal fry, breathy and falsetto) were analysed by [10] through EGG waveforms. Double pulsed patterns for vocal fry without any long closed phase were observed there. Single and triple pulsed patterns with a long closed phase were also found [11]. More supporting results of multiple pulses in a single cycle were reported for EGG waveforms [3]. The mean F_0 was found to be lower for triple pulsed than the double pulsed vibratory patterns.

The GAW provides spatial information regarding vocal fold movement. In particular, it reflects the duration of the open and closed intervals of the glottal cycle, and also the duration of the opening and closing [12]. The GAWs are not affected by resonance effect as seen in audio signals, which make them suitable for in depth investigation of voice production. Therefore we model vocal fry GAWs with different vibration patterns by amplitude modulation of the sinusoidal carrier using a modulator, which is obtained as a superposition of lower and upper envelope of the observed GAW.

II. METHODS

Voice samples are obtained from an available database of laryngeal high-speed videos of dysphonic and euphonic subjects with simultaneous audio recordings [13]. The voice samples are annotated by three listeners with regard to the presence or absence of vocal fry based on the impulsiveness of the voice samples, where the pulses are separately audible. The corresponding GAWs are extracted from high-speed videos using a seeded region growing segmentation algorithm implemented in an available tool. Annotators agreed on seven voice samples to contain vocal fry. A group of eight euphonic GAWs is used for comparison.

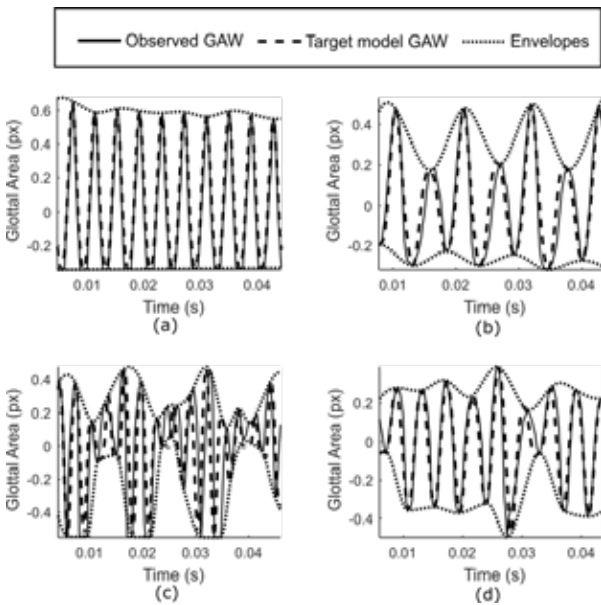


Figure 1: Observed and target models of (a) euphonic, (b) CAMS-2 (c) CAMS>2 and (d) AAMS GAWs.

Seven vocal fry GAWs are distinguished into two different types with regard to the meta cycles of the GAWs by visual inspection. Three out of seven vocal fry GAWs are observed to be acyclically modulated. This type is termed as Acyclically Amplitude Modulated Sinusoid (AAMS). The cyclically modulated GAWs are further sub-classified into two types, one of which is characterized by double-pulsing, that is termed as Cyclically Amplitude Modulated Sinusoid-2 (CAMS-2). The other type is characterized by more than two pulses per cycle. This type is termed as Cyclically Amplitude Modulated Sinusoid>2 (CAMS>2). Few or no closed phases are observed in this type. This suggests modelling the GAWs by means of a sinusoidal carrier ranging from +1 to -1, which is amplitude modulated. Examples of each of the three types of observed vocal fry GAWs, as well as of the euphonic GAW are shown in Fig. 1.

For modelling the GAWs, dominant peaks in the magnitude spectrum are picked to find coarse estimates of frequency candidates for the sinusoidal carrier. The frequency and the phase of the sinusoidal carrier are estimated by an optimization function, where the objective function minimizes the root mean squared (RMS) modelling error E between the model waveform and the observed GAW. E is obtained in dB with reference to the level of the observed GAW (Eqn. 1). The frequency and the phase candidates with the smallest modelling error are selected.

$$E = 20 \cdot \log_{10} \left(\frac{\sqrt{(g - \hat{g})^2}}{\sqrt{g^2}} \right) \text{ [dB]}, \quad (1)$$

where g is the observed GAW and \hat{g} is the estimated GAW. The upper and lower envelopes of the observed GAW are estimated by finding the upper and lower extrema, i.e., the peaks and valleys of the observed GAW respectively, and shape preserving piecewise cubic interpolation. The modulator m is obtained by superposition of the upper envelope and the lower envelope as,

$$m = \left(\frac{1+c}{2} \right) \cdot m_u + \left(\frac{-1+c}{2} \right) \cdot m_l, \quad (2)$$

$$c = \sin \left(\frac{2\pi n f_c}{f_s} + \varphi_c \right), \quad (3)$$

where m is the modulator, c is the carrier with frequency f_c and phase φ_c , f_s is the sampling frequency, m_l is the lower envelope and m_u is the upper envelope. GAWs are modelled by multiplying the estimated sinusoidal carrier with the estimated modulator. This model waveform is termed as *target model*

$$\hat{g}_t = c \cdot m. \quad (4)$$

As the euphonic GAWs have smaller amplitude modulation than the vocal fry GAWs, a *control model* \hat{g}_c is obtained by the same procedure used for modelling \hat{g}_t , except for keeping the lower and the upper envelopes constant by using their respective means. Model errors are determined separately for \hat{g}_t and \hat{g}_c . Target model error E_t and control model error E_c are obtained for each of the GAWs using Eqn. 1. For euphonic GAWs, the errors of the two models are approximately equal, because the envelopes of the GAWs fluctuate weakly only. For vocal fry GAWs, the E_c is larger than E_t , because the lower and the upper envelopes of the vocal fry GAW models are fluctuating more strongly. Hence, the difference between the

modelling errors is larger for the vocal fry subjects than for the euphonic subjects, which makes the groups distinguishable automatically.

In Fig. 1, we show for each of the four GAW types representative examples of observed GAWs, target model GAWs \hat{g}_t , and the estimates of the upper and lower envelopes used for obtaining \hat{g}_t . In Fig. 1.b and 1.c, the upper and the lower envelopes are cyclic resulting in cyclic modulator waveforms. On the other hand, acyclic envelopes result in acyclic modulators (Fig. 1.d). The magnitude spectra of the acyclic modulators are more homogeneous than the magnitude spectra of the cyclic modulators. The homogeneity of a magnitude spectrum is reflected by Shannon's Entropy

$$H = -\frac{\sum_{f=1}^N P(f) \log_2 P(f)}{\log_2 N}, \quad (5)$$

where $P(f) = \frac{S(f)}{\sum_f S(f)}$, is the probability distribution of the modulators' power spectrum $S(f) = |X(f)|^2$, $X(f)$ is the discrete fourier transform of the modulator and N is the total number of discrete frequencies f . To distinguish between CAMS-2 and CAMS>2 GAWs, the modulator-to-carrier frequency ratio $r_f = f_m/f_c$ is determined, where the modulator frequency f_m is the frequency corresponding to the maximal magnitude of the modulator's spectrum. Only frequencies below $0.95 \cdot f_c$ are considered.

III. RESULTS

Four features are extracted. The target model error E_t , the entropy H and the modulator-to-carrier frequency ratio r_f are obtained using the target model \hat{g}_t , and the control model error E_c is obtained using the control model \hat{g}_c . All the features are plotted against each to analyze variability between different voice types. In Fig. 2.a, target model errors E_t are plotted against the corresponding control model errors E_c . Differences between the modelling errors are smaller for euphonic GAWs than for vocal fry GAWs, owing to the amplitude modulation which is much stronger in the fry GAWs than in the euphonic GAWs. The mean and the standard deviation of the differences between the two model errors of euphonic GAW models are determined as 0.3115 dB and 0.2366 dB respectively. For vocal fry GAW models, they are 4.4880 dB and 1.2668 dB respectively, which makes the groups distinguishable. Sensitivity and specificity of distinguishing vocal fry from euphonic GAWs are estimated to be 0.9964 and 0.9995 respectively.

Fig. 2.b and 2.c show the entropy against E_t and E_c respectively. The entropy of the PSD is larger for

homogeneous PSDs than for inhomogeneous PSDs. For acyclic modulators, the distribution is more homogeneous than for cyclic modulators, which results in a higher entropy of acyclic modulators.

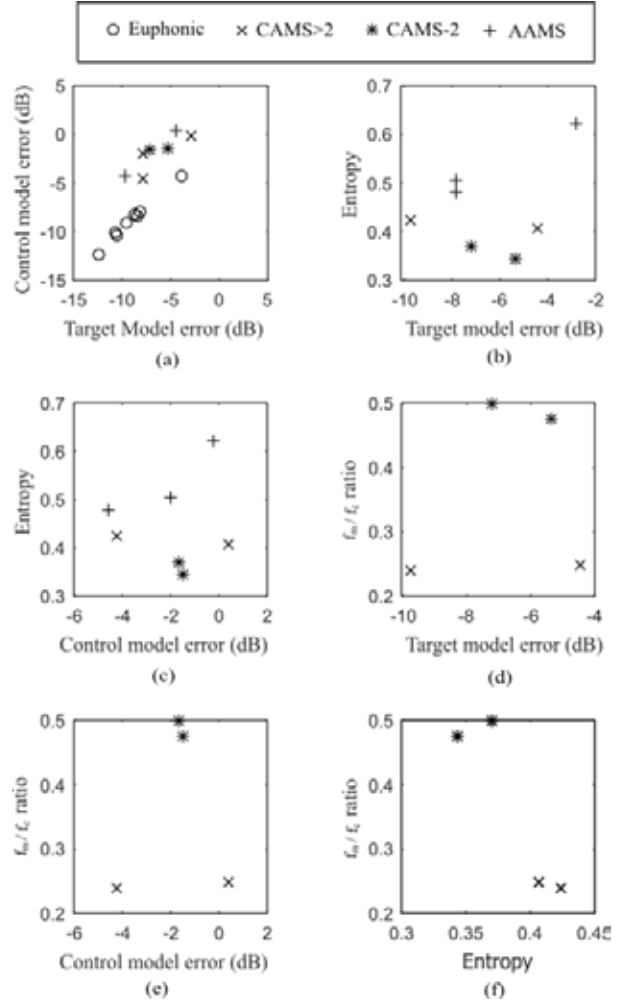


Figure 2: Control model error, target model error, entropy and f_m/f_c ratio are plotted against each other.

The modulator-to-carrier frequency ratio r_f , the modulator frequency f_m , and the carrier frequency f_c are shown in Tab. 1. From Fig. 2.d, 2.e, 2.f, and Tab. 1, it is seen that the r_f ratio is either approximately 0.5 or approximately 0.25, the former reflecting double pulsing, and the latter reflecting quadruple pulsing. In double pulsing, a meta cycle contains two pulses, whereas in quadruple pulsing, a meta cycle contains four pulses. This makes a clear distinction between the two subtypes of cyclically modulated vocal fry GAWs.

V. DISCUSSION AND CONCLUSION

We propose a prototypical method to model vocal fry and euphonic GAWs. Vocal fry and euphonia are modelled using a target model, which uses a sinusoidal

carrier multiplied by a modulator. The modelling error is determined to evaluate the fit of the model. Target model errors of euphonic and vocal fry target models are observed to be overlapping, which makes the two groups indistinguishable based on target model errors only. Hence, a control model (non-modulating model) is also used, which fits the euphonic GAWs better than vocal fry GAWs. The difference between the two model errors is found to be less than 1 dB for all the euphonic GAWs. For all seven vocal fry subjects, the difference is larger, which makes them distinguishable from the euphonic GAWs. Further, the target models are used for distinguishing different types of vocal fry GAWs.

Table 1: r_f ratio of the cyclically modulated vocal fry GAWs.

Subject code	f_m (Hz)	f_c (Hz)	r_f
JYW	102.10	411.13	0.2483
GPS	49.37	206.90	0.2386
OQP	93.98	188.72	0.4980
ISR	61.00	128.05	0.4764

For vocal fry GAWs, cyclic modulators are distinguished from acyclic modulators. The entropy of the modulator's PSD is determined from the target model for all the vocal fry speakers. The entropy observed for the three acyclic modulators is larger than the entropy observed for the four cyclic modulators. This is because the modulator's magnitude spectra of the acyclic modulators are more homogeneous than the modulator's magnitude spectra of the cyclic modulators. Cyclic modulators of the vocal fry GAW models are further classified into CAMS-2 and CAMS>2. These two are distinguished based on their r_f ratio.

The limitations of the presented approach are listed as follows. Firstly, single, double and multiple openings and closings of vocal folds are observed in a single meta cycle, which agrees with past studies [8, 10]. However, no long closed phases were observed in our data. Secondly, the waveforms are not clipped in our GAW modelling approach which might increase the modelling errors. Third, peak picking for the frequency estimation is not robust in case of noisy GAWs. In the future, waveform models for other types of vocal fry with a long closed phase should be proposed in order to establish a quantitative distinction between the signal types found in this study. In addition, distinction of vocal fry GAWs and (other) dysphonic voices should be proposed. Also, to test and improve our prototypical waveform model, more data will be needed.

VI. ACKNOWLEDGEMENT

The authors would like to thank the Division of Phoniatrics and Pediatric Audiology, University Hospital Erlangen, Germany, for providing the Glottis Analysis Tools (GAT).

REFERENCES

- [1] J. Oates, "Auditory-Perceptual Evaluation of Disordered Voice Quality," *Folia Phoniatrica et Logopaedica*, 2009,61:49–56
- [2] M. Blomgren, Y. Chen and H.R. Gilbert. "Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers," *The Journal of the Acoustical Society of America* 103.5, 1998, 2649-2658.
- [3] R. Appleman, and M. Bunch. "Application of vocal fry to the training of singers," *J Sing* 62.1 (2005): 53-59.
- [4] P. A. Keating, M. Garellek, and J. Kreiman, "Acoustic properties of different kinds of creaky voice," *ICPhS*, 2005.
- [5] Y. Horri. "Jitter and Shimmer in Sustained Vocal Fry Phonation," *Folia Phoniatrica et Logopaedica*, vol.37 (1985): 81-86
- [6] R.O. Gottliebson, L. Lee, B. Weinrich, and J. Sanders, "Voice problems of future speech-language pathologists," *Journal of Voice*, 21(6), 699-704.
- [7] Ylitalo, Riitta, and Britta Hammarberg. "Voice characteristics, effects of voice therapy, and long-term follow-up of contact granuloma patients," *Journal of Voice* 14.4 (2000): 557-566.
- [8] H. Hollien, G.T. Girard, and R.F. Coleman. Vocal fold vibratory patterns of pulse register phonation. *Folia Phoniatrica et Logopaedica*, (1977), 29(3), 200-205
- [9] P. Moore, and H. Von Leden. "Dynamic variations of the vibratory pattern in the normal larynx," *Folia Phoniatrica et Logopaedica* 10.4 (1958): 205-238.
- [10] D.G. Childers, and C. K. Lee. "Vocal quality factors: Analysis, synthesis, and perception," *the Journal of the Acoustical Society of America* 90.5 (1991): 2394-2410.
- [11] R.L.Whitehead, D.E. Metz, and B.H. Whitehead. "Vibratory patterns of the vocal folds during pulse register phonation," *The Journal of the Acoustical Society of America* 75.4 (1984): 1293-1297.
- [12] D. G. Hanson, M. D'Agostino, J. Jiang, and G. Herzon, "Clinical measurement of mucosal wave velocity using simultaneous photoglottography and laryngostroboscopy," *Annals of Otology, Rhinology & Laryngology*, 104(5),1996, 340-349.
- [13] P. Aichinger, I. Roesner, M. Leonhard, D. Denk-Linnert and B.Schneider-Stickler, "A database of laryngeal high-speed videos with simultaneous high-quality audio recordings of pathological and non-pathological voices," *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, vol.10, pp. 767–770, 2016.

MODELLING LONGITUDINAL PHASE DIFFERENCES IN A LUMPED AND DISTRIBUTED ELEMENTS VOCAL FOLD MODEL

C. Drioli¹, P. Aichinger²

¹ Department of Mathematics, Computer Science and Physics, University of Udine, Italy.

² Medical University of Vienna, Department of Otorhinolaryngology, Division of Phoniatics-Logopedics, Vienna, Austria.
carlo.drioli@uniud.it, philipp.aichinger@meduniwien.ac.at

Abstract: We discuss the representation of anterior-posterior (A-P) phase differences in vocal cord oscillations through a numerical biomechanical model involving lumped elements as well as distributed elements, i.e., delay lines. A dynamic glottal source model is illustrated in which the fold displacement along the vertical and the longitudinal dimensions is modeled using numerical waveguide components. In contrast to other models, in which the reproduction of longitudinal phase differences are impossible (e.g., in two-mass models) or not easy to control (e.g., in 3D 16-mass and multi-mass models in general), the one proposed here provides direct control over the amount of phase delay between folds oscillations at the posterior and anterior part of the glottis, while keeping the dynamic model simple and computationally efficient. The model is assessed by addressing the reproduction of oscillatory patterns observed in high-speed videoendoscopic data, in which A-P phase differences are observed, and of parameters related to the glottal area waveform. **Keywords:** High-speed video analysis, vocal folds dynamical modelling, voice quality characterization, voice disorders.

I. INTRODUCTION

Voice quality characterization is pivotal to the clinical care and science of voice disorders, because it aids the indication, selection, evaluation and optimization of medical treatment techniques. These techniques include voice therapy conducted by speech-language pathologist or logopedists, or phonosurgery, which includes surgery that aims at improving voice quality. A lack of understanding the relation between voice quality and vocal fold vibration patterns is observed. Existing models of the relations between different types of voice quality, glottal area waveform (GAW) patterns and auditory percepts are neither totally reliable nor exhaustive.

Video data acquisition and processing are recognized as essential tools for voice quality assessment and medical diagnosis. Recent research discussing connections between biomechanical modelling of the folds

and high-speed videoendoscopic or videokymographic techniques can be found in [1], [2], [3]. Voice source analysis through numerical models of the vocal folds oscillatory patterns is nowadays a mature research field, and reliable glottal models of different accuracy and complexity are available that mimic the underlying dynamics of the folds [4], [5], [6], [7], [8]. These models were originally intended to provide waveforms of phonatory acoustic emission. Validation via comparing them with high-speed video recordings provides an intermediate means of a reality check of these models.

Existing biomechanical models of the folds based on lumped elements (masses, springs and dampers) either cannot reproduce longitudinal phase differences (as the two-mass model) or do not enable controlling the phase difference in an easy and direct way (as, for example, in the 3D 16-mass model, in which the phase difference depends on a number of parameters). Conversely, we propose here an approach to fold edge modelling that allows direct control over the amount of phase delay between the folds, oscillations at posterior and anterior parts of the glottis. This edge displacement model is combined with a low-dimensional lumped-element scheme, which allows to keep the dynamic model simple and computationally efficient.

The proposed model provides a tool for both the simulation of different phonation modalities involving longitudinal phase differences, and the analysis and interpretation of videoendoscopic data for voice quality classification and medical diagnosis.

II. BIOMECHANICAL MODEL

The posterior inferior edge of each fold is represented by a single mass-spring system with stiffness k , damping r and mass m . Possible L-R asymmetry is taken into account by including two different single-mass systems, one for each fold. Superscript $\alpha \in \{l, r\}$ are used to indicate the left and the right fold respectively. The driving pressure P_m , and the resulting force F_m , are computed from the flow U_g and the inferior glottal area A_{g_i} using Newton's and Bernoulli's laws:

$$\begin{cases} m^\alpha \ddot{x}^\alpha(t) + r^\alpha \dot{x}^\alpha(t) + k^\alpha x^\alpha(t) = F_m^\alpha(t) \\ F_m^\alpha(t) = P_m(t) \cdot S_m^\alpha \\ P_m(t) = P_l - \frac{1}{2}\rho \frac{U_g(t)^2}{A_{g_i}(t)^2} \end{cases} \quad (1)$$

where S_m is the equivalent fold surface on which the pressure is exerted. The force F_m^l is perpendicular to S_m^l and oriented to the left, whilst F_m^r is perpendicular to S_m^r and oriented to the right.

The vertical phase difference of the vibration of the cord edges is essential for the modelling of self-sustained oscillations. It is represented by a distributed element introducing a delay of the displacement of the fold along the vertical axis. The propagation of the displacement along the lateral direction is represented by a propagation line introducing a delay $\tau_{sag}(y)$, y being the longitudinal position.

Inspired by the modelling approach in [6], we align the source of the displacement with a given point on the fold surface, and propagate the motion along the surface. Let x^{i-p} be the posterior displacement of the fold at the entrance of the glottis (inferior, posterior edge), and x^{s-p} the displacement at the exit (superior, posterior edge). A collision model f_X distorts the folds displacements and adds the offsets $x_0^l = -x_0$ and $x_0^r = x_0$ which are the vocal folds' resting positions¹. The displacements are computed as:

$$\begin{cases} x^{i-p,\alpha}(t) = f_X(x^\alpha(t), x_0^\alpha) \\ \quad = \begin{cases} x^\alpha(t) + x_0^\alpha & \text{if } x^l(t) < x^r(t) + 2x_0 \\ (x^l(t) + x^r(t))/2 & \text{otherwise} \end{cases} \\ x^{s-p,\alpha}(t) = f_X(x^\alpha(t - \tau_{sag}), x_0^\alpha) \end{cases} \quad (2)$$

The longitudinal A-P phase difference of the displacement is also modeled here by a distributed element. The propagation delay value τ_{lon} is selected according to the desired A-P phase difference. We also allow different L-R longitudinal phase differences (i.e., $\tau_{lon}^l \neq \tau_{lon}^r$). The displacements of the anterior edges denoted as x^{i-a} and x^{s-a} , are given by:

$$\begin{cases} x^{i-a,\alpha}(t) = x^{i-p,\alpha}(t - \tau_{lon}^\alpha) \\ x^{s-a,\alpha}(t) = x^{s-p,\alpha}(t - \tau_{lon}^\alpha) \end{cases} \quad (3)$$

Finally, a flow model converts the glottal area given by the fold displacements into the airflow at the entrance of the vocal tract. L is the length of the glottis, and the A-P axis is sliced into N_L sections, each one of length $\delta_L = L/N_L$. Due to projection, the glottal area at slice j is computed as the minimum of the area a_j^i at inferior vocal fold edge and the area a_j^s at the superior vocal fold edge. The total glottal area is then computed

¹Displacements are considered negative on the left of the medial axis, and positive on the right.

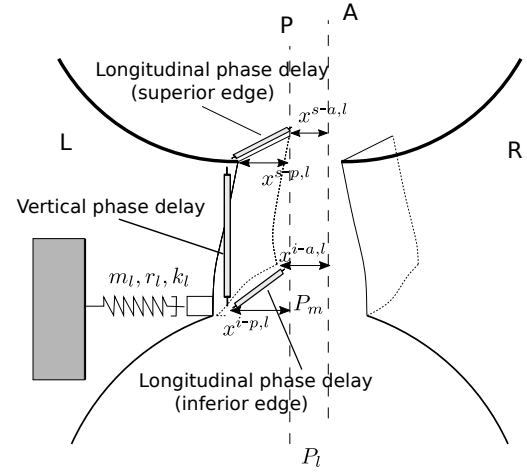


Fig. 1. Schematic view of the model: the vertical (inferior-superior) and longitudinal (anterior-posterior) phase differences of the fold displacement are modeled using three propagation lines.

as the sum of all areas along the A-P dimension, and the flow is assumed to be proportional to the total glottal area:

$$\begin{cases} x_j^{i,\alpha}(t) = x^{i-p}(t - \frac{j\tau_{lon}^\alpha}{N_L}) \\ x_j^{s,\alpha}(t) = x^{s-p}(t - \frac{j\tau_{lon}^\alpha}{N_L}) \\ a_j^i(t) = \delta_L(x_j^{i,l}(t) + x_j^{i,r}(t)) \\ a_j^s(t) = \delta_L(x_j^{s,l}(t) + x_j^{s,r}(t)) \\ U_g(t) = \sqrt{\frac{2P_l}{\rho}} \sum_{j=1}^{N_L} \min\{a_j^i(t), a_j^s(t)\} \end{cases} \quad (4)$$

where ρ is the air density, and P_l is the lung pressure.

The discretization of the equations (1)-(4) leads to a discrete-time system that is numerically solved to obtain an estimate of the glottal flow $U_g(nT_s)$ and of the folds displacements $x_j^{i,\alpha}(nT_s)$ and $x_j^{s,\alpha}(nT_s)$, at discrete time n , with sampling interval T_s and slice index j .

The sampling frequency $F_s = (1/T_s) = 22050$ Hz. The oscillatory patterns are visualized as if the folds were observed from above. The oscillation patterns are visually compared to high-speed video data.

In this preliminary investigation, we aim at reproducing observed oscillatory patterns of the folds observed in the video data qualitatively. We do not aim to achieve copy-synthesis at this time. An approximate estimate of the period was obtained from the video data by tuning the m and k parameters of the folds manually by trial and error. The natural frequency of the mass-spring system is $f_0 = 1/2\pi\sqrt{k/m}$, but vibration frequency may be different.

The direct control of delays, masses, spring constants enables simulation of various asymmetric vibratory patterns. Figure 2 illustrates the simulation from a

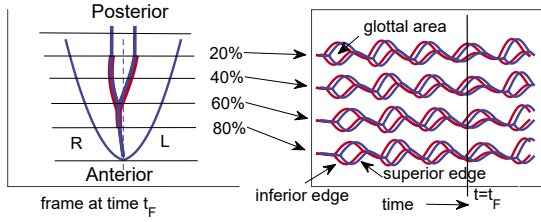


Fig. 2. Example of oscillatory pattern obtained by L-R mass unbalancing and L-R symmetric A-P phase differences. Left panel: the vocal folds' contours as seen from above; Right panel: time evolution of the folds. Displacements at four equally spaced locations along the longitudinal axis.

configuration in which the natural frequencies of the left and right fold are different, i.e., $f_0^l = 210$ Hz, and $f_0^r = 180$ Hz. The two longitudinal phase delays are $\tau_{lon}^l = \tau_{lon}^r = 1.8$ msec. Given that the resulting glottal cycle length is approximately 5 msec, the maximal A-P phase difference is approximately 130 degrees. With regard to the L-R difference of masses, the resulting oscillation is characterized by paramedian collision.

III. OBSERVED ASYMMETRIC PATTERNS: ANALYSIS AND MODELLING

In this section, the model is assessed qualitatively by empirically tuning its parameters to replicate some typical oscillatory patterns observed in high speed videoendoscopic data, in which A-P phase differences, and paramedian collisions due to the mass unbalancing, are observed. Figure 3 shows a complicated vibration pattern affected by L-R asymmetry, as well as A-P longitudinal phase delay (upper panel), and its imitation provided by the model. The left and right natural frequencies were set to respectively $f_0^l = 210$ Hz, and $f_0^r = 180$ Hz, and the two longitudinal phase delays were set to $\tau_{lon}^l = \tau_{lon}^r = 1.8$ msec.

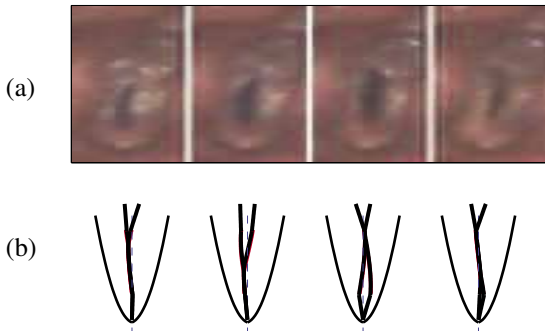


Fig. 3. (a) A selection of frames within one cycle from an high-speed video (HSV) (subject: S2), and (b) a selection from the model simulation. The vibration pattern is characterized by both L-R asymmetry, as well as A-P phase delay.

As a first attempt to quantitatively assess the model capability of fitting AP asymmetric vibration, a set of

measures related to the glottal area waveform (GAW) are computed². To this aim, we refer to the left and the right hemi-GAW ($hGAW^L$ and $hGAW^R$) defined as the time-varying area of the left and the right half of the glottis, and satisfying $hGAW^L + hGAW^R = GAW$. Similarly, we refer to the anterior and the posterior hemi-GAW ($hGAW^A$ and $hGAW^P$) defined as the time-varying area of the anterior and the posterior half of the glottis, and satisfying $hGAW^A + hGAW^P = GAW$. For each one of these waveforms, the instants corresponding to maximum excursions, i.e., T_i^R , T_i^L , T_i^A , T_i^P , are obtained by picking peaks (where i is the cycle index), as shown in Fig. 4. Finally, measures of L-R and A-P asymmetries are computed as $\Delta T_i^{LR} = T_i^R - T_i^L$, and $\Delta T_i^{AP} = T_i^A - T_i^P$.

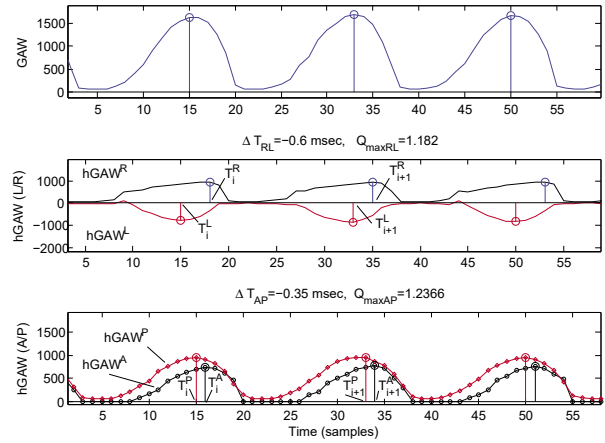


Fig. 4. Result of peak picking procedure on GAW (top), L-R hGAWs (center), and A-P hGAWs (bottom).

A selection of five recordings ($S1 - S5$) from the Laryngeal High-Speed Video Database of Pathological and Non-Pathological Voices described in [11] is used, in which several examples of patterns with A-P phase differences are observed. The peak picking based hGAW analysis is applied to both the natural video data, and the model simulations thereof. Results are reported in Tables I and II (average over 10 periods). The Tables also report the average value of the $hGAW^R$ vs $hGAW^L$ and $hGAW^A$ vs $hGAW^P$ ratios, denoted Q_{LR} and Q_{AP} respectively, although no attempt was made to obtain the same values of the HSV data.

Sample $S1$ is characterized by a slight L-R asymmetry and low A-P phase difference (a double pulsing is also observable as a small secondary peak in the A-P hGAWs, however we disregard this component in the present investigation). In the model, the left and right fold oscillating frequency were set slightly unbalanced and the two longitudinal phase delays were equally

²Videos are graphically segmented to obtain the spatio-temporal vibration patterns as described in [10].

TABLE I
THE HGAW-BASED ASYMMETRY ANALYSIS ON HSV DATA.

Subj.	ΔT^{LR} (msec)	ΔT^{AP} (msec)	Q_{AP}	Q_{LR}
S1	0.2	0.15	1.03	1.36
S2	-0.98	0.64	0.97	1.0
S3	0.65	-0.26	1.0	1.1
S4	0.01	-0.4	1.11	1.16
S5	-0.6	-0.35	1.18	1.23

TABLE II
THE HGAW-BASED ASYMMETRY ANALYSIS ON SIMULATIONS.

Setup	ΔT^{LR} (msec)	ΔT^{AP} (msec)	Q_{AP}	Q_{LR}
M1	0.13	0.11	0.88	1.0
M2	-0.53	0.45	1.2	0.99
M3	0.38	-0.11	0.71	1.0
M4	0.03	-0.39	0.91	0.99
M5	-0.49	-0.22	0.77	1.2

set to $\tau_{lon}^l = \tau_{lon}^r = 10$ samples, resulting in an A-P phase delay of approximately 0.11 msec. In sample *S2*, Figure 3, a complicated vibration pattern due to L-R mass differences, as well as A-P phase differences, is observed. The left and right fold oscillating frequency were set to respectively $f_0^l = 210$ Hz, and $f_0^r = 180$ Hz to obtain a negative ΔT^{LR} value, and the two longitudinal phase delays were set to $\tau_{lon}^l = \tau_{lon}^r = 40$ samples, resulting in a A-P phase delay of approximately 0.45 msec. Sample *S3* is characterized by a moderate P-to-A delay, resulting in a negative ΔT^{AP} , and a moderate L-R asymmetry. The natural oscillating frequency were set to $f_0^l = 180$ Hz, and $f_0^r = 210$ Hz to obtain a positive ΔT^{LR} value, and the two longitudinal phase delays were set to $\tau_{lon}^l = \tau_{lon}^r = 10$ samples. A negative value of ΔT^{AP} was successfully obtained by propagating the fold displacement from the P-end of the longitudinal propagation line. Sample *S4* is characterized by P-to-A delay and L-R symmetry. The left and right folds' oscillating frequencies were equally set to $f_0^l = f_0^r = 210$ Hz, and the two longitudinal phase delays were equally set to $\tau_{lon}^l = \tau_{lon}^r = 34$ samples, resulting in an A-P phase delay of -0.39 msec. Finally, sample *S5* has an L-R asymmetric A-P negative phase delay, reproduced by setting $\tau_{lon}^l = 1$ sample, and $\tau_{lon}^r = 34$ samples, providing a A-P phase delay of approximately 108 degrees only on the left fold and correctly matching the sign of ΔT^{LR} and ΔT^{AP} .

IV. CONCLUSIONS

We discussed the modelling of longitudinal A-P phase differences in vocal folds oscillations by means of a lumped and distributed-elements vocal fold model. The distributed elements (delay lines) represent the longitudinal propagation of the fold displacement. We

address the reproduction of the oscillatory patterns observed in high-speed video recordings of the folds, including vertical and longitudinal phase differences and left-right fold mass unbalancing. The model was assessed qualitatively, by empirically tuning its parameters to replicate some oscillatory patterns observed in high speed videoendoscopic data. Asymmetry measures derived from the peak analysis of the L-R and A-P hemi-GAWs were compared to those obtained from the HSV data. The comparisons suggest that it is possible to independently control the L-R asymmetries and A-P phase differences by tuning the left and right mass unbalancing, and the longitudinal propagation delay.

V. ACKNOWLEDGEMENTS

This work was supported by the Austrian Science Fund (FWF): KLI 722-B30. The authors also want to thank Jean Schoentgen for a fruitful discussion.

REFERENCES

- [1] A. P. Pinheiro, D. E. Stewart, C. D. Maciel, J. C. Pereira, and S. Oliveira, "Analysis of nonlinear dynamics of vocal folds using high-speed video observation and biomechanical modeling," *Digital Signal Processing*, vol. 22, no. 2, pp. 304–313, 2012.
- [2] M. Döllinger, P. Gómez, R. R. Patel, C. Alexiou, C. Bohr, and A. Schützenberger, "Biomechanical simulation of vocal fold dynamics in adults based on laryngeal high-speed videoendoscopy," *PLOS ONE*, vol. 12, no. 11, pp. 1–26, 11 2017. [Online]. Available: <https://doi.org/10.1371/journal.pone.0187486>
- [3] C. Drioli and G. L. Foresti, "Accurate glottal model parametrization by integrating audio and high-speed endoscopic video data," *Signal, Image and Video Processing*, vol. 9, pp. 1451–1459, 2015.
- [4] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *The Bell Syst. Tech. J.*, vol. 51, no. 6, pp. 1233–1268, July-August 1972.
- [5] T. Koizumi, S. Taniguchi, and S. Hiromitsu, "Two-mass models of the vocal cords for natural sounding voice synthesis," *J. Acoust. Soc. Am.*, vol. 82, no. 4, pp. 1179–1192, October 1987.
- [6] I. R. Titze, "The physics of small-amplitude oscillations of the vocal folds," *J. Acoust. Soc. Am.*, vol. 83, no. 4, pp. 1536–1552, April 1988.
- [7] X. Pelorson, A. Hirschberg, R. R. van Hassel, and A. P. J. Wijnands, "Theoretical and experimental study of quasisteady-flow separation within the glottis during phonation. Application to a modified two-mass model," *J. Acoust. Soc. Am.*, vol. 96, no. 6, pp. 3416–3431, December 1994.
- [8] J. C. Lucero, J. Schoentgen, J. Haas, P. Luizard, and X. Pelorson, "Self-entrainment of the right and left vocal fold oscillators," *The Journal of the Acoustical Society of America*, vol. 137, no. 4, pp. 2036–2046, 2015.
- [9] C. Drioli, "A flow waveform-matched low-dimensional glottal model based on physical knowledge," *J. Acoust. Soc. Am.*, vol. 117, no. 5, pp. 3184–3195, May 2005.
- [10] J. Lohscheller and U. Eysholdt, "Phonovibrograph visualization of entire vocal fold dynamics," *The Laryngoscope*, vol. 118, no. 4, pp. 753–758, 2008.
- [11] 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 *Proc. Int. Conf. Lang. Resour. Eval.*, vol. 10, 2016, pp. 767–770.

EXTRACTING KINEMATIC VOCAL FOLD PARAMETERS FROM VIDEOKYMOGRAMS VIA SIMULATION OF CLINICAL DATA

S. Bulusu¹, S. P. Kumar², J. G. Svec³, P. Aichinger¹

¹ Medical University of Vienna, Department of Otorhinolaryngology, Division of Phoniatrics-Logopedics, Vienna, Austria

² SSN College of Engineering, Department of Biomedical Engineering, Kalavakkam – 603110, Chennai, India

³ Palacky University, Faculty of Science, Department of Biophysics, Voice Research Laboratory, Olomouc, Czech Republic

sridharbulusuvie@gmail.com, pravinkumars@ssn.edu.in, jan.svec@upol.cz, philipp.aichinger@meduniwien.ac.at

Abstract: This paper proposes the extraction of vocal fold parameters from videokymographic images. A previously developed model of vocal fold vibrations is generalized to include left-right phase differences and paramedian collisions of the vocal folds. A model fitting error minimization procedure is implemented in order to extract kinematic parameters. 55 clinical and 50 synthetic kymograms are used to evaluate the procedure using the “Structural Dissimilarity Index Measure” (DSSIM) and the “Cross Uncorrelation” (CUC) as error measures. After fitting the clinical kymograms, probability density functions (PDFs) of the model parameters are obtained. The PDFs are used to generate synthetic kymograms with random parameters. The synthetic kymograms are used to evaluate the performance of the fitting procedure by measuring the errors between the randomly generated parameter values and those obtained through the fitting procedure. The relative error ranged between 0.052% and 44.95%.

Keywords: Videokymograms, vocal fold kinematics, parameter extraction

I. INTRODUCTION

Quality of human voice strongly depends on kinematic parameters of vocal fold vibration, and disturbance of vocal fold vibration may be a sign of voice disorders. Studying vocal fold vibration patterns is therefore an important task that aids voice quality assessment and may help revealing the cause of a voice problem.

This paper explores a kinematic mucosal wave model which enables simulating the motion of the vocal folds [1]. The model uses 10 adjustable parameters controlling the vocal fold vibration kinematics. In addition to the cross sectional shapes of the vocal folds, the model produces kymograms, showing the vibratory pattern of the vocal fold as if observed laryngoscopically. These were shown to

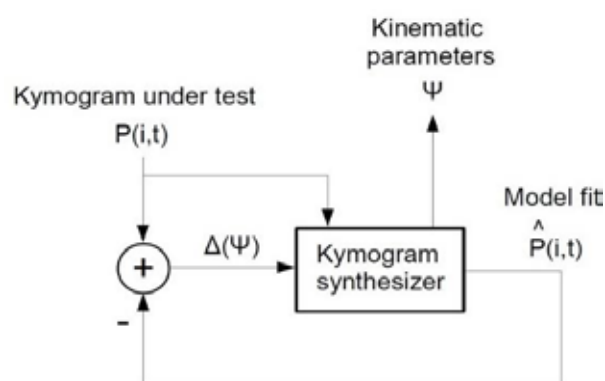


Fig. 1: Block diagram for parameter estimation. The model fit $\hat{P}(i,t)$ is subtracted from the kymogram under test $P(i,t)$. The error and the kymogram under test are fed into the synthesizer and a new model fit is created, updating the kinematic parameters ψ .

be comparable to clinically recorded vibratory patterns [1].

In this work we propose extracting vocal fold vibration parameters from clinical high-speed videokymographic images using the kinematic mucosal wave model [2]. The model parameters are estimated by minimizing the error between the synthetically generated kymogram and the clinical kymogram. PDFs are obtained. Using the PDFs 50 synthetic kymograms are randomly generated and fitted in order to evaluate the performance of the fitting.

II. METHODS

A. Corpora

Two corpora are used. The first corpus contains 55 clinical kymograms, which were recorded during sustained phonations obtained from VKG examinations of patients at the Voice and Hearing Centre Prague during the years 2015-2017.

The PDFs of the parameters resulting from the parameter estimation procedure are used to synthesize

Table 1: Distributions for each of the parameters obtained using DSSIM and CUC as the error measure. N denotes a normal distribution; the arguments are mean, standard deviation, lower bound and upper bound. U denotes a uniform distribution; the arguments are the lower and upper bounds.

Parameter	DSSIM	CUC
f_0 (Hz)	N(238.1, 54.3,70,400)	N(237.4, 53.88,70,400)
w_h (cm)	N(0.043, 0.017, 0, 0.15)	N(0.053, 0.017, 0, 0.15)
ϕ_R (degrees)	N(0.084, 1.29, -3.14, 3.14)	N(-0.11, 1.34, -3.14, 3.14)
ϕ_L (degrees)	N(0.193, 1.38, -3.14, 3.14)	N(0.06, 1.40, -3.14, 3.14)
A_u (cm)	U(0.05, 0.1)	U(0.05, 0.1)
A_l (cm)	U(0.02, 0.1)	U(0.02, 0.1)
ϕ_v (degrees)	U(10, 120)	U(10, 120)
α (degrees)	N(9.98, 9.97, -10, 40)	N(6.81, 10.11, -10, 40)
d_{ul} (cm)	U(0.2, 0.5)	U(0.2, 0.5)
E (1)	U(0.025,100)	U(0.025,100)

a synthetic corpus of 50 kymograms. Since the values of their parameters are known, they are used as reference for evaluating the fitting procedure.

B. Parameter Extraction

The used kinematic model of laryngeal kymographic images [1] is generalized to allow for phase differences between the left and the right vocal folds. Each point on the surface of the vocal fold is simulated to be moving in a circle. The phases and radii are chosen in such a way as to create a wave-like motion from bottom to top.

To extract the parameters, clinical kymograms are cropped at first such that only vocal fold tissue is visible and the glottal midline is in the middle.

The vibration frequency is estimated from the kymogram by an autocorrelation method. Using this frequency estimate and default values for the other parameters, an initial model fit is obtained.

Third, the remaining parameters are estimated as follows: Prior to the error calculation, the brightness of the fitted histogram is matched to that of the kymogram under test [5]. The error obtained from the kymogram under test and the synthetic kymogram is minimized using Golden section search combined with parabolic interpolation [6, 7]. The optimization of the parameters is carried out one by one in a loop. After optimizing the last parameter the loop is reiterated if the loop index is below five and the error has not reduced. The order of optimization in each loop is: f_0 , ϕ_R , ϕ_L , w_h , A_u , A_l , ϕ_v , α , d_{ul} and E .

In order to obtain a measure that reflects the significance of each fitting parameter, the cumulative

decrease in model error is calculated for each parameter.

In this study two different error measures are used independently. The first is the Structural Dissimilarity Index Measure (*DSSIM*), which is based on the Structural Similarity Index Measure (*SSIM*) and given by equations (1) to (4)

$$DSSIM(Y, Z) = \frac{1-SSIM(Y, Z)}{2} \quad (1)$$

Where Y and Z are the matrices containing the pixel intensities for all rows and columns of the two images respectively

$$SSIM(Y, Z) = \frac{(2\mu_Y\mu_Z+C_1)(2\sigma_{YZ}+C_2)}{(\mu_Y^2+\mu_Z^2+C_1)(\sigma_Y^2+\sigma_Z^2+C_2)} \quad (2)$$

And

$$C_1 = (0.01 \cdot L)^2 \quad (3)$$

And

$$C_2 = (0.03 \cdot L)^2 \quad (4)$$

where L is the dynamic range, i.e., the difference of maximum and minimum possible value, μ_Y is the mean of the image Y , σ_{YZ} is the covariance of images Y and Z , and σ_Y is the variance of image Y .

The second error measure is the Cross Uncorrelation (CUC) calculated from the normalized Cross Correlation (CC) [8] from equations (5) and (6):

$$CUC(Y, Z) = \frac{1-CC(Y, Z)}{2} \quad (5)$$

where CC is given by:

$$CC(Y, Z) = \frac{\sum_{ab}(Y_{ab}-\mu_Y)(Z_{ab}-\mu_Z)}{\sqrt{\sum_{ab}(Y_{ab}-\mu_Y)^2 \sum_{ab}(Z_{ab}-\mu_Z)^2}} \quad (6)$$

Y_{ab} denotes the pixel intensity of image Y in the a^{th} row and the b^{th} column and μ_Y the mean intensity of image Y .

Since the normalized CC ranges from -1 (worst possible fit) to 1 (best possible fit), the CUC will take values from 0 (best possible fit) to 1 (worst possible fit). The SSIM ranges from 0 to 1 and therefore the DSSIM ranges from 0 to 1/2.

III. RESULTS AND DISCUSSION

Table 2: relative error for each parameter calculated as the ratio of the standard deviation of the error divided by the possible range of the parameter

Parameter	DSSIM	CUC
f_0 (Hz)	0.052%	0.51%
w_h (cm)	3.56%	6.18%
ϕ_R (degrees)	2.49%	2.77%
ϕ_L (degrees)	2.42%	13.86%
A_u (cm)	3.86%	7.35%
A_l (cm)	20.37%	27.02%
ϕ_v (degrees)	18.51%	25.7%
α (degrees)	17.21%	18.12%
d_{ul} (cm)	31.87%	38.11%
E (1)	44.95%	39.29%

for the fitted PDFs of the parameters. The parameters' means are somewhat different with regard to the used error measure. This indicates a dependence of the estimated parameter values on the error measure used.

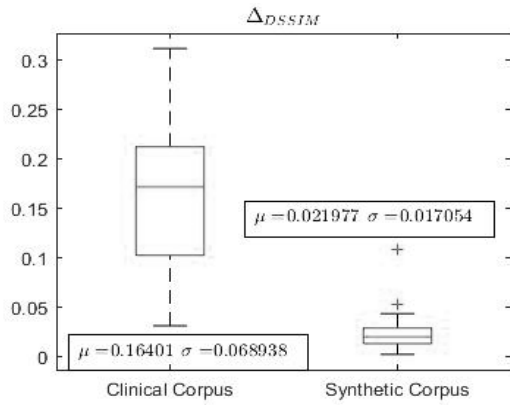


Fig. 2: Boxplot of total error for clinical and synthetic corpus when using DSSIM as error measure. μ is the mean error and σ the standard deviation of the error.

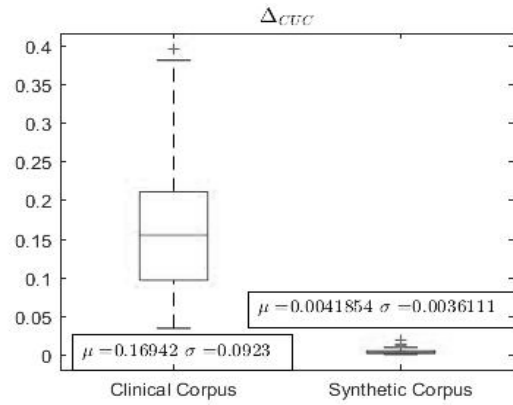


Fig. 3: Boxplot of total error for clinical and synthetic corpus when using CUC as error measure. μ is the mean error and σ the standard deviation of the error.

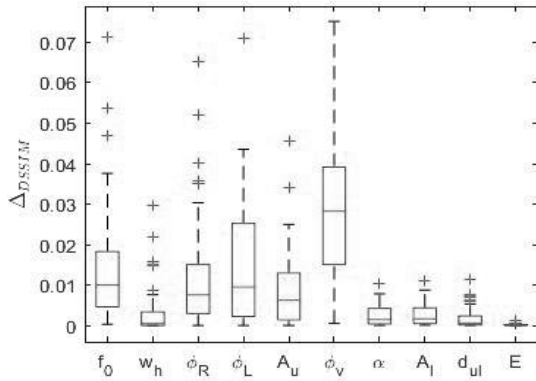


Fig. 4: Boxplot of cumulative improvement for each parameter when using DSSIM as error

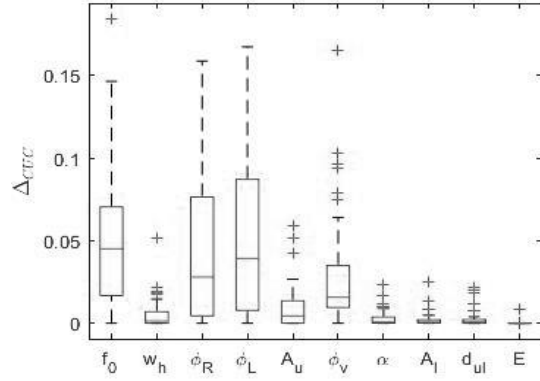


Fig. 5: Boxplot of cumulative improvement for each parameter when using CUC as error

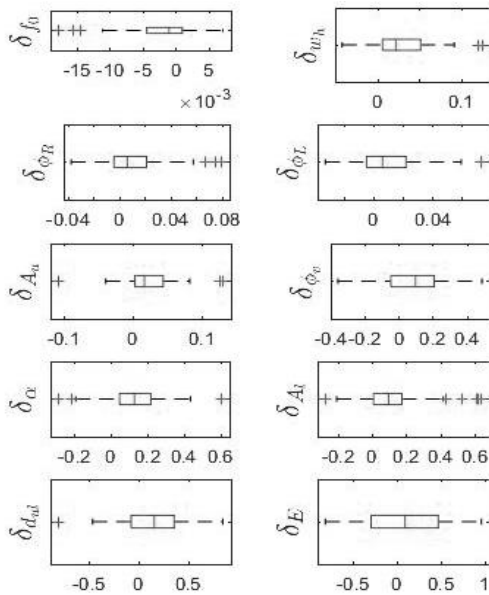


Fig. 6: Box plot of error relative to the range for each parameter when using DSSIM as error measure

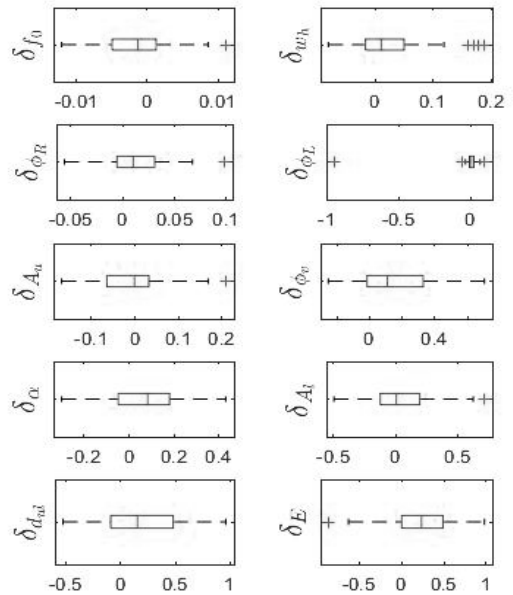


Fig. 7: Box plot of error relative to the range for each parameter when using CUC as error measure

Table 2 shows the relative estimation error for each parameter. The relative error is lowest for f_0 , probably due to the fact that f_0 is estimated via an autocorrelation method prior to the estimation using the fitting procedure, and highest for E and d_{ul} . The relative error is higher for CUC than for DSSIM, except for the parameter E . A_l , ϕ_v , α and d_{ul} have high relative errors, probably because they control primarily the lower part of the vocal fold that is partly hidden and becomes visible in the kymogram only during the closing phase. The large error of the extent of the mucosal wave beyond the glottal edge E is likely due to the fact that this parameter introduces relatively small changes in the shading of the upper surface and the error measures used here are not sensitive to this. Fig. 2 and 3 show the total error for the clinical and the synthetic corpus when using DSSIM and CUC as error measures respectively. The error for the clinical corpus is much higher than that for the synthetic corpus. This indicates that the clinical the clinical kymograms' features are not covered by the model. Indeed, for simplicity, some of the parameters (i.e., A_u , A_l , ϕ_v , α , d_{ul} , E) were considered to be identical for the left and right vocal folds, while they could be different in reality. Figs. 4 and 5 show the cumulative improvement of the error for each parameter in the clinical corpus. Surprisingly, these two figures clearly depict that the order of significance depends on the error measure used. However, in general, the parameters can be classified into five parameters of major importance and five of less importance, which are the same for both error measures used. The five most important parameters are f_0 , w_h , ϕ_R , ϕ_L , A_u , and the others are the less important parameters.

Figs. 6 and 7 show the boxplots of the errors relative to the parameter range. Not all errors are symmetrically distributed around zero, indicating bias of the fitting procedure. When using DSSIM as the error measure, the error boxes of parameters w_h , A_u , A_l , and α did not include zero. When using CUC as the error measure the only parameter error box which did not include zero was E .

V. CONCLUSION

Since the error is much higher in the clinical corpus than in the synthetic corpus, one can conclude that not all existing vibration patterns are captured by the selected model parameters. Also, not all parameters impact the kymograms strongly. This is important for creating future models and selecting parameters that reflecting variability in clinical observations. Moreover, when using different error measures, the parameter means are significantly different except for f_0 . When using the DSSIM error measure the errors were more biased, but had smaller standard deviation

as compared to when using CUC. This suggests that the DSSIM introduces systematic error, whereas CUC introduces random errors mainly. Further research is therefore necessary to improve the fitting procedure. It is to be noted that the synthetic corpus was generated purely based on the parameter distributions of the clinical corpus. This means, that physically unreasonable kymograms could also have been generated. This can be improved if the relations between the parameters are considered as well, which can be done for example by employing a method which takes into account the covariance between the parameters.

VI. ACKNOWLEDGEMENTS

The authors thank Dr. Jitka Vydrova from the Voice and Hearing Centre Prague for providing the clinical videokymographic images. This work was supported by the Austrian Science Fund (FWF): KLI 722-B30. In the Czech republic, the project was supported by the Czech Science Foundation (GA CR) project no. 19-04477S.

REFERENCES

- [1] S.P. Kumar, and J.G. Svec, "Kinematic model for simulating mucosal wave phenomena on vocal folds," *Biomed. Signal Process*, vol. 49, pp. 328–337, 2019.
- [2] Q. Qiu, and H. Schutte, "Real-time kymographic imaging for visualizing human vocal-fold vibratory function," *Rev. Sci. Instrum*, vol. 78(2), p. 024302, 2007.
- [3] R.C. Scherer, D. Shinwari, K.J. De Witt, C. Zhang, B.R. Kucinski, and A.A. Afjeh, "Intraglottal pressure profiles for a symmetric and oblique glottis with a divergence angle of 10 degrees," *J. Acoust. Soc. Am*, vol. 109(4), pp.1616-1630, 2001.
- [4] S. Li, R.C. Scherer, M. Wan, and S. Wang, "The effect of entrance radii on intraglottal pressure distributions in the divergent glottis," *J. Acoust. Soc. Am*, vol. 131(2), pp. 1371-1377, 2012.
- [5] R.C. Gonzalez, and R.E. Woods, "Histogram Matching (Specification)," in *Digital Image Processing*, 2nd ed., Prentice Hall, NJ: Upper Saddle River, 2002, pp. 75-146.
- [6] G.E. Forsythe, M.A. Malcolm, and C.B. Moler, "*Computer Methods for Mathematical Computations*," Prentice Hall, NJ: Englewood Cliffs, 1976.
- [7] R.P. Brent, "*Algorithms for Minimization without Derivatives*," Prentice-Hall, NJ: Englewood Cliffs, 1973.
- [8] H.B. Mitchell, "Image Similarity Measures," in *Image Fusion*, Springer, Berlin, Heidelberg: 2010, pp. 167-185.

VOCAL FOLD OSCILLATORS AT LARGE ASYMMETRIES

J. C. Lucero¹, X. Pelorson², A. V. Hirtum²

¹ Dept. Computer Science, University of Brasília, Brazil

² LEGI, UMR CNRS 5519, Grenoble Alpes University, Saint-Martin-d'Hères, France

lucero@unb.br, xavier.pelorson@univ-grenoble-alpes.fr, annemie.vanhirtum@univ-grenoble-alpes.fr

Abstract: This paper investigates threshold conditions of the vocal fold oscillation in the presence of a natural frequency asymmetry. Theoretical expressions for the subglottal threshold pressure and frequency are derived from a simple dynamical model of the vocal folds, and compared to data measured from a mechanical replica of the larynx under both symmetrical and asymmetrical configurations. The results demonstrate good agreement between theory and experiments, and show that the oscillation threshold is sensitive to the asymmetry with distinct behaviors between regions of low vs. high asymmetry.

Keywords: Vocal folds, oscillation threshold, asymmetry, mechanical replica

I. INTRODUCTION

The vocal folds constitute a pair of coupled oscillators that, in normal configurations, oscillate with in-phase synchrony. Pathological conditions create right-left asymmetries which hamper the oscillation and may induce complex entrainment regimes and other nonlinear phenomena [1].

In a recent theoretical study [2], asymmetries in the stiffness of the vocal folds were analyzed and different behaviors at small vs. large asymmetry were detected. At small asymmetry, the oscillation threshold value of the subglottal pressure increases with the asymmetry. The curve relating threshold pressure and right/left stiffness ratio has a “U”-shape characteristics, with the minimum at the symmetric configuration. Mathematically, the threshold corresponds to a Hopf bifurcation. At large asymmetry, on the other hand, the threshold pattern changes to a double Hopf bifurcation and the oscillation threshold pressure assumes a constant value.

The detected difference may have relevant consequences, not only for the understanding of the dynamics of the vocal fold oscillation, but also for the application of the oscillation threshold pressure as a parameter for clinical diagnosis [3]. Thus, the present study has the purpose of exploring further the effect of

asymmetries on the oscillation threshold. At the same time, it will seek validation of the theoretical results by using data collected from a mechanical replica of the vocal folds.

II. METHODS

A. Theoretical model

The vocal folds are represented as a coupled system of two one-degree-of-freedom oscillators of the form

$$\begin{aligned}\ddot{x}_r + \beta(1 + x_r^2)\dot{x}_r + \omega_r^2 x_r &= \alpha(\dot{x}_r + \dot{x}_l) \\ \ddot{x}_l + \beta(1 + x_l^2)\dot{x}_l + Q^2 \omega_r^2 x_l &= \alpha(\dot{x}_r + \dot{x}_l)\end{aligned}\quad (1)$$

where x is the normalized tissue displacement and subindices r, l designate the right and left folds, respectively, β is the damping, ω_r is the natural angular frequency of the right vocal fold, Q is a coefficient of natural frequency asymmetry, and α is the aerodynamic coupling [2].

Further,

$$\alpha = \frac{S^2 P_s}{k_t a_0 c M}, \quad (2)$$

where S is the area of the medial surface of the vocal folds, P_s is the subglottal pressure, k_t is a transglottal pressure coefficient, a_0 is the glottal area at rest, c is the mucosal wave velocity and M is the vocal fold mass. Coefficient k_t depends on the glottal area and is modeled as

$$k_t = \frac{E}{a_0} + F, \quad (3)$$

where E and F are empirical coefficients [4].

The oscillation threshold pressure P_{th} is obtained by a standard stability analysis of Eqs. (1). In the case of a low asymmetry, the analysis yields

$$P_{th} = P_0(1 + F'a_0) \left[1 + \left(\frac{\omega \Delta}{\beta} \right)^2 \right], \quad \text{for } \Delta \leq \frac{\beta}{\omega}, \quad (4)$$

where $P_0 = cME\beta/(2S^2)$, $F' = F/E$, ω is the angular frequency of the oscillation, and Δ is a normalized

asymmetry coefficient which maps $Q \in [-\infty, +\infty]$ into $\Delta \in [-1, +1]$ and is given by

$$\Delta = \frac{1-Q^2}{1+Q^2}, \quad (5)$$

and

$$\omega = \omega_r \sqrt{\frac{1+Q^2}{2}} \quad (6)$$

is the angular frequency of the oscillation.

In the case of a large asymmetry, the threshold pressure is

$$P_{th} = 2P_0(1 + F'a_0), \text{ for } \Delta \geq \frac{\beta}{\omega}, \quad (7)$$

and the angular frequency of the oscillation is given by the real solutions to

$$\omega^4 - [(1+Q^2)\omega_r^2 - \beta^2]\omega^2 + Q^2\omega_r^4 = 0 \quad (8)$$

Eq. (4) models the threshold pressure when both folds oscillate in synchrony at the same frequency given by Eq. (6). The threshold pressure has a minimum at $\Delta = 0$ (full symmetry) and increases with $|\Delta|$. Eq. (7), on the other hand, represents the threshold when the folds oscillate independent of each other, each one at its own frequency given by Eq. (8). In this case, the threshold pressure is independent of Δ .

B. Data

Oscillation threshold data were collected from a mechanical replica of the vocal folds. The replica and data collection method has been described in detail elsewhere [5,6]. Briefly, the replica consists of two parallel latex sleeves filled with water under pressure and supported by a metallic structure. The latex sleeves represent the vocal folds in a 3:1 scale. Air from a pressure reservoir is blown through a third latex sleeve, representing the glottal passage, situated in-between the two vocal fold sleeves and perpendicular to them.

Values of oscillation threshold pressure and frequency were obtained by increasing the air pressure upstream of the replica from zero until an oscillation of the sleeves was detected. The time instant of oscillation onset was determined by spectral analysis of the upstream pressure signal, and the mean upstream pressure and oscillation frequency at that instant were computed. The glottal area at rest was determined from pictures taken by a digital camera, calibrated with a benchmark grid.

Three experiments were performed, in which measures of the above parameters (pressure, frequency and area) were taken at various values of internal (water) pressures of the fold sleeves, in symmetrical and asymmetrical configurations. In addition, the mechanical constants of one of the vocal fold sleeves

(natural frequencies and bandwidths) were measured for various internal pressure values, by means of a shaker in conjunction with a laser vibrometer.

III. RESULTS

A. Frequency response

The frequency response of the vocal fold sleeve is shown in Fig. 1. The first natural frequency (f_1) has a linear relation with the internal pressure (h), and is well fitted by the equation $f_1 = 52.7 + 11.1h$. Its bandwidth (b_1) was simply approximated with the constant value $b_1 = 13.1$ Hz. Thus, we have $\beta = 2\pi b_1 = 82.3 \text{ s}^{-1}$.

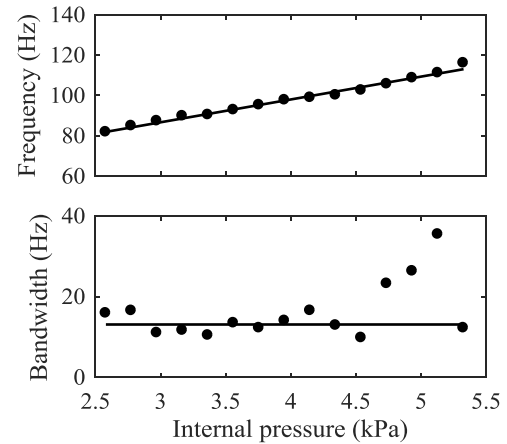


Fig. 1. First natural frequency (top) and bandwidth of a vocal fold sleeve. The solid lines show the fitted approximations.

B. Experiment 1

In this experiment, the internal pressure of both folds was varied simultaneously between 4 kPa and 8.5 kPa, keeping a symmetrical configuration. The results are shown in Fig. 2.

The threshold pressure and frequency are well approximated by Eqs. (4) and (6), respectively, with $\Delta = 0$, $P_0 = 635$ Pa, $F' = -0.0535 \text{ Pa/cm}^2$, and $\omega = \omega_r = 2\pi f_1$. However, the oscillation frequency seems to follow a nonlinear relation which is not captured by the natural frequency f_1 reported in subsection III.A. A possible cause may be that the nonlinearity becomes evident at large values of internal pressure, whereas the frequency response was measured at a lower pressure range.

The glottal area decreases with the internal pressure, due to the increase in volume of water held inside each vocal fold sleeve. The measured values are well approximated by the quadratic equation $a_0 = -0.77h^2 + 8.26h - 13.7$, and this equation was applied in Eq. (4) when computing the threshold pressure. Note

that the increase in threshold pressure with the internal pressure is a direct consequence of the glottal area variation.

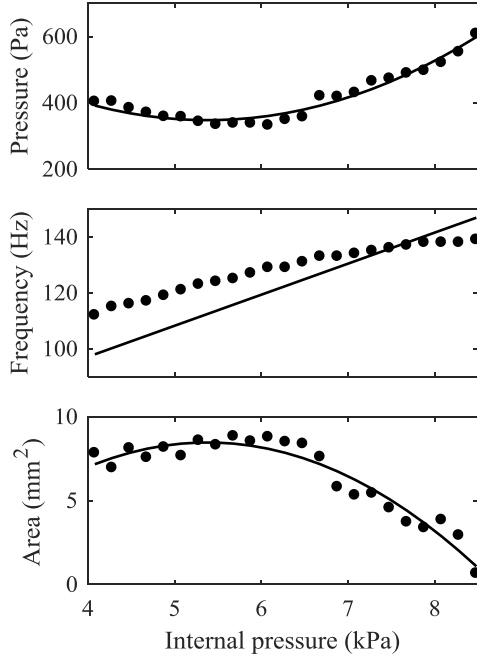


Fig. 2. Results of Experiment 1. Top: oscillation threshold pressure, middle: oscillation frequency, bottom: glottal area at rest. The solid lines show the fitted approximations.

C. Experiment 2

In this experiment, the internal pressure of one fold was fixed at 6.0 kPa whereas the internal pressure of the other was varied between 4.1 kPa and 8.3 kPa. The intention was to obtain both negative and positive values of the asymmetry coefficient Δ , and the results are shown in Fig. 3.

The threshold pressure seems to follow well the pattern given by Eq. (7), with the same values for P_0 and F' as in Experiment 1, even at low asymmetry. Eq. (4), on the other hand, is unable to reproduce the data. A possible interpretation is that the vocal folds are slightly coupled and still act as independent oscillators in this region. The points where both curves intersect mark the limits between the theoretical low and high asymmetry regions.

The oscillation frequency is well modeled by Eq. (6), within the low asymmetry region, with $\omega_r = 2\pi f_2$, where $f_2 = 231$ Hz is the second natural frequency at 6.0 kPa of internal pressure [6]. In the high asymmetry region, Eq. (8) produces two values, one for each fold. At positive asymmetry, the measured value seems to be the lower of both frequencies, whereas at negative asymmetry, it seems to be the average.

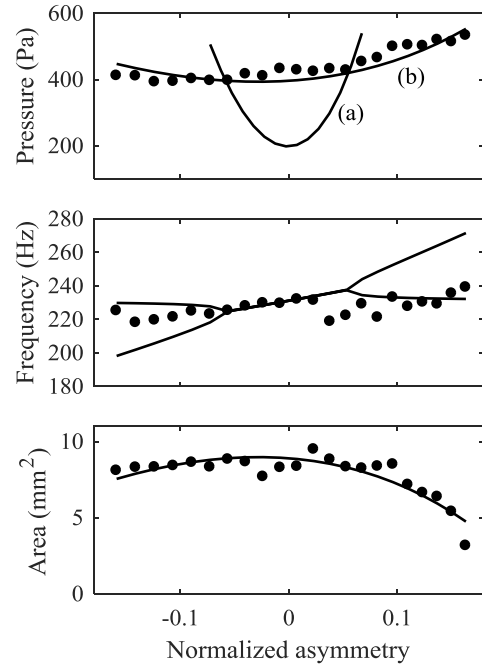


Fig. 3. Results of Experiment 2. Top: oscillation threshold pressure, middle: oscillation frequency, bottom: glottal area at rest. The solid lines show the fitted approximations. In the top panel, curve (a) is produced by Eq. (4) and curve (b) by Eq. (7).

The glottal area may be fitted similarly as in Experiment 1, with the quadratic equation $a_0 = -0.60h^2 + 6.80h - 10.3$.

C. Experiment 3

In this experiment, the internal pressure of one fold was fixed at 2.5 kPa whereas the internal pressure of the other was varied between 2.5 kPa and 5.4 kPa. The intention was to obtain larger values of the asymmetry coefficient Δ , and the results are shown in Fig. 4.

For the threshold pressure, again Eq. (7) provides a better approximation than Eq. (4), although values for $\Delta > 0.15$ seem to follow a different variation pattern.

The oscillation frequency is modeled by Eqs. (6) and (8), as in Experiment 2, but this time using $\omega_r = 2\pi f$, with $f = 95$ Hz. This value is a bit higher than the first natural frequency $f_1 = 80$ Hz measured at 2.5 kPa in the frequency response experiment. With this value, the low asymmetry region is well fitted, and in the high asymmetry region the measured values match the higher of the two theoretical frequencies.

The glottal area may be fitted as in the previous experiments, with the quadratic equation $a_0 = -0.26h^2 + 1.94h - 16.4$.

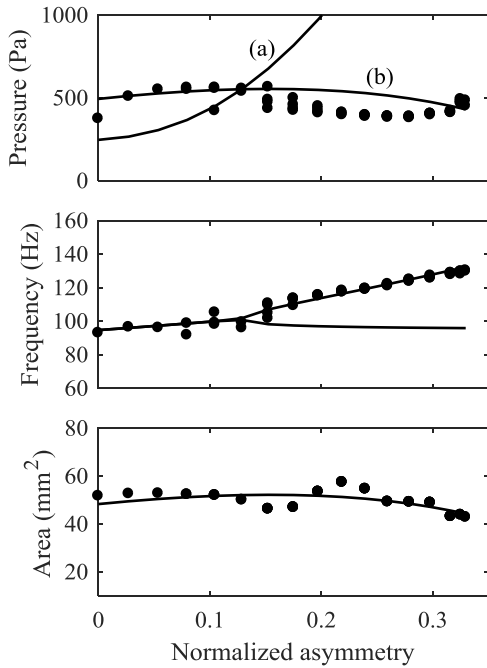


Fig. 4. Results of Experiment 3. Top: oscillation threshold pressure, middle: oscillation frequency, bottom: glottal area at rest. The solid lines show the fitted approximations. In the top panel, curve (a) is produced by Eq. (4) and curve (b) by Eq. (7).

IV. DISCUSSION AND CONCLUSIONS

The subglottal threshold pressure of the vocal fold oscillation has been interpreted as a measure of ease of phonation and proposed as a diagnostic parameter for vocal health [3]. Therefore, it is important to understand its behavior, particularly in abnormal laryngeal configurations.

Our results show a complex pattern for this parameter. According to the adopted theoretical model, distinct behaviors at low vs. high asymmetry are related to the synchronization vs. desynchronization of the vocal fold oscillators. At large asymmetries and low coupling, the vocal folds act as independent oscillators at their own frequency and constant oscillation threshold pressure. If the asymmetry is reduced or the coupling (pressure) increased, then the synchronized action of the vocal fold facilitates the oscillation and

the threshold pressure is reduced up to the minimum at full symmetry.

The model shows agreement with the collected data, except for curves (a) in Figs. 3 and 4. A difficulty when comparing the theory with the experiments is that variations of the internal pressure of the vocal folds affect not only their natural frequency but also their volume, introducing variations in both their separation and oscillating mass which are not contemplated by the theoretical model.

ACKNOWLEDGMENTS

This work was done while Jorge C. Lucero was a visiting researcher of Grenoble Alpes University at LEGI/CNRS. Jorge C. Lucero was also supported by CNPq (Brazil).

REFERENCES

- [1] P. Mergell, P., H. Herzel, and I. R. Titze, "Irregular vocal-fold vibration – High-speed observation and modeling," *J. Acoust. Soc. Am.*, vol. 108, pp. 2996-3002, 2000.
- [2] J. Lucero, J. Schoentgen, J. Haas, P. Luizard, and X. Pelorson, "Self-entrainment of the right and left vocal fold oscillators," *J. Acoust. Soc. Am.*, vol. 137, pp. 2036-2046, 2015.
- [3] I. R. Titze, S. S. Schmidt, and M. Titze, "Phonation threshold pressure in a physical model of the vocal fold mucosa," *J. Acoust. Soc. Am.*, vol. 97, pp. 3080-3084, 1995.
- [4] L. P. Fulcher and R. C. Scherer, "Phonation threshold pressure: Comparison of calculations and measurements taken with physical models of the vocal fold mucosa," *J. Acoust. Soc. Am.*, vol. 130, pp. 1597-1605, 2011.
- [5] J. Haas, P. Luizard, X. Pelorson, and J. C. Lucero, "Study of the effect of a moderate asymmetry on a replica of the vocal folds," *Acta Acust. Acust.*, vol. 102, pp. 230-239, 2016.
- [6] P. Luizard and X. Pelorson, "Threshold of oscillation of a vocal fold replica with unilateral surface growths," *J. Acoust. Soc. Am.*, vol. 141, pp. 3050-3058, 2017.

PHYSICAL STUDY OF THE INFLUENCE OF LEFT-RIGHT VOCAL FOLDS ANGULAR ASYMMETRY ON PHONATION

A. Van Hirtum¹, A. Bouvet¹, X. Pelorson¹, I. Tokuda²

¹ LEGI, UMR CNRS 5519, Grenoble Alpes University, Grenoble, France

² Dep. Mech. Eng., Ritsumeikan University, Kyoto, Japan

annemie.vanhirtum@univ-grenoble-alpes.fr, anne.bouvet@univ-grenoble-alpes.fr, xavier.pelorson@univ-grenoble-alpes.fr, isao@fc.ritsumei.ac.jp

Abstract: Dysphonia is often due to a level difference induced by left-right vocal folds angular asymmetry. In the case of unilateral vocal fold paralysis, angular asymmetry can occur as the normal vocal fold is positioned in the transverse plane whereas the paralyzed vocal fold is rotated in the sagittal plane as its posterior edge is moved in the superior direction. The effect of angular asymmetry (up to 25 degrees) between the left and right vocal fold on the auto-oscillation is experimentally studied using mechanical replicas. For all replicas, it is observed that as full vocal folds contact ceases, increasing angular asymmetry induces a decrease of the signal-to-noise ratio, an increase of the total harmonic distortion rate and an increase of the oscillation threshold pressure. These general tendencies are in agreement with clinical findings reported for vertical level difference during phonation. The increase of oscillation threshold pressure is in accordance with a physical study on level difference by spacing parallel vocal folds so that an analogy is proposed relating the oscillation threshold pressure to angular asymmetry.

Keywords: angular level difference, VF replicas, unilateral vocal fold paralysis, upstream pressure time series analysis

I. INTRODUCTION

Dysphonia is often due to a vertical level difference (up to approximately 3 mm [1]). In the case of unilateral vocal fold paralysis (UVPF), angular asymmetry arises as the normal vocal fold (VF) is positioned in the transverse plane whereas the paralyzed vocal fold is rotated in the sagittal plane so that its posterior edge is moved in the superior direction while its anterior edge remains fixed.

Recently [2], the impact of level difference (LD) between parallel VF's, hereafter referred to as parallel LD, was investigated but VF tilting (the paralyzed VF) was neglected. Therefore, in this work, the influence of tilting of a single VF is studied, i.e. angular LD. It is aimed to systematically study the potential influence of

asymmetry angle α on the auto-oscillation of mechanical replicas. It is noted that for other VF properties, such as tension and shape, left-right VF symmetry conditions are maintained.

II. METHODS

VF replicas: Three deformable silicone VF replicas are used, labeled M5, MRI and EPI. Replicas consist of a succession of molding layers. Concretely, two (M5), three (MRI) and four (EPI) layers are assessed. Each layer aims to reproduce the properties (Young's modulus, thickness) of the different VF constituents: the thyroid muscle (vocalis), the superficial layer of the lamina propria (Reinke's space), the underlying ligament of the lamina propria without tension and the epithelial layer. Besides the structural complexity, replicas are casted using two different geometric molds. Replicas are illustrated in Fig. 1.

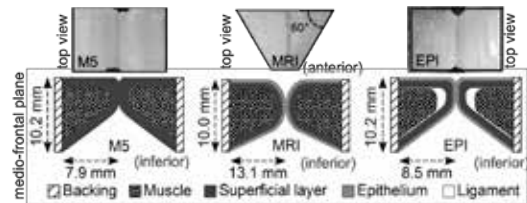


Fig. 1: illustration of VF replicas ($\alpha=0^\circ$).

Mechanical resonances of VF replicas: Mechanical resonance properties of VF replicas are assessed for angular symmetry ($\alpha=0^\circ$) from frequency response functions between an excitation (shaker fed with sinus from 80 Hz up to 350 Hz) and its response (modulation of light passing through the glottal aperture). For all replicas, two mechanical resonances are detected around 143 Hz and around 263 Hz.

Imposing angular asymmetry: The normal VF is fixed in the transverse plane whereas the paralyzed VF is rotated so that its posterior edge is lifted in the superior direction resulting in right-left VF asymmetry angle α as depicted in Fig. 2. A total of 13 different

asymmetry angles are assessed ranging from angular symmetry ($\alpha=0^\circ$) up to $\alpha=24.6^\circ$. This produces an angular LD up to 7.5 mm, which is twice the value reported during phonation [1]. Three VF contact regimes occur: total contact (regime I), partial contact (regime II) for medium α and no contact (regime III) for large α .

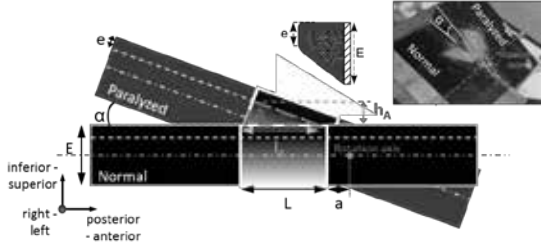


Fig. 2: angular level difference.

Fluid-structure interaction experiments: Flow is supplied to the glottal replicas and upstream pressure P_u is sampled at 10 kHz. The effect of angular asymmetry on the fluid-structure interaction and auto-oscillation is characterized from P_u analysis. Upstream pressures associated with auto-oscillation onset (P_{On}) and offset (P_{Off}) are sought as are spectral features: oscillation frequency (f_0), second (f_1) and third (f_2) harmonics, signal-to-noise ratio SNR and total harmonic distortion rate THD within a steady state portion. A typical $P_u(t)$ time series is plotted in fig. 3.

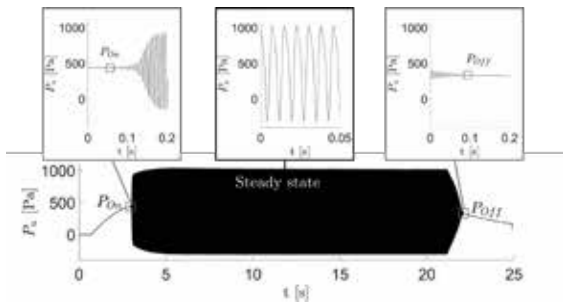


Fig. 3: example of time series $P_u(t)$ (EPI, $\alpha=0^\circ$).

III. RESULTS

A. Geometric characterization of gap for angular LD

Geometrical parameters, which do not depend on imposed asymmetry angle α , are indicated in Fig. 2: minimum (e) and maximum (E) VF thickness, VF length (L) and rotation axis position (a). Geometrical reasoning allows a geometric characterization of angular LD as a function of asymmetry angle α for triangular gap area $A(\alpha)$, base of the triangular gap $l_A(\alpha)$ and hence degree of

VF contact $G(\alpha) = 1 - l_A(\alpha)/L$ as well as triangular gap height $h_A(\alpha)$. Triangular gap base $l_A(\alpha)$ increases with α so that $G(\alpha)$ yields 1 for small α (contact regime I: full contact), $0 < G(\alpha) < 1$ for medium α (contact regime II: partial contact) and 0 for large α (contact regime III: no contact). Two critical angles are defined indicating the shift from contact regime I to II ($\alpha_{I,II}$) and from II to III ($\alpha_{II,III}$), respectively. The degree of contact $G(\alpha)$ and glottal gap area $A(\alpha)$ for assessed asymmetry angles α is shown in Fig. 3. As the e -value for EPI yields about $\frac{1}{3}$ of the e -value for M5 and MRI, plotted curves show the influence of e on the gap geometry. Indeed curves of M5 and MRI are in close agreement, whereas the gap for EPI increases more rapidly. It is noted that contact regime III is only reached for the EPI replica.

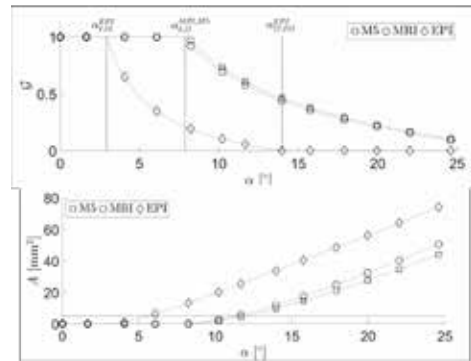


Fig. 4: degree of contact $G(\alpha)$ and gap area $A(\alpha)$. Contact regimes (I, II and III) and critical angles are indicated for all replicas.

B. Fluid-structure interaction with angular LD

Oscillation threshold pressures: Detected auto-oscillation onset $P_{On}(\alpha)$ and offset $P_{Off}(\alpha)$ threshold pressures are illustrated in Fig. 5 for the EPI and MRI replicas. Onset pressure minima $\min(P_{On}(\alpha))$ occur for $\alpha > 0^\circ$: within regime I (full contact, $G(\alpha)=1$) for the MRI replica and within regime II (partial contact, $0 < G(\alpha) < 1$) for the M5 and EPI replicas. Nevertheless, the decrease is limited to less than 13% for all replicas.

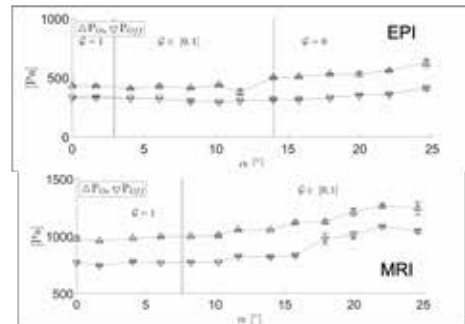


Fig. 5: onset $P_{On}(\alpha)$ and offset $P_{Off}(\alpha)$ pressures.

Onset pressure maxima, $\max(P_{On}(\alpha))$, occur near the largest assessed asymmetry angle $\alpha \approx 24.6^\circ$ and the relative increase yields at least 29.5% for all replicas. For all replicas, the onset pressure increases for $\alpha \geq 15^\circ$. For $\alpha < 15^\circ$ the shape of the $P_{On}(\alpha)$ curves varies between replicas so that it depends on the detailed geometry and layer composition of each replica. It is observed that more realistic replica's (MRI and EPI) exhibit a more monotonous tendency than the M5 replica. For all replicas hysteresis between onset and offset pressures is observed as expected. The degree of hysteresis is only marginally affected by α in contact regimes I and II (all replicas: M5, MRI and EPI), whereas it is seen to be reinforced in contact regime III (only EPI replica).

Harmonic oscillation frequencies: Auto-oscillation frequencies are illustrated in Fig. 6 for the EPI and MRI replicas. For angular symmetry ($\alpha=0^\circ$) detected oscillation frequencies f_0 and f_1 are in close approximation with the first and second mechanical resonance frequencies. For $\alpha > 0^\circ$, fundamental frequency f_0 decreases monotonously with α . The overall decrease of f_0 relative to its value for $\alpha=0^\circ$ is negligible (less than 3%) in contact regime I (full contact) and becomes more pronounced (>10%) in regimes II (partial contact) and III (no contact). Second harmonics f_1 are detected for all replicas and all asymmetry angles α in regimes I and II. For the EPI replica a third harmonic f_2 is detected for all α in regime III, indicating that the harmonic contents enriches with α .

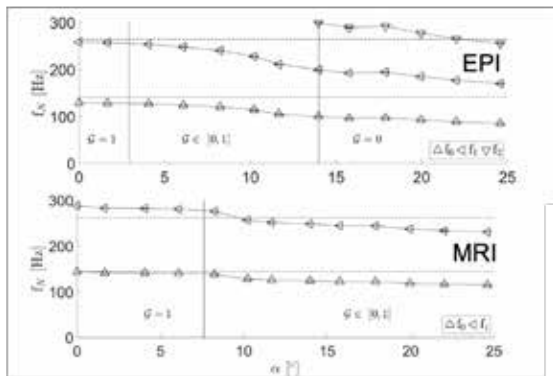


Fig. 6: oscillation frequencies. Mechanical resonances are indicated (horizontal dashed lines).

Global spectral features THD and SNR: That harmonics grow with α is confirmed considering the total harmonic distortion rate (THD) plotted in Fig. 7.

THD increases for all replicas with α . The increase is most prominent in contact regime II ($0 < G < 1$) as it yields about 10 dB for the EPI replica and about 25 dB for the M5 and MRI replicas. The signal-to-noise ratio (SNR) is plotted as a function of α in Fig. 7 as well. It is observed for all replicas that increased THD is associated with a reduced signal-to-noise (SNR) ratio with about 15~dB.

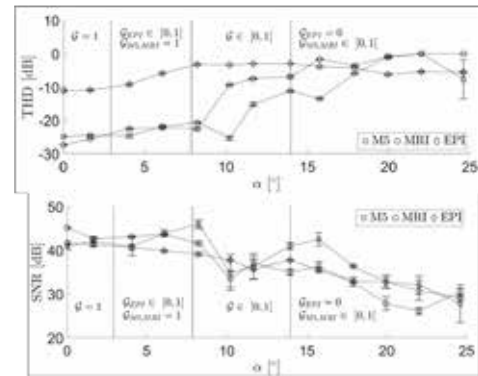


Fig. 7: total harmonic distortion rate THD and signal-to-noise ratio SNR as a function of α .

Parallel LD analogy from P_{on} : In [2] a parallel level difference was imposed by varying the distance (dE) in the inferior-superior direction between parallel transverse planes containing the upper surfaces of each VF. The glottal leakage due to parallel level difference is thus rectangular while it is triangular in the current study on angular level difference. Consequently, when imposing parallel LD only regime I (full contact) and regime III (no contact) occurs, whereas also partial contact (regime II) happens when imposing asymmetry angle α .

Some major tendencies associated with increasing asymmetry angle α , such as decreasing fundamental frequency f_0 and overall increasing onset pressure P_{on} , were also reported when increasing dE [2]. Therefore, it is sought to propose a parallel level difference analogy aiming to relate angular level difference in terms of parallel level difference, *i.e.* to establish the relationship $dE(\alpha)$. A-priori, different approximations of the sought relationship can be considered: $dE(\alpha) \approx h_A(\alpha)$, $dE(\alpha) \approx h_L(\alpha)$ with $h_L(\alpha) = A(\alpha)/L$ and $dE(\alpha) \approx h_{LA}(\alpha)$. These approximations yield maxima for $\alpha = 24.6^\circ$, which are within the dE-range (up to 7.5 mm) assessed in [2]. Moreover, both h_L and h_{LA} are of the same magnitude (up to 3~mm) as vertical level differences reported during human phonation [1].

A theoretical expression for oscillation onset threshold pressure P_{on} as a function of parallel level difference dE was derived in [2] based upon the small amplitude approximation of the vocal folds oscillation

and neglecting potential sub- and supra-glottal acoustic coupling. Fitting of the theoretical expression using parallel level difference approximation $dE(\alpha) \approx h_{IA}(\alpha)$ resulted in the best fit accuracy (coefficient of determination $R^2 > 81\%$) for all replicas. It follows that $dE(\alpha) \approx h_{IA}(\alpha)$ results in the sought analogy. Measured and fitted $P_{on}(h_{IA}(\alpha))$ are shown in Fig. 8.

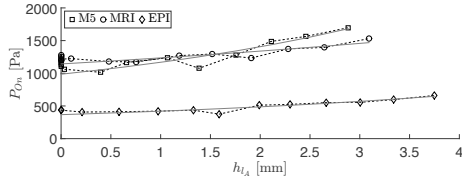


Fig. 8: Measured and fitted $P_{on}(h_{IA}(\alpha))$.

IV. DISCUSSION

Characterized auto-oscillation features for all three replicas show that angular asymmetry within contact regime I (full contact) remains without consequences as onset pressures P_{on} , oscillation frequencies f_0 , THD and SNR are in close approximation to values observed for angular symmetry at $\alpha=0^\circ$. In contact regime II, up to $\alpha \leq 15^\circ$ voice quality parameters as THD and SNR start to deteriorate as THD increases and SNR decreases. For $\alpha > 15^\circ$ all extracted voice parameters are affected.

Some of the described observations relate with medical findings reported on human subjects in case of UVFP so that the studied angular asymmetry provides a potential explanation. Reduced SNR due to air leakage in contact regime II and III through the glottal gap is consistent with a breathy voice description. Increased oscillation onset pressure P_{on} in contact regimes II and III is consistent with vocal fatigue. The decrease of fundamental frequency f_0 on the other hand does not follow the unnatural high-pitched voice associated with phonation in the case of UVFP. This might be due to several reasons such as the increased contribution of observed higher harmonics on the perceived pitch, mechanical property changes to the paralyzed vocal fold which are not assessed in the current work or compensatory strategies involving supra-laryngeal structures by human speakers. As auto-oscillation does not cease for the assessed range of asymmetry angles, it follows that varying α within the assessed range could not reproduce aphonia. Finally, it is noted that α -ranges for which $h_A \leq 3$ mm, which is the typical range of vertical LD reported during human

phonation [1], extend to within contact regime II for all three replicas: $\alpha \leq 13^\circ$ for EPI and $\alpha \leq 17^\circ$ for both EPI and MRI. Therefore, partial contact (regime II) is of particular importance for phonation. Nevertheless, as few measurements of vertical level difference during phonation in case of UVFP are reported in literature, observed features for the assessed range of α -values provide a useful reference for further studies. Moreover, as glottal imaging technologies continue to improve so that *in-vivo* measurement of VF geometrical parameters might become possible, derived geometrical expressions $A(\alpha)$, $l_A(\alpha)$, *etc.* can be applied to estimate geometrical features and to characterize VF contact (G and regime I, II or III) during phonation in the case of UVFP.

V. CONCLUSION

An experimental study is presented assessing the effect of angular level difference on the auto-oscillation for three mechanical VF replicas. Spectral oscillation features (harmonic frequencies, THD and SNR) and upstream threshold pressures (oscillation onset and offset) are analyzed. The same tendencies were observed for all three replicas. Increasing asymmetry angle α so that VF's are no longer in full contact (contact regimes II and III) alters observed features: SNR decreases, THD increases, oscillation threshold pressures increase, fundamental frequencies decrease and higher harmonics emerge. Geometrical details of each VF replica determine the leakage area and critical asymmetry angles associated with a shift in contact regime. Expressions are derived to quantify these features from geometrical VF parameters. Apart from the decrease in fundamental frequency, observed tendencies are in agreement with those reported in clinical studies on vertical level difference so that the same geometrical reasoning and expressions might be applied to characterize the glottal gap during human phonation in the case of UVFP. A parallel level difference analogy relating angular and parallel level difference is proposed.

REFERENCES

- [1] Y. Oyamada, E. Yumoto, K. Nakano, H. Goto, "Asymmetry of the vocal folds in patients with vocal fold immobility," *Arch Otolaryngol head neck surg*, vol. 131 pp. 399-406, 2004.
- [2] I. Tokuda, R. Shimamura, "Effect of level difference between left and right vocal folds on phonation: Physical experiment and theoretical study," *J Acoust Soc Am*, vol. 142 pp. 482-492, 2017.

THEORETICAL AND EXPERIMENTAL MODELING OF LESIONS OF THE VOCAL FOLDS

X. Pelorson, A. Bouvet, A. Van Hirtum

LEGI, UMR CNRS 5519, Grenoble Alpes University, Grenoble, France
xavier.pelorson@legi.cnrs.fr, anne.bouvet@legi.cnrs.fr, annemie.vanhirtum@legi.cnrs.fr

Abstract: In this study a simple physical model of the vocal folds auto-oscillation during speech is tested against experimental data obtained on a replica of the larynx. It is shown that all parameters except one can be directly measured on the replica. Examples of results corresponding to two different pathological configurations are shown and discussed. To a certain extent, it appears that the theoretical model allows to explain, at least qualitatively, the experimental data.

Keywords: Vocal folds, mechanical replica, pathology.

I. INTRODUCTION

The vocal folds at the larynx constitute a biomechanical oscillator that acts as the main sound source in voicing. Under appropriate conditions, the air flow that passes through the glottis induces their self-sustained oscillation. This oscillation, in turn, modulates the airflow which causes the generation of the acoustical waves that, after they propagated inside the vocal tract and radiate at the lips, we perceive as voice.

Some frequent pathologies, especially amongst teachers and singers, are due to an alteration of the vocal folds tissues. This can happen, for instance, in the presence of a localized growth at the surface of one, or both, vocal fold (cysts, nodules, polyps) or of larger alterations (vocal fold scar, sulcus vocalis). Depending on the severity of the pathology, the consequences range from limited voice disorders to aphonia [1].

Physical modeling is of interest for a better understanding of the mechanisms underlying such vocal folds pathologies and to provide tools for analysis and diagnostic, for the prediction of surgery events or for the design of vocal folds prosthesis.

A theoretical model of phonation will first be presented. It consists of a biomechanical description of the vocal folds elasticity in a modal approach coupled with a quasi-steady viscous flow model for the airflow and with a linear 1-D description of the acoustics of the

vocal tract. This theoretical model can then modified in order to account for each pathology of the vocal folds in the most plausible physiological way.

The outcomes of the theoretical model are compared with experimental results obtained on a mechanical replica of the larynx and of the vocal tract.

II. METHODS

Theoretical model

The theoretical model is inspired by the work of Avanzini [2]. The geometry of the vocal folds is controlled by two mass points, one at the entrance and one at the exit of the glottis as shown in Fig 1.

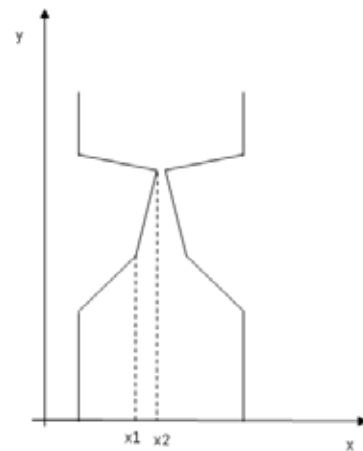


Fig. 1 : Definition of the vocal folds geometry

In the z-direction, the width of the vocal folds, L_g , is assumed to be constant.

The motion of the entrance point, x_1 is predicted using a second order differential equation:

$$\frac{d^2 x_1}{dt^2} + (1 + \gamma)2\pi\Delta f \frac{dx_1}{dt} + (2\pi f)^2 x_1 = \frac{F}{m}$$

where f and Δf are respectively the vocal fold mechanical resonance frequency and its bandwidth, m

is the vibrating mass of the vocal folds and γ the contact parameter defined by :

$$\gamma = L_c/L_g$$

where L_c is the length of the eventual contact between the two vocal folds. Note that $\gamma = 1$ (full contact between the vocal folds) corresponds to the critical damping value.

The motion of the point at the outlet of the glottis, x_2 , is assumed identical to x_1 but delayed by a time constant τ :

$$x_2(t) = x_1(t-\tau)$$

The force, F , generated by the airflow on the vocal folds is calculated under the assumptions of quasi steady incompressible flow accounting for viscous losses as a corrective term [3].

Lastly, acoustical coupling with the sub- and supra-glottal cavities is neglected in this study.

Experimental Set-up

The set-up is illustrated in Fig. 2.

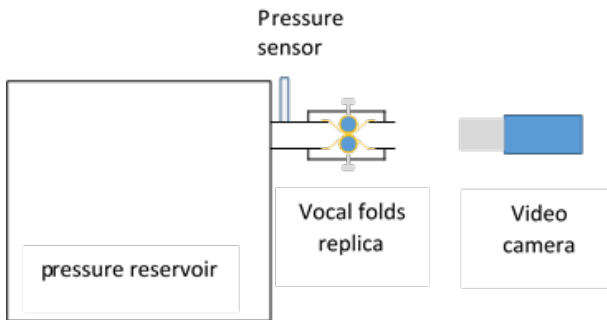


Fig. 2 : Schematic view of the experimental set-up.

It consists of a large air pressure reservoir filled with acoustical foam, in order to damp acoustical resonances. An up-scaled (by a factor 3:1) vocal tract replica is connected to the pressure reservoir using a 10 cm long uniform tube. The displacement of the vocal folds replica is recorded using a high speed video camera (Mikrotron, EoSens Cube7). After calibration, using calibrated optical grids, this allows for a measurement of the glottal area.

A pressure sensor (Kulite XCS093) is mounted upstream of the replica. Calibration of the pressure sensor is made again a water meter with an accuracy of +/- 5 Pa.

The vocal folds replica consists of latex tubes (internal diameter 11mm and length 80 mm) filled with water under controllable pressure. Changing the water pressure inside the latex tubes allows to modify the stiffness of the replicas but also modify the glottal area at rest because as the latex tubes are inflating or deflating. A latex tube passes between each vocal folds replica and connects the upstream region to the downstream one, as shown on Fig. 3.

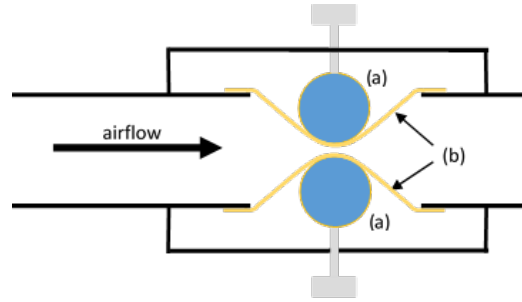


Fig. 3 : Vocal folds replica. (a) latex tubes filled with water, (b) connecting latex tube.

This experimental set-up allows to measure the parameters needed to feed the theoretical model :

- the geometrical dimensions (length, width, thickness, initial glottal area) of the vocal folds,
- the mechanical resonance frequency, f , and its bandwidth, Δf , measured with the technique described in [4],
- the delay τ , can be estimated using high speed video imaging by tracking points on the surface of the vocal folds replica

The only parameter that cannot be directly measured is the vibrating mass, m .

III. RESULTS

A. Example 1 : localized perturbation

In order to simulate a localized affection of one vocal fold, such as a nodule, sphere of led (diameter 2.75 mm, mass 0.1 g) is inserted between the connecting latex tube ((b) in Fig 3) and one vocal fold replica ((a) in Fig. 3). The threshold pressure, defined as the minimum air pressure needed to obtain a self-sustained oscillation of the vocal folds, is measured as a function of the water pressure inside the vocal folds. An example of measurement results is shown in Fig. 4. The theoretical predictions are also displayed on the same figure. In the calculations, it was assumed that the mass, m was 20 % of the total mass of the replica.

Other examples of results can be found in [5].

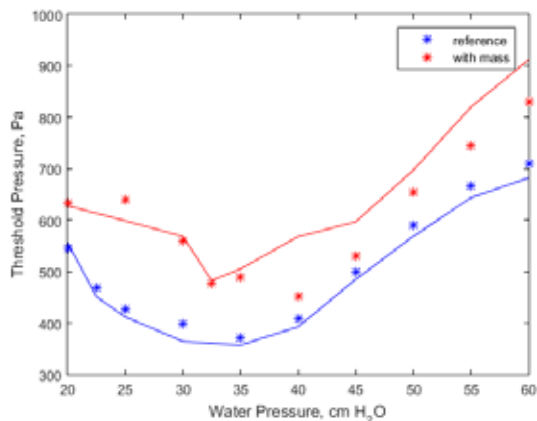


Fig. 4 : Threshold of oscillation as a function of the water pressure inside the vocal folds replica. In blue : reference case (without perturbation), in red with added mass. * are measurements points, solid line is the theoretical prediction.

From Fig 4, one can observe that the addition of a mass at the surface of one vocal fold increases significantly (of order of 15 %) the threshold pressure. The measured fundamental frequency of oscillation is not affected by the presence of the mass. This can be understood by the fact that the added mass is much smaller than the mass of the replica.

The theoretical prediction for the reference case, without added mass, fits well the experimental data thanks to the particular choice of the mass, m . Using the same value for m and accounting for the added mass, the theoretical curve predicts the increase of the threshold pressure. The mean RMS deviation with the experimental data is of order of 10 %.

B. Example 2. Paralyzed vocal fold

In this example, a complete paralysis of one vocal fold is experimentally simulated by covering one vocal fold replica with a metal sheet preventing thus any oscillation. As for the previous example, the reference configuration, without the metal sheet, is compared with the experimental data. An example of results is presented in Fig 5.

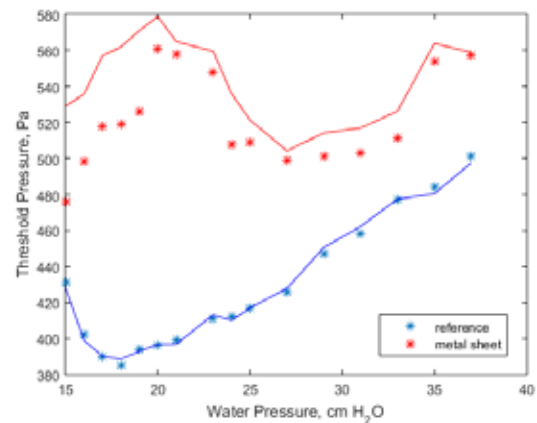


Fig. 5 : Threshold of oscillation as a function of the water pressure inside the vocal folds replica. In blue : reference case (without perturbation), in red with metal sheet. * are measurements points, solid line is the theoretical prediction.

In the reference configuration, without perturbation, the threshold pressure measurement exhibits the same kind of concave shape observed in Fig. 4. Preventing the oscillation of one vocal fold by adding a metallic sheet significantly increases the threshold pressure (of order of 15 %) but also modifies the shape of the curve. This is particularly the case for the lowest values of the water pressure (between 15 and 20 cm H₂O). This effect might be due to the fact that the presence of the metallic sheet not only prevents oscillation of one vocal fold but also increases the initial glottal area.

The theoretical predictions agree qualitatively with the experimental data. Quantitatively, large discrepancies appear especially for low water pressure configurations. The presence of turbulence of the flow, which is not accounted in the theoretical model, might be an explanation for these departures.

IV. DISCUSSION & CONCLUSIONS

The theoretical model developed in this paper has the advantage of relying on a limited number of parameters. Compared with the popular model of Ishizaka and Flanagan [6] the present description requires 4 times less parameters. Further, all these parameters are explicit and all, but one, can be directly measured on a replica of the larynx.

By definition, this theoretical model only accounts for a single mechanical mode and thus is not suitable for the prediction of frequency jumps which are frequent in voice pathology observations. The

extension to include more mechanical modes is under development.

Despite its simplicity and its limits, the theoretical model shows however good prediction results when compared with experimental data obtained on a mechanical replica of the larynx. This conclusion was only illustrated by two relatively simple configurations : a small localized mass addition and a total paralysis, and needs to be confirmed, or not, using a wider case of configurations.

Acknowledgments : This work was partially funded by the ArtSpeech project (ANR-15-CE23-0024). The help of Mohammad Ahmad for the experimental data is acknowledged.

REFERENCES

- [1] K. Verdolini, C. A. Rosen, and R. C. Branski, *Classification Manual for Voice Disorders—I* (Psychology Press, New York, 2014)
- [2] F. Avanzini, "Simulation of vocal fold oscillation with a pseudo-one-mass physical model," *Speech Commun.*, 50, 95–108. 2008.
- [3] J. Haas, P. Luizard, X. Pelorson, and J. C. Lucero, "Study of the effect of a moderate asymmetry on a replica of the vocal folds," *Acta Acust. Acust.* 102(2), 230–239. 2016
- [4] N. Ruty, X. Pelorson, A. Van Hirtum, I. Lopez-Arteaga, and A. Hirschberg, "An in vitro setup to test the relevance and the accuracy of low-order vocal folds models," *J. Acoust. Soc. Am.* 121(1), 479–490. 2007
- [5] Luizard, P. and Pelorson, X. "Threshold of oscillation of a vocal fold replica with unilateral surface growths". *The Journal of the Acoustical Society of America* , 141(5), 3050–3058.2017
- [6] K. Ishizaka and J. L. Flanagan, "Synthesis of voice sounds from a two-mass model of the vocal cords," *Bell Sys. Tech. J.* 51, 1233–1268. 1972.

VOICE SOURCE EFFECTS OF FUNDAMENTAL FREQUENCY VARIATION

J Sundberg^{1,2}

¹ Department of Speech Music Hearing, School of Electrical Engineering and Computer Science, KTH, Stockholm,
² University College of Music Education Stockholm Sweden

Abstract: Introduction

Model work by Titze and associates has shown that vocal tract resonances, or formants, can interact with the voice source, i.e., the pulsating glottal airflow. The effect occurs when a formant is close in frequency to a harmonic partial. When the formant is just below a partial its amplitude is increased and when it is just above the partial its amplitude is decreased. The paper will present results of attempts to document such effects by measurements.

Methods

Audio signal was recorded by omnidirectional microphone a few cm from the side of the lip opening of trained singer voices phonating glissando tones on a constant vowel as in normal singing and with artificial dampening of the formants by a piece of cotton in front of the lip opening. The signal was inverse filtered using the Sopran software and the voice source waveform and spectrum were analysed.

Results

Preliminary results show that the amplitude of a source spectrum partial is attenuated in the vicinity of the first formant. The effect is reduced when the formant is damped.

Conclusions

Interaction between the first formant and the voice source can affect the spectrum of the voice source and is dependent on the damping of the formant.

Keywords: Formant, voice source, spectrum

The relationship between the derivative of the flow glottogram and its spectrum was analysed in terms of a waveform model by Fant and associates [1]. The model produces a spectrum characterized by a smooth envelope sloping between -12 and -18 dB/octave depending on the maximum flow declination rate, a parameter controlled mainly by the subglottal pressure. Using computerized models, Titze and associates have analysed the consequences of the interaction between the glottal voice source and the vocal tract filter. The model predicts that in an ascending pitch glide, the interaction increases the amplitude of a spectrum partial when its frequency is just below the frequency of a formant, the reason being that the reactance is inductive in this frequency range. By contrast, a partial's amplitude should decrease above the formant frequency, because here the reactance is there compliant. Titze and associates analyzed the radiated spectrum of pitch glides performed when the subjects' vocal tracts were lengthened with hard-walled tubes of lengths between 5 and 19 cm [2]. Tracking of the amplitude of individual partials through the glide showed an asymmetric change of the level of the partials that passed a formant in a pitch glide.

This phenomenon would be caused by the vocal tract resonances affecting the glottal airflow. Hence, it seemed worthwhile to analyze these effects not only in the radiated spectrum, but also in the voice source itself. The aim of the present study was to analyze how the level of single voice source spectrum partials was affected when passing a formant.

I. INTRODUCTION

The pulsating glottal airflow, generated when the vocal folds chop the airstream from the lungs, is a clinically and pedagogically highly relevant aspect of voice production. Voice disorders are typically caused by a disturbed function of the vocal folds, and efficient control of the acoustic properties of the glottal airflow is crucial to speech production and even more in singing. The glottal airflow during voicing consists of quasi-triangular pulses surrounded by flat segments, see Fig. 1. The pulses and the flat segments correspond to the open and closed phases of the vocal fold vibration cycle, respectively.

II. METHODS

Five trained singers, one female, four male, sang, on the vowel /a/ or /ae/, glide tones covering their pitch ranges. As source filter interaction can be assumed to depend on the damping of the formants, two of the subjects repeated the task with a piece of loose cotton, approximately 6x6x2 cm, in front of and almost touching the lips.

The audio signal was picked up by an OM1 condenser microphone (Line Audio design, Sweden), held near the cheek at a distance of about 5 cm from the center of the lip opening. The audio signals, digitized at 16000 Hz by

a Focusrite Scarlet 2i2 AD converter, were recorded by the Sopran software and stored as wav files in a PC computer.

The recordings were analyzed by the Inverse Filter module of the Sopran software. It displays the input and the filtered waveforms and spectra in quasi-real time. For the manual tuning of the inverse filters two criteria were applied, (i) a closed phase as ripple-free as possible and (ii) a source spectrum envelope as free as possible of local dips and peaks near the formant frequencies. It turned out that the filter settings that yielded a flow glottogram meeting these criteria at one fundamental frequency could be used for the entire glide. The program saves the inverse filtered vowel in a new track of the wav file.

The next step was to analyze the spectrum of the stored voice source signal using the Spectrum module of the Sopran software. The analysis bandwidth was set to 80 Hz yielding a time window of 50 ms. The levels of all partials below 1000Hz were measured and the values stored in log file together with the partial's frequency and the time coordinate.

III. RESULTS

Fig. 2 shows, as a function of time, the variation of sound level for the partials below 1000 Hz for singer L. Level dips up to about 10 dB can be observed in partials 2, 3, 4, and 5, when their frequencies coincided with the frequency of the first formant. Interestingly, the dip of partial 4 at $t=0.876s$ was accompanied by a simultaneous dip in partial 5. Similarly, the dip of partial 5 at $t=0.625s$ was accompanied by a simultaneous dip of partial 6. Figure 3 shows the same data as function of frequency, the vertical dashed representing the frequency of the first formant.

Holding a piece of cotton in front of the lip opening will add resistance to the vocal tract resonator, widening the formant bandwidths. In addition, a minor lowering of the formant frequencies could be expected. An estimation of the magnitudes of these effects could be made by comparing the filter settings used for the inverse filtering of the normal and damped vocal tracts. The cotton was found to lower the first formant by about 10% and to widen its bandwidth by a factor of about 1.6. Figure 4 shows the associated effects observed for partials 3 and 4 in the female singer. When the first formant was damped, the dips in the level curves were about 5 dB less deep and clearly less sharp.

IV. DISCUSSION

This preliminary study has shown clear effects on the voice source of the vocal tract resonance, i.e., of source filter interaction. In this sense, these findings are in agreement with the work of Titze and associates; the amplitudes of source spectrum partials increased just above the first formant. On the other hand, they clearly

decreased with frequency just below the formant. Also, the decrease tended to be symmetrical on both sides of the formant, while the source filter interaction could be expected to produce an increase of the level below the formant and a decrease above it. The reason for this discrepancy between predicted and observed effects should be further studied in the future.

The source-filter interaction should affect the shape of the glottal flow pulses, making them tilt to the right. This should add to the amplitude of the voice source. A trend in this direction could be discerned in terms of an amplitude increase with frequency for many source spectrum partials in the region below the first formant, as illustrated in Fig. 3. However, the trend was reversed to a decrease about 100Hz below the formant; in this region the amplitudes tended to decrease with rising frequency.

On the other hand, several partials failed to recover completely on the other side of the formant. This might be an effect of the reactance shifting from inertive to compliant at the formant frequency.

The experiment was run with singer subjects only. Such subjects are experts in both phonatory and articulatory control. This was advantageous since they were able to keep the articulation and thus the formant frequencies constant throughout the pitch glide. Thus the same inverse filter setting could be used. On the other hand, it is possible that such subjects have learnt to avoid or minimize also the effects of source filter interaction, such that their voice source keeps the same acoustic characteristics independent of combinations of partials and formant frequencies. If so, it would be interesting to understand how this can be achieved.

V. CONCLUSION

Amplitudes of the source spectrum partials tend to decrease with frequency just below the first formant and to increase with frequency just above it. This would be an effect of source filter interaction.

REFERENCES

- [1] G. Fant, J. Liljencrants and Q. Lin, "A four-parameter model of glottal flow," *Speech Transmission Laboratory Quarterly Status and Progress Report 4/85*, KTH, pp. 1-3, 1985.
- [2] L. Maxfield, A. Palaparthi and I. Titze, "New evidence that nonlinear source-filter coupling affects harmonic intensity and fo stability during instances of harmonics crossing formants" *Journal of Voice*, vol. 31: 2 pp. 149-156, 2017.

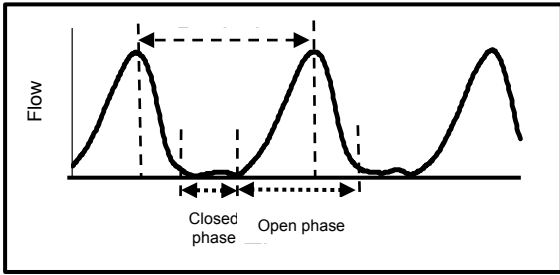


Figure 1. Flow glottogram showing glottal flow versus time.

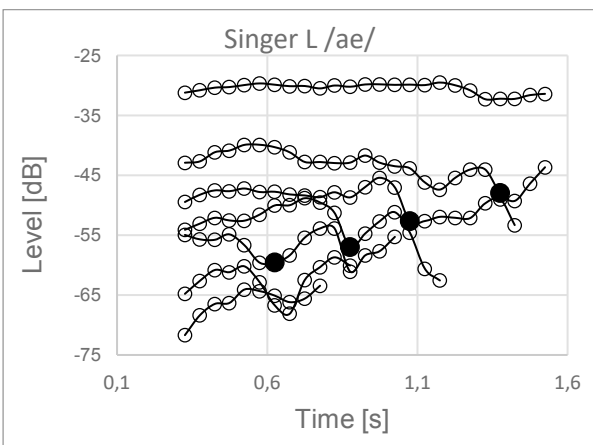


Fig. 2. Variation of sound level for the source spectrum partials below 1000 Hz for singer L. Filled circles show the values observed when the partial coincided with the first formant.

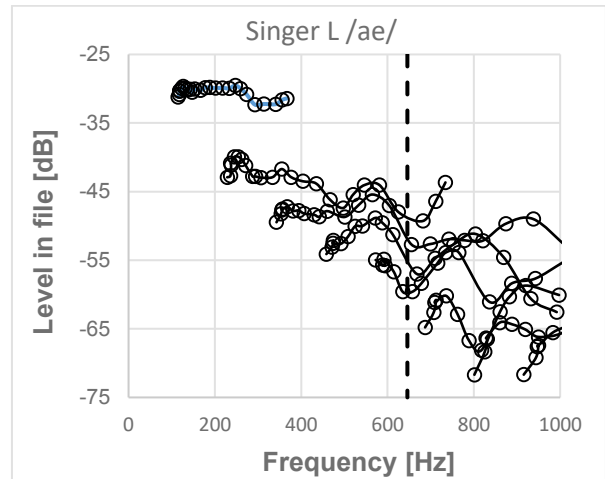


Fig. 3. Levels of the lowest source spectrum partials as function of frequency. The dashed vertical line shows the frequency of the first formant.

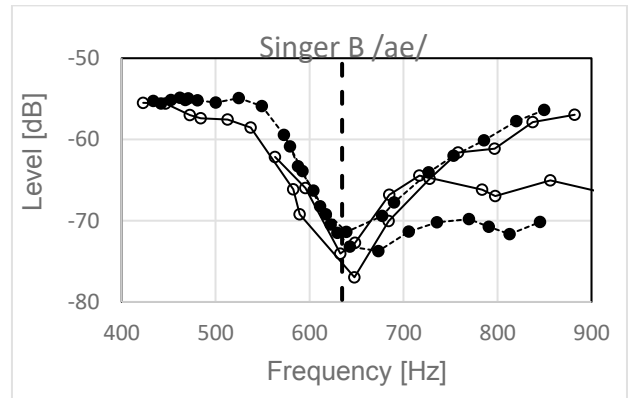


Fig. 4. Variation of sound level for source spectrum partials 3 and 4 for singer B. Open and filled circles refer to phonation without and with a piece of cotton in front of her lip opening. The dashed vertical line represents the frequency of the first formant.

**SESSION VII - COMET SESSION
ACTOR'S AND ACTRESS' VOICES**

SPECTRAL SPECIFICITIES OF ACTING VOICE IN PROFESSIONAL ACTRESSES

P. H. DeJonckere¹, H. Stoffels²

¹ Federal Agency for Occupational Risks, Brussels, Belgium

² Utrecht University, Clinical Speech & Hearing Sciences, Utrecht, The Netherlands
ph.dejonckere@outlook.com

Abstract: Research about actor's and actresses' voices is an extension of research about the singing voice. Male classical singers and (to less extent) actors show, when performing, a typical 'singer's (or actor's) formant', generated by a clustering of formants 3, 4, and 5. Whether actresses demonstrate a similar phenomenon is controversial. Our design compared the voice acoustics of 18 Dutch professional actresses reading a selected dramatic text (from Euripides' Medea) with respectively their actress' voice and a loud 'neutral' voice. Both recordings were acoustically investigated by LTAS and perceptually evaluated. The blinded evaluators could trace - practically without error - which recording was the acting one. Actresses exhibit significantly more spectral energy in the 3-4 kHz part of the LTAS when they are acting. Moreover, the actresses showed significantly more reinforcement between 5-8 kHz in the acting condition.

Keywords: Actress' voice, Actor's formant, LTAS.

I. INTRODUCTION

Research about actor's and actresses' voices is an extension of research about the singing voice. Male classical singers and (to less extent) actors show, when performing, a typical 'singer's formant' [1], generated by a clustering of formants 3, 4, and 5. Similarly, in professional male actors, a formant peak has been found at around 3500 Hz, although the increase in energy was smaller than that found in the singers [2 – 6]. Whether actresses demonstrate a similar phenomenon is not clear.

Master, De Biase and Madureira (2012) [7] compared 30 actresses with 30 non-actresses reading a short text at normal and loud volume. They did not find an actor's formant in these actresses, and concluded by suggesting that - in females - voice projection is likely more related to glottal settings rather than to resonances in the vocal tract.

II. METHODS

In order to control as far as possible biasing factors, our own design compared 18 Dutch experienced professional actresses (average 32,8 years) reading a selected, highly dramatic passage (from Euripides' tragedy Medea, act 1) of about 100 s (244 words) with respectively their acting voice and a loud 'neutral' voice. The subjects had the opportunity to first familiarize with the text (Fig. 1). All actresses were in good vocal health and none had voice complaints. Both digital recordings (sampling frequency 44.1 KHz) of each actress (mouth-microphone distance = 30 cm) were acoustically investigated by computing a LTAS (0 – 10.000 Hz) and were perceptually evaluated by 3 blinded raters. Raters had to define if the recording pertained to an acting voice or to a non-acting voice.



Fig. 1: Experimental setup : Actress reading the Medea-passage.

LTAS is a spectrum where the voiced parts of a continuous recording are averaged instead of measuring only a short sound fragment with a fast Fourier spectrum. The LTAS stabilizes after around 40 seconds [8]. All LTAS curves were normalized, which means that the highest peak of the spectrum was set to 100 % of the y-axis, and 'slices' of 1.000 Hz were created for statistical comparisons of acoustical energy levels. The Alpha-ratio [9], which is a measure of spectral balance, defined as a ratio of energy below and

above 1000 Hz, was also calculated for each LTAS. There are two variants of the Alpha-ratio: the (0 to 1 KHz) / (1 to 10 KHz) variant, and the (0 to 1 KHz) / (1 to 5 KHz) variant. Fig. 2 shows a graphical illustration example of the (0 to 1 KHz) / (1 to 10 KHz) variant.

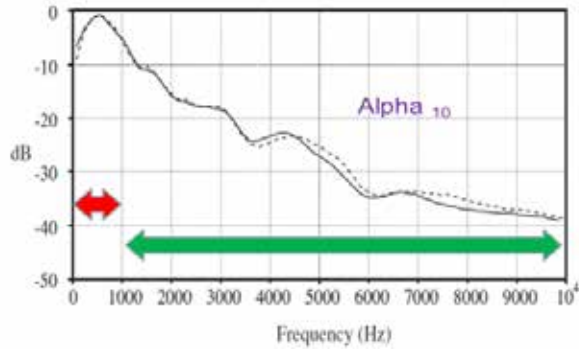


Fig. 2

Fig. 2: Graphical illustration of the (0 to 1 KHz) / (1 to 10 KHz) variant of the Alpha-ratio.

Mean speaking frequency and total duration were computed from the recordings. Intensity was monitored at 1 m. of the lips.

III. RESULTS

A. Perceptual experiment

All 'acting' voices could be correctly identified by two of the raters. The third one failed twice in identifying the acting voice.

B. LTAS

The comparative normalized average LTAS for the two voicing conditions (acting and non-acting) are presented in Fig. 3.

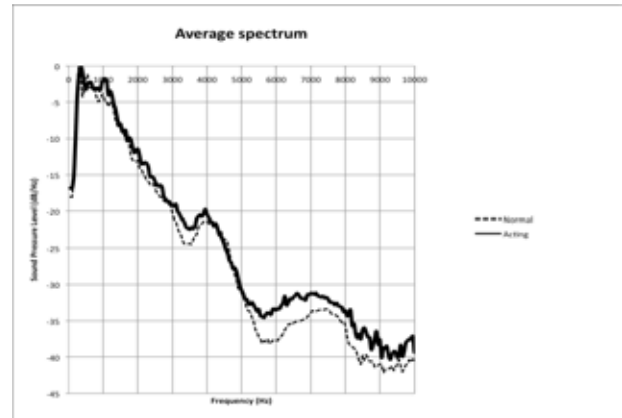


Fig. 3

Fig. 3: Normalized average LTAS of all actresses (n = 18) when reading normally (dotted line) and when reading whilst acting (continuous, bold line).

The differences between the 2 LTAS-curves are shown in Fig. 4.

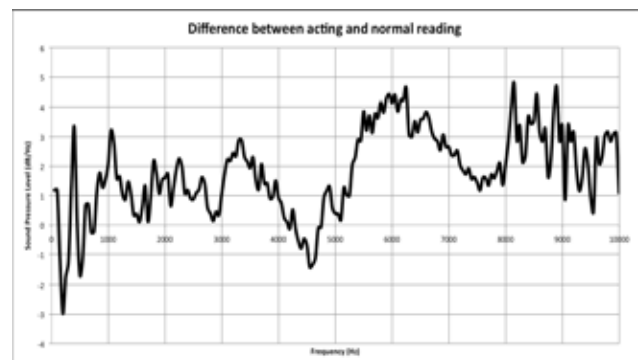


Fig. 4

Fig. 4 Differences between the acting- and the normal condition based on normalized, averaged LTAS from the whole group (n = 18).

A comparison slice by slice (1 KHz per slice) demonstrates significant differences for slices 3-4 KHz ($p = .003$), 5-6 KHz ($p = .001$) and 6-7 KHz ($p < .0001$).

The Alpha-ratio is significantly different between the two voicing conditions for the (0 to 1 KHz) / (1 to 10 KHz) variant, but not for the (0 to 1 KHz) / (1 to 5 KHz) variant.

C. Mean Fo and duration

Mean speaking frequency was in average 225.0 Hz (SD: 20.29) in the normal voicing condition and 250.7 Hz (SD: 23.57) in the acting condition. The total duration of reading was 102.05 s. (SD: 13.12) in the normal condition and 111.39 (SD: 11.85) in the acting condition. These differences are statistically significant, respectively with a $p < .001$ and a $p < .05$.

IV. DISCUSSION

The design of this experiment aimed at optimizing the comparison of the two voicing conditions by using the same subjects (all professional actresses) in both conditions, by selecting an emotionally expressive passage, and by asking the subjects to produce a louder than normal voice in the non-acting condition.

There are important inter-individual variations, but statistically, there is significantly more acoustical energy in the 3-4 KHz slice when the subjects are acting, even if this difference does not look like a true 'formant', as observed in male singers, and even in male actors. This difference is probably related to resonances in the vocal tract.

Further, a clear difference between the two voicing conditions is also noticed in the higher part of the spectrum, particularly in the slices 5-6 and 6-7 KHz. This difference is also reflected by the Alpha-ratio (0 to 1 KHz) / (1 to 10 KHz) variant. The higher acoustical energy level observed in the acting conditions is likely related to a glottal mechanism, possibly a longer closed phase. The chosen passage is a very assertive one, tending to elicit a tense, hypertonic style phonation style. Even if was tried to control both voicing conditions for intensity, in average the acting condition was slightly louder.

The acoustical differences are confirmed by the perceptual experiment: the acting voices sounded obviously different from the non-acting voices. The average pitch and average duration of reading were also slightly higher in the acting condition.

V. CONCLUSION

The actresses' acting voice sounded clearly different from their normal voice, as pointed out by ratings of blinded listeners.

Actresses showed significantly more acoustical energy in the 3-4 KHz zone of the long-time averaged spectrum in the acting condition, compared to the non-

acting condition, although this difference does not look like a true 'formant', as observed in male singers, and even in male actors. The difference is probably related to resonances in the vocal tract. Moreover, the actresses showed also more energy in the zone 5 – 8 KHz in the acting condition, which is likely related to a glottal mechanism, possibly a longer closed phase.

In acting condition, the mean speaking frequency was enhanced and the speech rate slightly slowed.

REFERENCES

- [1] J. Sundberg, "The acoustics of the singing voice" *Scientific American*, vol. 236(3), pp. 82 – 91, 1977.
- [2] T. Leino, T. "Long-term average spectrum study on speaking voice quality in male actors" in *SMAC93: Proceedings of the Stockholm Music Acoustic Conference*, A. Friberg, J. Iwarsson, E. Jansson & J. Sundberg Eds. Stockholm: Royal Swedish Academy of Music 1994, pp. 206-210.
- [3] T. Nawka, L.C. Anders, M. Cebulla and D. Zurakowski, "The Speaker's Formant in Male Voices." *Journal of Voice*, vol. 11, pp. 422 - 428. 1997.
- [4] R. Pinczower, and J. Oates, "Vocal Projection in Actors: The Long-Term Average Spectral Features That Distinguish Comfortable Acting Voice From Voicing With Maximal Projection in Male Actors." *Journal of Voice*, vol. 19, pp. 440-453, 2005.
- [5] S. Master, N. G. de Biase, B. M. Chiari and A. M. Laukkanen, "Acoustic and Perceptual Analyses of Brazilian Male Actors' and Nonactors' Voices": Long-term Average Spectrum and the 'Actor's Formant'", *Journal of Voice*, vol. 22, 146 - 154, 2008.
- [6] Leino, T., A.-M. Laukkanen & V. Radolf, "Formation of the Actor's/Speaker's Formant: A Study Applying Spectrum Analysis and Computer Modeling", *Journal of Voice*, vol. 25, 150-158, 2011.
- [7] S. Master, N.G. de Biase and S.J. Madureira, "What about the "actor's formant" in actresses' voices?" *Journal of Voice*, vol. 26(3), pp. 117 – 122, 2012.
- [8] T. Cleveland, J. Sundberg, R.E. and Stone, "Long-Term-Average Spectrum Characteristics of Country

Singers During Speaking and Singing”, *Journal of Voice*, vol. 15, pp. 54 – 60, 2001.

[9] B. Frokjaer-Jensen and S. Prytz, “Registration of Voice Quality” *Technical Review Bruel & Kjaer* Nr. 3, pp. 3 – 17, 1976.

SHORT TERM EFFECT OF ‘SEMIOCCLUDED VOCAL TRACT EXERCISES’ ON HEALTHY ACTORS’ VOICES

V. Di Natale¹, G. Cantarella², C. Manfredi³, A. Ciabatta², C. Bacherini³, P. H. DeJonckere⁴

¹ Università degli Studi di Milano, Milano, Italy

² Department of Otolaryngology, Fondazione IRCCS Ca’ Granda Ospedale Maggiore Policlinico, Milano, Italy

³ Department of Information Engineering, Università degli Studi di Firenze, Firenze, Italy

⁴ Federal Agency for Occupational Risks, Brussels, Belgium

dinatale.vale@tiscali.it

giovanna.cantarella@policlinico.mi.it

claudia.manfredi@unifi.it

annaclara.ciabatta@policlinico.mi.it

ph.dejonckere@outlook.com

Abstract: The aim of this study was to investigate the effect of a 10-minutes warm-up protocol with semi-occluded vocal tract exercises (SOVTE) on actors without voice complaints. A short dramatic passage was audio-recorded at 4 time points. Between the second and the third recording the actors performed the exercises, while between the third and the fourth they performed in a show. The voice quality was acoustically and auditory-perceptually analysed at each time point by blinded raters. Self-assessment parameters anonymously collected pre and post exercising were also evaluated. No statistically significant differences on perceptual ratings and acoustic parameters were found between pre/post exercise session and males/females. Statistically significant improvement was found in the self-assessment parameters concerning comfort in production (males), sonorousness, vocal clarity and power (both males and females). The proposed vocal warm-up with the SOVTE protocol created may thus be effective in determining a self-perceived improvement in comfort, voice quality and power.

Keywords: theater actors, vocal warm-up, SOVTE

I. INTRODUCTION

Theater actors are voice professional with high vocal demands. These lead to laryngeal hyperfunction and perceived vocal fatigue [1,2]. In a study of Rangarathnam et al. [3], auditory-perceptual and aerodynamic measures of theater actors’ voices significantly deteriorated after six weeks of stage performances and rehearsals, indicating a voice quality worsening.

In this population voice problems have shown to lead to both poor psychological and occupational issues [1,4].

Previous studies on the effects of semi-occluded vocal tract exercises (SOVTE) showed that phonating

while reducing the diameter of the vocal tract at the level of the tongue and the lips can result in a more effective and efficient vocal production on healthy, disordered and singers’ voices [5–8].

According to Titze [9] and Guzman et al. [10], this effect is due to an increased oral pressure.

SOVTE can be divided into groups according to the resistance created by the semi-occlusion, which directly relates to the amount of oral pressure which is generated [9, 10].

Moreover, Amarante Andrade et al. [11] reported that SOVTE can also be classified according to the presence or absence of a vibratory component at the semi-occlusion level. In line with this model, exercises performed with a second source of vibration in addition to the vocal folds (named fluctuating) are characterized by a predominant massage effect of the vocal tract. Exercises performed with only the vibration of the vocal folds (steady) promote, instead, an easy phonation.

To the best of our knowledge, only one existing study has determined the effects of SOVTE on the actor’s voice, while most of the literature focused on singers’ and disordered voices. Leino et al. [12] found an increase of the speaker’s formant (a peak around 3.5 kHz in the Long Term Average Spectrum-LTAS of trained speaking voice) as well as a better perceived voice quality on a single actor after 30 minutes of SOVTE training. LTAS provides a visual display of the average frequency distribution of the energy of a continuous speech sample. It thus yields some information about the quality of the voice signal.

The aim of the present study was to determine the acoustical, auditory and self-perceived short-term effects of a vocal warm-up based on a protocol of different SOVTE in a population of actors with healthy voices.

II. METHODS

The study was performed on professional theater's actors on stage with a play, without voice complaints and with at least three years of working experience.

Difficulties in performing the exercises would have led to participants' exclusion.

All participants agreed to participate in the study by signing a written consent form.

4 audio-recordings (R1, R2, R3, R4) were made for each actor while interpreting and reading aloud with acting voice a selected short passage of "Hamlet" as translated in Italian and imagining an audience of 400 people.

The recording tool consisted of a SM58 dynamic microphone (Shure, Niles, Illinois, US) coupled with an US-322/366 external sound card (Tascam, Santa Fe Springs, California, US) and a VivoBook A551LB laptop (Asus, Taipei, Taiwan), Audacity® (Audacity Team, 2017) set at a 44.1 kHz sampling rate and 16 bit resolution. A constant mouth-to-microphone distance of 15 cm was kept with the use of a metal spacer.

A timeline of the data collection procedures is displayed in Fig. 1.

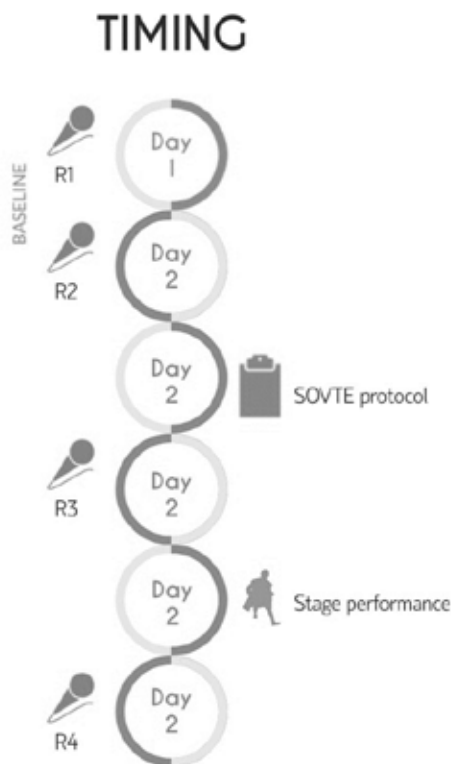


Figure 1. Timeline of the data collection procedures.

The exercise session consisted of a series of 10 minutes of SOVTE progressing from high to low resistance and including both steady and fluctuating tasks (Lax Vox with tube immersion of 3 cm in water,

20 vocalisations on /u/; straw phonation, 20 x sustained /u/; 10 lip trills, 10 tongue trills and 10 hummings).

R1, R2, R3 and R4 were analysed perceptually for voice quality by 5 blinded experts in voice analysis through a 100 mm Visual Analogue Scale.

Moreover, the recordings were acoustically evaluated with BioVoice [13] for duration, voiced/unvoiced selection, %voiced, %unvoiced, dynamic range, LTAS, jitter, quality ratio and mean, max, min, standard deviation of F0 (fundamental frequency) and of F1-F5 (formants). Praat [14] was used to analyse shimmer and noise-to-harmonic ratio.

Anonymously collected self-assessment parameters (comfort in production, sonorousness, expressiveness, pleasantness, vocal clarity and power) were also evaluated. For this analysis, participants were asked to fill in a questionnaire indicating, for each parameter, if their voice was worse, the same or better after the vocal warm-up as compared to its quality before SOVTE.

Data analysis was performed both together and separately for males and females with SPSS version 25 (IBM, Armonk), the significance level was set to 0.05.

The auditory-perceptual evaluations of the 5 raters were averaged as well as all the data of R1 and R2, which made the baseline.

Student t test and Cohen's d effect size were used to statistically and clinically analyse pre-post session changes in the perceptual and acoustic outcomes as well as the differences between males and females.

Binomial distribution and Fisher exact test were applied to analyse the self-assessment ratings.

Bonferroni correction was applied to statistically significant results.

III. RESULTS

The data analysis showed no statistically significant differences both for the perceptual ratings and for the acoustic parameters between pre/post exercise session and males/females. Due to space limitations, the results of acoustical analysis are not detailed here. Fig. 2 displays the perceptual rating for all actors. Similar results were found separately analyzing male and female data.

This analysis showed a high inter-individual variability. For instance, the two examples of LTAS reported in fig. 3 and 4 show an opposite result, as in the former we can notice a decrease in the energy level around 3.5 kHz, while in the latter an increased energy peak around the same frequencies after the exercises.

Statistically and clinically significant improvement was found in the parameters concerning comfort in production (males, $p < 0.001$), sonorousness, vocal clarity and power (both males and females, $p < 0.05$).

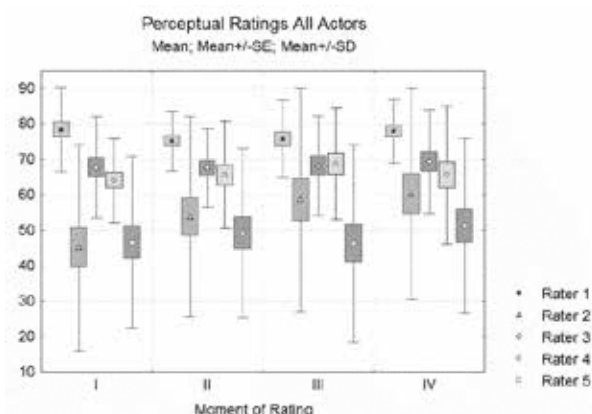


Figure 2. Mean, standard error (SE) and standard deviation (SD) of the actors' perceptual evaluation at each time point and for each rater.

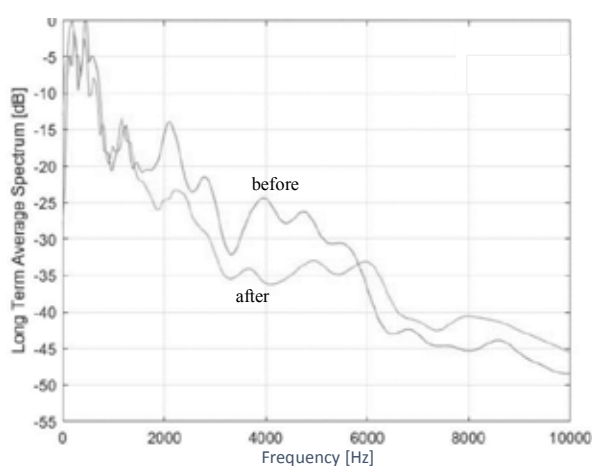


Figure 3. Energy level [dB] across Frequency [Hz] before and after the exercise session in a single actor.

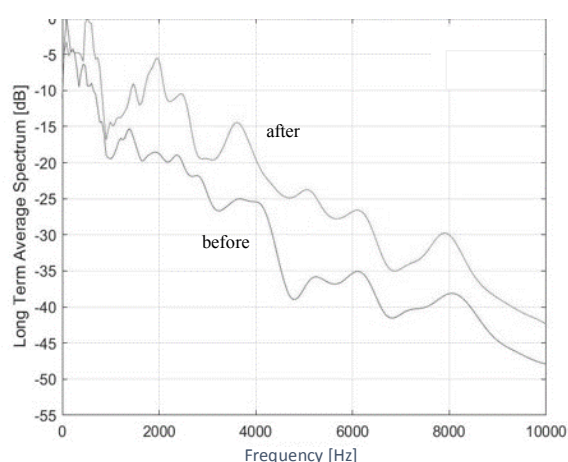


Figure 4. Energy level [dB] across Frequency [Hz] before and after the exercise session in a single actor.

IV. DISCUSSION

The present study aimed at finding the short-term effects of a SOVTE warm-up protocol on the actor's voice.

After 10-minutes of vocal warm-up, no statistically nor clinically significant differences were found on either the acoustic and auditory-perceived voice quality parameters. These findings differ from the ones obtained in the previous work of Leino et al. [12] on one single actor. As the exercises performed in this previous study were 30 minutes in length, it might be that 10 minutes of vocal warm-up is not enough to create an effect at the acoustic and auditory-perceptual levels. The present study was conducted in a real environment and during the stage performance period rather than in an artificial context (voice laboratory). A shorter protocol was preferred being more feasible before the stage performances and to avoid a tiring effect on the voice. The conditions in which the present data were collected, however, are more generalizable to the real context.

The findings of this study show that a 10-minutes vocal warm-up with SOVTE is enough to determine a self-perceived improvement in comfort for males and in sonorousness, voice clarity and power for both males and females. This result is in line with the previous literature on the effects of SOVTE on the singer's voice [7].

This result is anyway relevant being the human voice an extremely complex mechanism whose study requires a multidimensional approach. The primary goal of clinicians dealing with the voice is indeed to help patients and professional users to fully use their voice without efforts, but at the same time expressively and strongly, thanks to the use of comfortable and effective tools.

V. CONCLUSION

In actors, vocal warm-up with a 10-minutes SOVTE protocol is effective in determining a self-perceived improvement in comfort, voice quality and strength.

REFERENCES

- [1] J.A. Kitch, J. Oates, "The perceptual features of vocal fatigue as self-reported by a group of actors and singers", *J Voice*, vol. 8, pp. 207–214, 1994.
- [2] M.Z. Lerner, B. Pashover, L. Acton, N. Young, "Voice disorders in actors", *J Voice*, vol. 27(6), pp. 705–708, 2013.
- [3] B. Rangarathnam, T. Paramby, G.H. McCullough, "Prologues to a Bad Voice': Effect of Vocal Hygiene Knowledge and Training on Voice Quality Following Stage Performance", *J Voice*, vol. 32(3), pp. 300–306, 2018.

- [4] B. Timmermans, M.S. De Bodt, F.L. Wuyts, A. Boudewijns, G. Clement, A. Peeters, P.H. Van de Heyning, "Poor voice quality in future elite vocal performers and professional voice users", *J Voice*, vol 16(3), pp. 372–382, 2002.
- [5] M.R. Kapsner-Smith, E.J. Hunter, K. Kirkham, K. Cox, I.R. Titze, "A Randomized Controlled Trial of Two Semi-Occluded Vocal Tract Voice Therapy Protocols", *J Speech, Lang Hear Res*, vol. 58, pp. 535–549, 2015.
- [6] M. Guzman, R. Jara, C. Olavarria, P. Caceres, G. Escuti, F. Medina, L. Medina, S. Madrid, D. Muñoz, A-M. Laukkanen, "Efficacy of Water Resistance Therapy in Subjects Diagnosed With Behavioral Dysphonia: A Randomized Controlled Trial", *J Voice*, vol. 31(3), pp. 385.e1-385.e10, 2017.
- [7] A.L.F. Mendes, R. Dornelas do Carmo, A.M.G. Dias de Araújo, L.R. Paranhos, C.S.O. da Mota, S.S.V. Dias, F.P.A. Reis, J.A. Aragão, "The Effects of Phonation Into Glass, Plastic, and LaxVox Tubes in Singers: A Systematic Review", *J Voice*, vol. 33(3), pp. 381.e1-381.e9, 2019.
- [8] M. Guzman, A-M. Laukkanen, P. Krupa, J. Horáček, J.G. Švec, A. Geneid A, "Vocal tract and glottal function during and after vocal exercising with resonance tube and straw", *J Voice*, vol. 27(4), pp. 523.e19-523.e34, 2013.
- [9] I.R. Titze, "Voice training and therapy with a semi-occluded vocal tract: rationale and scientific underpinnings", *J Speech, Lang Hear Res*, vol. 49(2), pp. 448–459, 2006.
- [10] M. Guzman, G. Miranda, C. Olavarria, S. Madrid, D. Muñoz, M. Leiva, L. Lopez, C. Bortnem, "Computerized Tomography Measures During and After Artificial Lengthening of the Vocal Tract in Subjects With Voice Disorders" *J Voice*, vol. 31(1), pp. 124.e1-124.e10, 2017.
- [11] P. Amarante Andrade, G. Wood, P. Ratcliffe, R. Epstein, A. Pijper, J.G. Svec, "Electroglottographic study of seven semi-occluded exercises: LaxVox, straw, lip-trill, tongue-trill, humming, hand-over-mouth, and tongue-trill combined with hand-over-mouth", *J Voice*, vol. 28(5), pp. 589–595, 2014.
- [12] T. Leino, A-M. Laukkanen, V. Radolf V, "Formation of the actor's/speaker's formant: A Study Applying Spectrum Analysis and Computer Modeling", *J Voice*, vol. 25(2), pp. 150–158, 2011.
- [13] C. Manfredi, D. Barbagallo, G. Baracca, S. Orlandi, A. Bandini, P.H. Dejonckere, "Automatic Assessment of Acoustic Parameters of the Singing Voice: Application to Professional Western Operatic and Jazz Singers", *J Voice*, vol. 29(4), pp. 517.e1-517.e9, 2015.
- [14] P. Boersma and D. Weenik, "Praat: doing phonetics by computer [Computer program] version 6.0.25" retrieved 6 march 2017 from <http://www.praat.org/>, 2017.

SINGING WHILE ACTING AND VICE-VERSA

Orietta Calcinoni¹

¹ VMPCT, Milano, Italy
orietta.calcinoni@gmail.com

Abstract: It is common opinion nowadays, that acting and singing are just two ways to perform in voice and that Artists trained in one voicing modality will use the other one with no problems, often alternating in the same performance. The Author reviews up-to-date knowledge and results about physiological and psychological processes involved and how instrumental data may help these voice professionals to cope their tasks or be aware of voice damage risks.

Keywords: acting voice, singing voice, vocal tract patterns.

I. INTRODUCTION

Vocal tract evolution. Gestures and vocalizations, as those identified by Goodall in chimpanzees, evolved through prehistorical men with a representative meaning, as for sacred ceremonies or representations. In parallel, vocal tract anatomy varied, gaining a wider “vowel triangle” and many more consonants than those possible in primates, Neanderthal men and... our newborn babies [1].

History of “vocal art”. Acting and singing share common roots. Ancient Egyptians, around 2000 BC, performed annually the story of God Osiris. Veda (2000 BC- 500 BC) contain some lines about priests-actors rules, and during Shang dynasty in China (1500 BC) began theatrical entertainments with music clowning and acrobatic displays. Dithyrambs hymns in honor of Dionysus are considered the earliest origins of drama, while tradition tells us that in 534-535 BC Thespis, a wandering bard, won the City Dionysia in Athen, reciting poetry on a wooden chart as if he was the characters whose lines he was reading. From this “world’s first actor” came the adjective “thespian”.

Anyway, “since last century, in Theatre History studies we find not much available beyond schematics about movement, mimic expressions, gestures and pretty nothing of the actual sound of an actor’s voice” [2].

It looks indeed curious that we “hear” Kean’s or Garrick’s voices from paintings and pictures, as well as now most advanced studies verify vocal tract’s mathematical model from mouth opening profile.[3]

Characterizing aspects in singing and acting voice. Since last century we observed an evolution in singing

voice performing, because agilities trills and warbles left place to correct pronunciation and articulation joint to interpretation of the verbal text as well as of the musical score. In acting voice, the prevalent actual target is to resemble normal speech as much as possible [4].

TABLE I - SINGING VS SPEAKING DIFFERENCES (From Nix J, modified)

SINGING VS SPEAKING (J NIX 1964)	SPEAKING	SINGING
INTELLIGIBILITY	PRIMARY GOAL	SACRIFIABLE FOR AESTHETICS
PITCH, DURATION, INTENSITY	VARY ACCORDING TO SPEAKER'S INTENTIONS	PRESCRIBED BY THE COMPOSER, MAY VARY IN DEFINED RANGES
INTENSITY RANGE	OVERALL NARROWER	IN AVERAGE WIDER
FREQUENCIES RANGE	OVERALL NARROWER	IN AVERAGE WIDER
RESPIRATORY PATTERN	LOWER TIDAL VOLUME IN AVERAGE-PHRASE	HIGHER TIDAL VOLUME IN AVERAGE-PHRASE
VOWELS DURATION (IN ENGLISH)	VOWELS TO CONSONANT TIME 3:1	VOWELS TO CONSONANT TIME 200:1 (10 sec vowel followed by 50 ms consonant)
CONSONANTS CO-ARTICULATION	FREQUENT	DEPENDS FROM SINGING STYLE: THE MORE TEXT-DRIVEN THE MORE CO-ARTICULATED
"PROJECTION"	NON NEEDED (OR SUPPORTED BY MIC) IN COMMON SPEECH IN ACTING "SIMILAR STAGE STRATEGIES" WITH SINGING	INCREASED SUBGLOTTIC PRESSURE, RESONANCE ADJUSTMENTS, VOCAL FOLDS ADDUCTION ADJUSTMENTS (LESS STEEP SPECTRAL SLOPE, MORE INTENSITY TO HIGHER PARTIALS)
PUSH RISK (OVERADDUCTING VOCAL FOLDS, OVERWORK BREATHING-WISE, CONSTRUCTIVE TENSIONS WITH REDUCED GOOD RESONANCE)	IN COMMON SPEECH FROM INEDUCATED PHONATION IN ACTING FROM EXAGGERATED PHONATION	IN SINGING, IF NOT OCCASIONAL AND MOTIVATED BY THE INTERPRETATION, HIGH RISK OF DAMAGE WHILE BRISK REDUCTION IN VOICE EFFICACY

Novak et al [5] demonstrated that a large amount of professional actors claimed vocal fatigue from laryngeal hyperactivity, characterized by supraglottic structures hyperadduction. [6] The relation of this pattern with Muscle Tension Dysphonia, glottal incompetence, vocal nodules and polyps was questioned later by an Experts’ review [7]

Purpose In literature there is a huge amount of works about singing voice, or voice impairment treatment. But not so many regarding the “difference” - *if any-* in acting voice vs common speech and singing voice, and how common or expert listener may perceive it. In producing an /æ/ vowel (as in “mad”), the vocal tract approximates a megaphone shape . The syllable /pae/ has been widely chosen as item to study voice production in neutral, breathy, flow and pressed patterns. Most of the studies involved EGG signal, but in this first experiment we looked for differences in

lips and vocal tract pharyngeal portion (glottis to tongue base and velum) movements vs in vivo spectral visualization and analysis of vocal production..

II. METHODS

Participants and protocol. 4 subjects, (2 males, 2 females, mean age 32) all Italian native, with both academic education and professional experience in acting and singing voice (mean 10 years) were requested to pronounce the syllable /pae/ in three different modalities: common “normal speech” voice, projected “acting” voice, singing voice; each in the four different types of phonation described by Sundberg [8] neutral/modal, breathy, flow/resonant, pressed. The /pae/ syllable permits to overcome native language problems, besides this vocal item is used in a large number of papers about phonation analysis. Better than common used spoken or musical phrases, it may also guarantee from hard attack and from cultural/emotional interferences in phonatory behavior of a subject. The projected speech was assumed as non emotional-linked “acting voice”, apart from discussions on Classical or Method Acting. The singing score chosen for singing task was the first line of “Summertime” from Porgy and Bess Gherwin’s opera, original score 1935 © by Gherwin Publishing Corporation. This score was chosen because of some previous papers [9, 10]. The examined subject might choose freely the singing style to use, but avoiding interpretation or stylistic excesses, using /pae/ instead of the words of the lyrics.

Method. The twelve items - speech, projection, singing, in neutral, breathy, flow, pressed - were pronounced in the three examination modes.

- during videoregistration of “non exaggerated” lips movement, under stroboscopic light, with 0° rigid optic fiber Fiegert, software Daisy© by Inventis srl
- in videorhinopharyngofibrosopy with flexible optic fiber Xion GmbH, always with stroboscopic light, same software.

In in vivo continuous registration of voice spectrum with VoceVistaPro© software.

III. RESULTS

The video and the .wav files from VoceVistaPro were analyzed frame by frame and segment by segment to look for intrasubject differences and intersubject analogies in different voicing patterns. In the sung task, the examined portion was the second “sustained” F, corresponding to “time” (*summer-time*).

TABLE II - LIPS VISUALIZATIONS

Tab. 2 Lips visualizations

Lips stroke speech	Neutral	Breathy	Flow	Pressed
M1 (Head hold still in the same position)	3 frames	11 frames (parted lips)	16 frames (lips a bit more opened)	8 frames (mouth wide opened)
M2 (head moving to underline flow and pressed)	13 frames	9 frames	8 frames	10 frames
F1	13 frames	20 frames	25 frames	7 frames
F2 (moving head very often)	5 frames (flabby lips)	5 frames (thinner lips)	20 frames (curling upper lip)	13 frames (thin upper lip, protruding lower lip)
Projected				
M1 (the only one maintaining lips)	12 frames (opens more than in “normal neutral speech”)	14 frames (opens more than in “ breathy speech”)	10 frames (opens more than in “flow speech”)	9 frames (opens more than in “pressed speech”)
M2	6 frames	10 frames	8 frames	12 frames
F1	8 frames	12 frames	8 frames (parted lips)	11 frames
F2	5 frames (plative)	9 frames (lips slowly a bit more opened)	6 frames (opens more than in “flow speech”)	6 frames (mouth wide opened quickly)
Sung /F (“time...”)				
M1	21 frames (lower lip covers mandibular arch)	35 frames (lower lip covers mandibular arch)	29 frames (parted lips lower lip covers mandibular arch, upper lip covers maxillar arch)	21 frames (mouth even more wide opened)
M2	32 frames	22 frames	28 frames (without closing mouth at the end)	23 frames
F1	28 frames	35 frames (open and closes slowly)	34 frames	41 frames
F2	13 frames	24 frames	10 frames	11 frames

One of the subjects (M1) holds head still, while two (M2 and F1) cannot avoid moving the head in some of the phonation types. Two of the subjects (M1 and F2) change clearly lips’ patterns with phonation types and voicing modalities.

TABLE III - VOCAL TRACT FIRST PORTION’S PATTERNS

Pharyngeal speak	Neutral	Breathy	Flow	Pressed
M1	Closed V rapidly opened, mild precontraction MPC	Closed V rapidly opening, less tight AR abduction	Slower V closure, widening OP	Slower V closure, widening OP, tightening AR
M2	Closed V rapidly opened	Semi-closed V, widening OP	More widened OP, TB moves forward, middle-ventricular narrowing AP EL	MPC contracts and V rapidly more widening
F1	Pharyngeal abduction AR and VFs Tightened HP and EL	Slowly reducing posterior parting narrowing EL	TB moves backward, narrowing AP EL	Keeps posterior
F2	Pharyngeal AR and VFs abduction, small posterior gap narrow HP	strong layers, wider angled opening, slower relaxation	Pharyngeal abduction, lowering larynx, closing OPV	Narrowing HP, lowering larynx, complete cord abduction, tube cord layer abduction
Projected				
M1	Same to speech but Widening OP	less tight AR abduction VFs to constant abduction	Widening OP and HP	Contracting OP and HP
M2	Closed V opens with MPC evident	Semi-closed V, widening OP all more slower	Widening laterolateral OP	V closes
F1	High OP contraction, mediating amplification	High OP contraction, mediating amplification, lowering larynx and semi-closed VFs	TB moves forward, raising larynx a little	TB moves forward, raising larynx a little, High OP contraction, mediating amplification
F2	Elevating V, no posterior gap, narrowing HP	Slower VFs “vocalist” abduction	Widening EL, narrowing HP	Narrowing EL, lowering HP, raising larynx
Sung				
M1 (every stable laryngeal position)	Widening OP, EL, and HP, small posterior triangle	Widening OP, EL, and HP, small posterior triangle, more widening OP and constant pressure VFs	Tightening a bit OP, lowering a bit larynx	Tightening OP and HP, widening all structures,
M2	Widened portion of the 1st part of VT, epiglottis opens vertically to show glottis	Everything visible, widened, for a longer time	TB more retro-positioned, wider profile, raising, tighter abduction	AR pressed, less tighter abduction, narrower epiglottis
F1	Keeps position of projected speech	Slower movements, TB backward	TB more retro-positioned, High OP contraction, mediating amplification	Semi-closed 1st part of VT, only AR partly visible
F2	Elevating V and raising larynx	Semi-closed OPV, widening medially all slower, no obvious no “vocalist” all ends in breath	Narrowing HP narrowing EL, raising larynx but more breath to sustain sound	Tightening V, OP HP and EL, raising larynx

V: Velum; MPC: Medium Pharyngeal Constrictor; AR: Arythenoids; OP: OroPharynx; TB: tongue base; HP: Hypopharynx; EL: EpyLarynx; VFs: Vocal Folds; VT: vocal tract. Projected voice shows in each subject some adjustments from former “neutral speech” pattern. M1 seems to choose an “open” pattern, with small adjustments for each task and a very stable laryngeal position. M2 progressively widens with some

constriction from tongue base and medium pharyngeal constrictor. F1 chooses to elongate VT first portion in projected and singing voice but activating medial pharyngeal constrictor and medializing amigdalae, progressively reducing laryngeal visualization. F2 modulates raising and lowering larynx more than widening 1st portion of VT. In VoceVistaPRO spectral analysis, F1 produces higher harmonics in all three pressed voicing tasks, while F2 recurs to higher harmonics in flow phonation with projected voice. M1 uses higher harmonics in pressed voice, but doubles voicing time in projected vs neutral speech and enhances higher harmonics reducing intensities while singing. M2 shows constantly higher harmonics in pressed voice but with risen intensities of lower harmonics in singing voice. Both obtain higher intensities than females, but projected voice use lower intensities than normal speech and singing voice..

IV. DISCUSSION

In literature we have studies about impairments in acting voice professionals [11,12], studies about prevention in acting voice [13] or studies about emotions perception in performing voice [14] Significant random intercepts were reported in all acoustic features, indicating a consistent tendency by some vocalists to exhibit higher or lower levels of these acoustic measures than other vocalists, even when controlling for the effects of their vocal experience background. [15] About singing voice, Herbst described two basic laryngeal adjustments, cartilaginous adduction and membranous medialization, in singers:

a. *cartilaginous adduction* i.e. adduction of the posterior glottis, with arytenoids squeezed together, permits strong high-frequency components (partials) distinguish pressed from breathy, were posterior gap from arytenoids set apart, results in high frequency partials of lesser strength;

b. *membranous medialization* i.e. medial bulging of the vocal folds, by thyroarytenoid muscle activity, permits adjustments between chest and falsetto, were in the first the acoustic spectrum contains string high-frequencies partials, while falsetto contains weaker overtones with less resonant sound. The degree of control of the two maneuvers permits to produce a variety of vocal timbres [16]. Our subjects choose different patterns in each voicing modality, so “projected” voice shows differences from “normal speech” voice. The different ways these few subjects modulate their vocal tracts prove more or less ergonomic value: M1 seems to use lips’ dynamics and stability in time to enhance projection, while F2 looks

at higher risk of vocal fatigue, changing continuously lips, vocal tract (and head) patterns.

TABLE IV - VOCE VISTA PRO IN FEMALE SUBJECTS

Tab.4. VoceVistaPro® in female subjects

Speech	Neutral Hz	Breathy Hz	Flow Hz	Pressed Hz		Neutral dB	Breathy dB	Flow dB	Pressed dB
F1	199.8	152.6	165.9	177.8	Higher harmonics in pressed	-27.4	-40.0	-31.2	-30.6
	181.1	287.5	323.6	354.8		-28.0	-37.9	-35.8	-25.2
	537.8	499.6	512.1			-33.5		-28.8	-19.6
	489.1		2319.9	2482.3		-45.1		-38.8	-35.6
F2	211.8	200.6	231.0	229.3	Higher harmonics in pressed	-26.3	-31.4	-22.2	-21.3
	668.8	393.2	689.3	684.5		-25.2	-23.4	-23.8	-17.5
	810.4	788.9	901.6	901.1		-22.7		-26.6	-20.6
	1955.7	1886.7	2064.2	2062.9		-26.2	-31.5	-23.9	-18.1
Projected F1	190.8	171.6	209.3	197.1	Higher harmonics in pressed	-29.6	-27.9	-26.8	-21.8
	184.5	346.6	626.6	617.8		-24.8	-21.1	-24.5	-27.2
	572.4	536.6	2304.5	2438.0		-28.3	-33.4	-20.4	-27.3
	2323.7	686.8	2513.8	2827.8		-36.4	-41.4	-26.6	-27.9
F2	826.7	238.4	217.3	213.1	Higher harmonics in flow	-19.3	-14.2	-22.0	-23.0
	1630.0	681.8	659.4	437.3		-19.2	-17.4	-19.6	-23.6
	1886.5	914.8	873.0	877.2		-20.9	-19.3	-19.1	-22.5
	1930.9	2070.0	1968.7	1968.8		-19.6	-20.1	-19.0	-20.0
Sung F1	291.3	280.5	283.3	371.1	Higher harmonics in pressed	-30.8	-31.1	-34.7	-26.6
	581.7	373.3	584.3	837.3		-43.8	-42.8	-33.1	-21.0
	866.6	864.9	871.7	1719.1		-34.6	-39.7	-29.1	-31.0
	2313.9		2312.0	2280.6		-43.1		-36.8	-27.9
F2	347.2	341.7	671.1	327.9	Higher harmonics in flow	-24.0	-26.2	-16.2	-18.0
	694.7	680.6	1010.0	882.3		-22.6	-23.6	-16.1	-16.8
	975.0	1371.6	1345.2	917.7		-26.5	-33.7	-24.3	-14.1
	1747.8	1768.8	1678.6	1638.8		-28.2	-30.1	-18.5	-14.3

TABLE V - VOCE VISTA PRO IN MALE SUBJECTS

Tab.5. VoceVistaPro® in male subjects

Speech	Neutral Hz	Breathy Hz	Flow Hz	Pressed Hz		Neutral dB	Breathy dB	Flow dB	Pressed dB
M1	Speech time more of 23%				Higher harmonics in flow and pressed				
	198.1	311.3	373.1	321.9		-36.1	-48.9	-39.1	-31.9
	1786.2		1780.4	1944.8		-41.1		-41.0	-31.7
	1804.7		1866.2	1770.1		-48.2		-41.4	-35.9
M2	Speech time double				Higher harmonics in pressed				
	381.2	381.3	487.2	311.2		-36.1	-48.9	-39.1	-31.9
	321.2	384.1	587.4	1411.7		-31.8	-41.7	-48.2	-39.1
	1786.9	479.1	1710.9	2281.3		-34.9	-47.4	-33.6	-31.8
3119.0	376.1	3445.1	2340.3		-38.9	-46.9	-34.4	-37.6	
Projected M1	187.9	184.2	302.0	129.1	Higher harmonics in pressed	-33.3	-27.8	-31.7	-24.2
	495.7	547.4	596.1	1823.3		-34.0	-29.0	-38.3	-30.9
	582.4	568.7	1781.2	2472.8		-34.2	-31.9	-34.7	-38.8
M2	Double time Speech time		Same as speech		Higher harmonics in pressed				
	582.9	582.4	211.8	329.8		-29.4	-39.6	-31.9	-23.9
	1761.1	421.2	886.3	2469.3		-37.6	-25.8	-29.8	-29.7
	3288.0	824.8	1411.0	2446.7		-33.7	-29.2	-31.6	-39.0
3214.8	828.2	1711.1	2286.0		-34.8	-41.6	-31.7	-37.9	
Sung M1	228.3	229.1	229.8	229.6	Higher harmonics adjustments reducing intensities	-28.9	-28.8	-29.9	-26.4
	695.6	695.8	691.3	695.6		-40.1	-33.8	-34.8	-34.1
	689.3	691.6	924.8	1411.2		-40.8	-43.4	-37.3	-33.7
	1688.9	1679.6	1620.1	2368.3		-44.8	-41.4	-31.1	-28.0
M2	147.9	146.1	146.1	480.6	Higher harmonics in pressed	-40.9	-46.1	-41.7	-33.1
	282.7	286.3	282.4	1368.3		-39.6	-41.9	-41.1	-37.7
	649.8	613.1	641.6	2127.9		-41.9	-48.9	-44.6	-33.2
	981.1	981.6	2862.2			-40.1	-41.6	-37.9	-31.6
1010.1					-43.2				

V. CONCLUSION

Even if nowadays “acting voice” must be as natural as possible, trained subjects use different objective and acoustic patterns when speaking, projecting voice or singing. These results must be controlled and confirmed with larger groups of subjects, and in wider instrumental approaches.

REFERENCES

- [1] Boë LJ, Berthommier F, Legou T, Captier G, Kemp C, et al. (2017) Evidence of a Vocalic Proto-System in the Baboon (*Papio papio*) Suggests Pre-Hominin Speech Precursors. *PLOS ONE* 12(1): e0169321. doi:10.1371/journal.pone.0169321. <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0169321>G.
- [2] Hamilton Ball R, “The Shakespeare Film as Record: Sir Herbert Beerbohm Tree,” *Shakespeare Quarterly* 3, no. 3 (1952): 227–36.) [3] Kimbrough A M, "The sound of meaning: theories of voice in twentieth-century thought and performance (2002). *LSU Doctoral Dissertations*533 https://digitalcommons.lsu.edu/gradschool_dissertations/533
- [4] M. Rossi, *Tension relaxation and vocal-body schema in singer and actor*. Proceedings of “Foniatra e Canto, Prosa e Foniatra” 2nd Nat Symp, Salsomaggiore 1987. Artegrafica Silva, Parma 1988
- [5] Novak A, Dlouha O, Capkova B, Vohradnik M. Voice fatigue after theater performance in actors. *Folia Phoniatr.* 1991; 43: 74–78.
- [6] Stager SV, Bielamowicz SA, Regnell JR, Gupta A, Barkmeier JM. Supraglottic activity: evidence of vocal hyperfunction or laryngeal articulation? *J Speech Lang Hear Res.* 2000; 43: 229–238
- [7] Behrman A, Dahl LD, Abramson AL, Schutte HK, Antero-posterior and medial compression of the supra glottis: signs of nonorganic dysphonia or normal postures? *J Voice* 2003; 17 (3): 403-410
- [8] J Sundberg, *Research Aspects on Singing*, Royal Swedish Academy of Music, 1981
- [9] Manfredi C et Al, Automatic assessment of acoustic parameters of the singing voice: application to professional Western Operatic and Jazz singers, *J Voice* 2015 Jul; 29(4): 517. e 1-9. Dos: 10.1016/j.voice.2014.09.014.
- [10] Calcinoni O, Dosimetric Analysis of Different Vocal Styles, *CoMeT Session* , PEVOC 11, Florence 2015.
- [11] Lerner MZ, Pashkover B, Acton L, Young N. Voice disorders in actors. *J Voice.* 2013; 27: 705-8; doi: 10.1016/j.jvoice.2013.05.006.
- [12] J Wendler, W Seidner, *Phoniatic care of actors and singers mirrored in our basic documentation*, Proceedings XIV Int Symp CoMeT, Padua 1987, *Acta Phon Lat*, Vol X, 3, 1988, 332-337
- [13] G Baracca, A Nacci, *Come si può fare prevenzione per la salute vocale nelle scuole di canto e recitazione?*- Proceedings Voci e Suoni di dentro e di fuori Symposium, Padua, 11-13 April 2019
- [14] Livingstone SR, Choi DH and Russo FA (2014) The influence of vocal training and acting experience on measures of voice quality and emotional genuineness. *Front. Psychol.* 5:156. doi: 10.3389/fpsyg.2014.00156
- [15] Walzak, P., McCabe, P., Madill, C., and Sheard, C. (2008). Acoustic changes in student actors’ voices after[
- [16] C Herbst, *Investigation of glottal configurations in singing*, Doctoral Dissertation, Olomouch, Czech Rep, 2011

THE SHOUTING VOICE ABILITY AND THE IMPORTANCE AS A 'FITNESS' PARAMETER FOR ALL VOICE USERS

Josef Schlömicher – Thier¹, Hannes Tropper¹, Ingolf Franke²,

¹ Austrian Voice Institute and International Voice Center Austria

² WEVOSYS medical technology GmbH, Baunach, Germany

Abstract: The term "shouting voice" is understood to mean a voice production that is physiologically and healthily produced and thus by the rules of vocal hygiene. The shouting voice is created by producing a short, swelling, powerful sound using increased breath pressure and complete glottis closure. This tone represents the highest and loudest vocal range of the chest register. The measured frequency position of the shouting voice in the vocal field coordinate system, often an indentation of the vocal field, corresponds to the position of the register transition between the chest and head register. According to the voice evaluation protocol of the European Laryngological Society (ELS), the maximum performance in men is 95dB and in women 90dB. This can be used as a possible guide in the determination of voice registers and to define the "vocal field architecture", as same as to define the "fitness of the voice. We analyzed the acoustic structure and the dynamic of the shouting in four everyman shouters in the play "Everyman" at the Salzburg Festival and we will underline the hypothesis, that a professional shouter has an "Shouting Formant" similar to the resonance strategy of singers and the ability of an extraordinary shouting dynamics (113 – 120 dB A)

Keywords: Shouting Formant, Shouting Dynamics, Leap Interval, Fitness of the Voice

I. INTRODUCTION

1. History of Everyman

The dramatist Hugo von Hofmannsthal completed his reworking of the 15th century English morality play "Everyman" (Jedermann) in 1911. Later with Max Reinhard and Richard Strauss he founded the Salzburg Festival and the first production of Jedermann to be performed there was under the direction of Max Reinhard on 22.08.1920 on Cathedral Square. Alexander Moissi was cast as the first Jedermann. The Cathedral Square in Salzburg still offers a splendid

open-air backdrop for this morality play, only in the case of heavy rain does the Everyman Play move to the Large Festival Hall.

2. The Story

A rich squire, businessman, is extremely greedy and stingy. He lets his debtors be thrown into debtor's prison, good deeds happen rarely, but if then only to avoid a troublesome situation. So it is that to avert annoying lamentations he consents to the support of the family of his debtor, whom he has hard-heartedly thrown into debtor's prison. A splendid feast is prepared, where he is escorted by the Buhlschaft. Here, however, he begins to feel poorly. He hears bells ringing and loud voices calling his name - "Jedermann, Jedermann! ", things that he alone among the party-goers can hear.

3. The Shouting Voice

The term "shouting voice" is understood to mean a voice production that is physiologically and healthily produced and thus by the rules of vocal hygiene. The shouting voice is created by producing a short, swelling, powerful sound using increased breath pressure and complete glottis closure. This tone represents the highest and loudest vocal range of the chest register. The dynamics of the shouting voice defines the "fitness of the voice," the maximum performance in men is 95dB and in women 90dB. This is opposed to the term "**Screaming**", created by an exaggerated frequency, roughly creaking, fidgeting in the falsetto, and often emotionally charged.

II. METHODS

Shouting” Locations



There are four Jedermann shouters for each performance. They are positioned at different distances. Two shouters are located behind the Cathedral arches (left and right), A third shouter is in the Franciscan Church tower 150 meters away. The fourth shouter is located on the "Katz" at the Fortress Hohensalzburg and has to cope with the longest distance of about 500 meters

Material


Everyman Shouting Casting 2012

The measurement of the Salzburg Everyman-Shouter was carried out with the Lingwaves Phonetogram, with the lip microphone distance of 30 cm. With the measured at the Shouters a Shouting level values between 113 to 120 dB(A). It turns out that only Shouters with a level value of at least 113 dB (A) were accepted by the director because they were able to handle the resonating space of the Cathedral Square well.

Shouter	Dynamics in dB	F0
1	120 dB	349Hz, f4,
2	118 dB	380Hz, fis4
3	118 dB	306 Hz, dis4
4	113 dB	289 Hz, d4

The acoustic Analyse of the of the four shouters shows a significant energy increase in the range of 2000-3000 Hz in professional Shouter (Everyman Shouter) and could in analogy to singer formant (2800-3500 Hz) be named as the shouting formant (2000-3000 Hz).

8	3	14	118	459	ais4
5		12	115	368	fis4
12		12	106	345	f4
11		10	118	341	f4
4		9	113	353	f4
6		8	115	353	f4
9		6	114	287	d4
10		4	110	368	fis4
13			111	483	h4



III. RESULTS

The Measurement of four Shouters 2019 (System Lingwaves)

There were two Bariton Shouter (349Hz, 380Hz, f4, fis4), and two Bass Shouters (289Hz, 306Hz, d4, dis4). There was a shouting interval of a quart between the Baritone Shouters (380 Hz, fis4) and the Bass Shouters (289 Hz d4).

IV. DISCUSSION

It was surprising, that these middle Shouting Frequencies of the casting 2012 and 2019 are in the Leap Interval, where according the presentation of Don Miller, Jan Svec and Harm Schutte the characteristic leap interval (CLI) for the chest – falsetto leaps in a given voice is situated [1]. This region is called the first passagio for women and second passagio for men. The measured frequency position of the shouting voice in the Phonetogram coordinate system, corresponds to the position of the register transition between the chest and head register ([2], [3]) Schutte 1980, Gross 1981, Klingholz 1985) These two transitions show the phenomenon of a zone where the chest registers is left or reached again from above. This zone is between D4 (294 Hz) and F4 (349 Hz). It is precisely these register breaks that are artificially formed during yodelling. The Position of the Shouting Frequency is the Zone of the upper end of the chest register.

V. CONCLUSION

Only Shouters with a level value of at least 113 dB (A) were accepted by the director of the Everyman Play, because they were able to handle the resonating space of the Cathedral Square well. All the shouters have had a gifted shouting ability and a special "fitness of the voice." The measured frequency position of the shouting voice in the Phonetogram coordinate system corresponds to the position of the register transition between the chest and head register. The acoustic analysis shows a significant energy increase in the range 2000-3000 Hz in professional Shouter (Everyman Shouter) compared to a shouter with a weak shouting voice (Nonprofessional Shouter) and could in analogy to singer formant (2800-3500 Hz) be named as the shouting formant (2000-3000 Hz).

REFERENCES

- [1] D.G. Miller, J.G. Švec, HK Schutte. *Measurement of characteristic leap interval between chest and falsetto registers*. J Voice 2002; 16(1):8-19
- [2] J. Calvet, G. Malhiac. *Courbes vocales et mue de la voix*. J Franc Otorhinolaryngol 1952; 1:115-124
- [3] F. Klingholz, F. Martin, (1983) *Die quantitative Auswertung der Stimmfeld- Messung*. Sprache – Stimme – Gehör 1983; 7:106-110

THE ACTORS VOICE: LAUGHING, CRYING AND SHOUTING IN THE MRI

Bernhard Richter¹, Louisa Traser¹, Michael Burdumy¹, Matthias Echternach², Claudia Spahn¹

¹ Freiburger Institute for Musicians Medicine Medical Faculty, Freiburg University and University of Music, Freiburg, Germany

² Division of Phoniatics and Pediatric Audiology, Department of Otolaryngology, Munich University Hospital and Faculty of Medicine, Munich University (LMU), Germany

bernhard.richter@uniklinik-freiburg.de, louisa.traser@uniklinik-freiburg.de, burdumym@posteo.de, claudia.spahn@uniklinik-freiburg.de, matthias.echternach@med.uni-muenchen.de

Abstract: Summary of the presentation

I. INTRODUCTION

Laughing is catching for other humans and promotes a feeling of togetherness within the group. There is a branch of science which deals with gelotology – laughing – from Greek "gélōs" laugh. The act of laughing is accompanied by a forceful exhalation and repeated exhaling thrusts. The individual audible sound pattern when laughing is quite variable and is influenced by processes in the vocal tract, for example the placing of the tongue and the length and form of the resonance cavity. A distinction is made between various laughing types – rhythmic laughing, soft laughing, commentary laughing, and shouting laughing. Each type itself has various characteristics.

Just as laughing, crying is a central form of expression among human beings. The German word for crying "Weinen" derives from the Germanic root "wai" – related to "whine" or "woe", and refers to the smooth transition which can take place between sighing and crying. Crying usually expresses the emotions of sadness, angst or melancholy, but one can also cry for joy or anger. There are many different forms of crying, for example childlike crying, suppressed crying, or dramatic crying.

Another important vocal expression for actors on stage is shouting. When shouting, all three structural elements of the voice organ distinctly intensify their activities.

In a bunch of movies we are able to show the different configurations of the tongue and cavities in the vocal tract – and the voice box as well – during various types of laughing, crying and shouting.

II. METHODS

Nowadays, with the help of modern investigative techniques from high-tech medicine – including dynamic magnetic resonance tomography which can be performed with 24 frames per second (Burdumy et al. 2015) –, it is possible to observe most of the structures of the "voice instrument" inside the body which are involved in sound production. With the methods applied we are able to observe detailed movements of the respiratory system, the larynx and the vocal tract while an actor is performing Laughing, Crying and Shouting.

III. DISCUSSION

During the presentation those film-clips will be discussed with the audience.

DISTORTED VOCALITY: SUPRAGLOTTIC MANAGEMENT IN ACTING AND DUBBING

Franco Fussi¹, Eleonora Bruni²

¹ Ravenna, Italy

² Roma, Italy

ffussi@libero.it

Abstract: Summary of the presentation

I. INTRODUCTION

Actors often have to interpret moments of anger or characters with an altered and rough voice and are subjected to vocal malmenage. The procedures of

euphonic use of distorted vocal modalities by the actors are examined, according to a specific protocol of falsocordal and arytenoid activation with low glottic impact. Performative examples of training are also offered.

SESSION VIII
**VOCAL FOLDS PARALYSIS/
ABNORMALITIES**

PHONOSURGICAL TREATMENT OF BILATERAL LARYNGEAL PARALYSES

Giovanna Cantarella^{1,2}, Alessandra D'Onghia¹, Michele Gaffuri², Lorenzo Pignataro^{1,2}

¹Department of Clinical Sciences and Community Health, Università degli Studi di Milano; ²Department of Otolaryngology and Head and Neck Surgery – Fondazione IRCCS Cà Granda Ospedale Maggiore Policlinico di Milano, Italy

giovanna.cantarella@policlinico.mi.it, michele.gaffuri@policlinico.mi.it, alessandra.donghia@unimi.it, lorenzo.pignataro@unimi.it

Abstract: Bilateral vocal fold paralysis (BVFP) is a challenging clinical condition. The purpose of this study is to analyze the efficacy of a mini invasive permanent laterofixation of one vocal fold in widening the glottic respiratory space sparing the phonatory and sphincteric functions of the larynx. A mini-invasive modification of the permanent Lichtenberger laterofixation technique was performed to treat bilateral laryngeal immobility in 16 adult patients. Most patients had post-thyroidectomy BVFP (PT- BVFP); 1 had also posterior glottic scarring (PGS) blocking the cricoarythenoid joints, and 1 had Guillain-Barré syndrome. The patients were followed up with videolaryngostroboscopy, Maximal Phonation Time (MPT) measurement, Voice Handicap Index (VHI) questionnaire, and GRBAS perceptual evaluation. Airway patency was restored in all patients with PT-BVFP. No postoperative dysphagia or severe dysphonia occurred in any patient. The voice quality of the patients remained stable or improved after surgery in all except 2 cases, presenting moderate breathiness, and long-term stability was confirmed by MPT, GRBAS, and VHI (p ranging between 0.004 and < 0.001). The technique proved to be successful in restoring airway patency for BVFP secondary to peripheral lesion of the recurrent laryngeal nerves and allowed to spare the laryngeal phonatory and sphincteric functions.

Keywords: bilateral vocal fold paralysis; airway obstruction; vocal fold lateralization; phonosurgery

I. INTRODUCTION

Bilateral vocal fold paralysis (BVFP) is a life-threatening clinical condition that can be due to paralysis or fixation of the vocal folds and is associated with significant morbidity and disability. BVFP is most commonly iatrogenic, due to surgical trauma, mainly thyroidectomy. Other less frequent causes of BVFP are non-surgical trauma, neurological diseases, and extralaryngeal malignancies [1]. The consequent airway obstruction may sometimes be so severe to end up in emergency tracheotomy. Lichtenberger [2]–[4] described a suture laterofixation technique performed to displace laterally one vocal fold using a specifically designed needle carrier which allows to pass a stitch

from inside the larynx, through the thyroid ala to the skin, and he defined the procedure as *endoextralaryngeal lateralization* (EExLL). The purpose of this study is to demonstrate the efficacy of a modified mini invasive permanent laterofixation in widening the glottic respiratory space while sparing the phonatory and sphincteric functions of the larynx.

II. METHODS

Patients: Sixteen adult patients with BVFP were treated by means of a modified laterofixation technique. Four patients had undergone previous procedures to enlarge the glottic area and five patients required previous tracheostomy.

Assessment methods: Airways videolaryngostroboscopy to objectivate the degree of laryngeal obstruction and to allow to assess the changes achieved after surgery at follow-up.

Functional voice assessment was by Maximum Phonation Time (MPT) measurement and GRBAS perceptual scale. Subjective assessment with validated questionnaires to evaluate voice (VHI 30)[5], airway's function (MRC and COPD)[6], [7], swallowing function (EAT 10) [8].

Spirometry could be obtained in 2 cases pre and post-operatively.

Surgical technique: A mini invasive submucosal posterior cordotomy was performed and then non absorbable sutures were inserted through the thyroid ala cartilage and looped around the paralyzed VF at the level of the vocal process using a Lichtenberger's needle carrier for widening the glottic space.

All patients were clinically re-examined after one month, three months, six months and then every 12 months.

III. RESULTS

An adequate airway patency was achieved in 14 patients, preserving voice and swallowing functions. Two patients needed revision surgery to further improve the surgical outcome.

The procedure failed to achieve a satisfactory airway in the patient affected by Guillain Barre' syndrome, who remained tracheotomy-dependent and in the one with posterior glottic scarring.

No postoperative dysphagia or severe dysphonia occurred post-operatively in any patient. The voice quality of the patients remained stable or improved after surgery in all except 2 cases, presenting moderate breathiness, and long-term stability was confirmed by MPT, GRBAS, and VHI (p ranging between 0.004 and < 0.001)

In the following section we present one clinical exemplificative case.

Clinical case

Patient n. 13, female, aged 56, affected by post thyroidectomy BVFP.

Her clinical respiratory condition was worsened by concurrent obstructive sleep apnea syndrome and elevated BMI (28.67) in follow-up by the unit of Pneumology and by the Unit of Bariatric Surgery of our Hospital.

She had undergone a previous laryngeal surgery (right posterior cordectomy with temporary tracheostomy, then removed) as attempt to enlarge the respiratory space without any significant improvement.

The patient complained of severe dyspnea, which was chronic at rest and worsened by minimum exercise, dysphagia, especially for liquids, and dysphonia. The voice quality was also negatively influenced by the effects of previous laryngeal surgery.

Besides the assessment protocol described in section II, she underwent pre-operative spirometry which demonstrated a variable extrathoracic respiratory obstruction [9]

Spirometric indexes considered were[10]:

- FEV1 (Forced expiratory volume in 1 second L/s);
- FIV1 (Forced inspiratory volume in 1 second L/s);
- FEV50 (Forced expiratory volume at time intervals of 0.5 seconds L/s)
- PIF (Peak inspiratory flow L/s)
- PEF (Peak expiratory flow L/s)
- Empey index [11] : FEV1/PEF (mVUmin); a value > 10 indicating significant obstruction.
- FEF 50/ FIF 50: ratio of maximal expiratory flow and maximal inspiratory flow at 50% of vital capacity. This ratio suggests obstruction with a value < 0.3 or > 1

In the following tables and figures we show the pre and post-operative findings.

Fig. 1 shows the preoperative videolaryngoscopy in inspiration: a very narrow respiratory space can be seen. Fig. 2 shows the endoscopic result after six months, the glottic respiratory space is much wider.

Fig 1

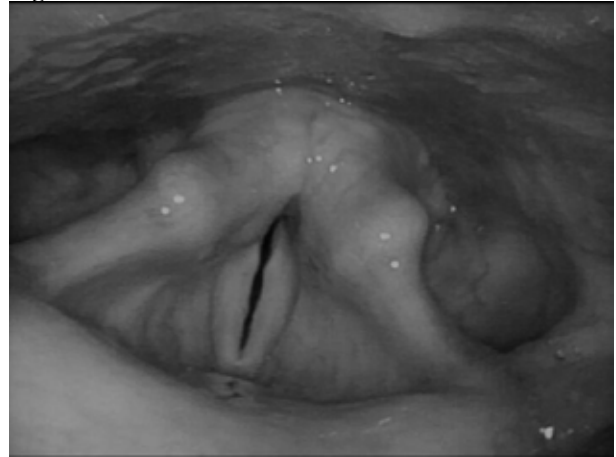


Fig 2

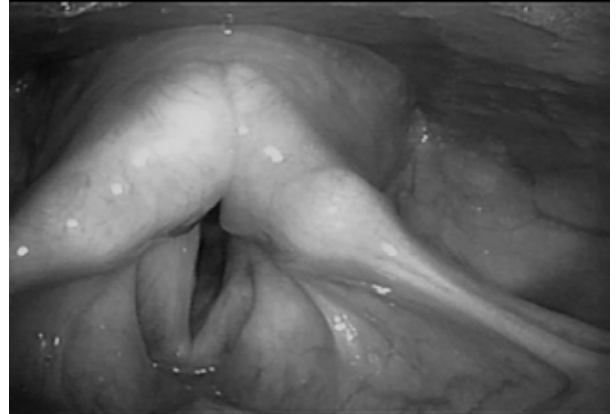


Fig. 3 and Fig. 4 show the spirometric flow – volume loop respectively pre and post-surgery. In Fig 4 the widening of the flow-volume loop both in inspiration and expiration indicates a decrease of the upper airways' variable extrathoracic obstruction.

Fig. 3

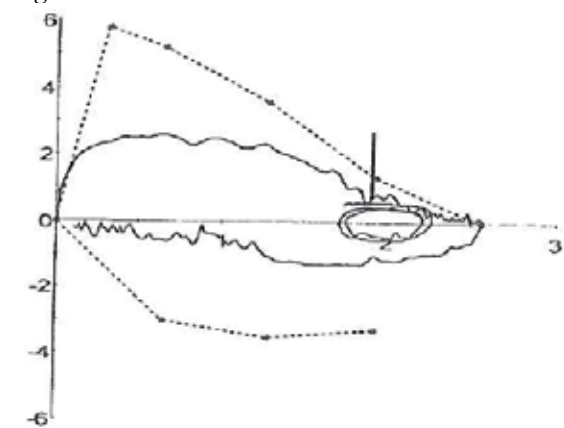


Fig. 4

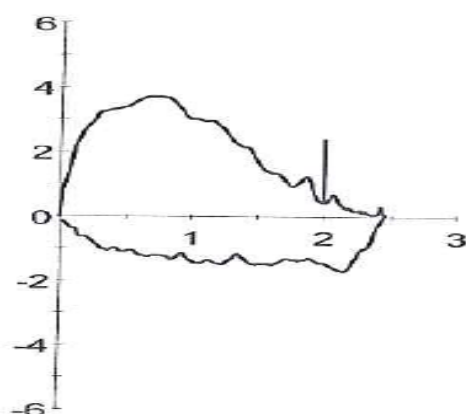


Table 1 shows pre and post-surgery values of spirometric measurements; there is an improvement for both inspiratory and expiratory volumes. A normalization of the Empey's Index is also noticed.

Table 1

	Pre-op	Post-op
FEV1 (L)	1,93 (89 %)	2,01 (94%)
PEF (L/s)	2,55(44 %)	3,9 (64%)
FEV50 (L/s)	2,27(63%)	2,75 (77%)
PIF (L/s)	1,28(33%)	1,70(44%)
FIVC (L)	2,47(80%)	2,50(83%)
FEV1/PEF (mVUmin)	12,61	9,34
[Empey's index]		
FEF 50/ FIF 50	1,92	1,84

Table 2 shows the pre and post-operative scores of the questionnaires about the laryngeal functions and the objective measurement of maximum phonation time. The improvement found with all questionnaires demonstrates a significant positive impact on patient's quality of life related to the respiratory, phonatory and sphincteric laryngeal functions after surgery.

Table 2

	Pre-op	Post-op
MRC	3	2
COPD	47	24
EAT10	10	2
VHI	55	10
MPT (s)	7	9

IV. DISCUSSION

BVFP is a life-threatening condition which has a severe impact on the quality of life of the affected patient. Tracheotomy has traditionally been the first line management for BVFP. A surgical widening of the endolaryngeal airway usually causes a severe

deterioration of voice quality and often of swallowing efficiency. The ideal treatment should allow to spare voice quality and give the possibility of being adjustable to each single case.[12]

Our technique allowed to widen the airway lateralizing in a permanent way the posterior third of one vocal fold in order to keeping the integrity of the phonatory part (the two anterior thirds) of the operated vocal fold. In our series as an overall improvement of voice quality and swallowing was achieved.

Our modified technique is a mini-invasive modification of the standard Lichtenberger description. [13]. The main modification consists of a posterior mini cordotomy with submucosal vaporization of the vocal muscle.

Our results demonstrated that there is no need for a wide tissue and arytenoid excision to achieve a wider respiratory space.

V. CONCLUSION

The modified vocal fold laterofixation technique is a safe, effective and well-tolerated approach for BVFP.

This technique allows to obtain an immediate respiratory benefit, stabilized within three weeks, while preserving the patients' voice quality.

VI. REFERENCES

- [1] L. H. S. Rosenthal, M. S. Benninger, and R. H. Deeb, "Vocal fold immobility: A longitudinal analysis of etiology over 20 years," *Laryngoscope*, vol. 117, pp. 1864–70, 2007.
- [2] G. Lichtenberger, "Endo-extralaryngeal needle carrier instrument.," *Laryngoscope*, vol. 93, no. 10, pp. 1348–50, Oct. 1983.
- [3] G. Lichtenberger and R. J. Toohill, "Endo-extralaryngeal suture technique for endoscopic lateralization of paralyzed vocal cords," *Oper. Tech. Otolaryngol. Neck Surg.*, vol. 9, no. 3, pp. 166–171, Sep. 1998.
- [4] G. Lichtenberger, "Reversible immediate and definitive lateralization of paralyzed vocal cords.," *Eur. Arch. Otorhinolaryngol.*, vol. 256, no. 8, pp. 407–11, 1999.
- [5] N. C. Jacobson BH, Johnson A, Grywalski C, Silbergleit A, Jacobson G, Benninger MS, "The Voice Handicap Index (VHI): development and validation.," *Am J Speech Lang Pathol*, 6 66–70., no. 6, pp. 66–70, 1997.
- [6] S. A. R. Nouraei *et al.*, "Sensitivity and responsiveness of the medical research council dyspnoea scale to the presence and treatment of adult laryngotracheal stenosis," *Clin. Otolaryngol.*, vol. 33, no. 6, pp. 575–580, Dec. 2008.
- [7] S. A. R. Nouraei *et al.*, "Validation of the Clinical COPD Questionnaire as a psychophysical outcome measure in adult laryngotracheal stenosis," *Clin. Otolaryngol.*, vol. 34, no. 4, pp. 343–348, Aug. 2009.

- [8] D. M. Cheney, M. Tausif Siddiqui, J. K. Litts, M. A. Kuhn, and P. C. Belafsky, "The ability of the 10-item eating assessment tool (EAT-10) to predict aspiration risk in persons with dysphagia," *Ann. Otol. Rhinol. Laryngol.*, 2015.
- [9] S. Leitersdorfer, G. Lichtenberger, and I. Kovacs, "Assessment of the results of glottis-dilating operations using lung function tests.," *Eur. Arch. Otorhinolaryngol.*, vol. 259, no. 2, pp. 57–9, Feb. 2002.
- [10] S. Leitersdorfer, G. Lichtenberger, A. Bihari, and I. Kovács, "Evaluation of the lung function test in reversible glottis-dilating operations.," *Eur. Arch. Otorhinolaryngol.*, vol. 262, no. 4, pp. 289–93, Apr. 2005.
- [11] J. C. Acres and M. H. Kryger, "Clinical significance of pulmonary function tests: Upper Airway Obstruction*," *Chest*, vol. 80, pp. 207–211, 1981.
- [12] N. Sapundzhiev *et al.*, "Surgery of adult bilateral vocal fold paralysis in adduction: History and trends," *European Archives of Oto-Rhino-Laryngology*. pp. 1501–1514, 2008.
- [13] G. Lichtenberger, "Comparison of endoscopic glottis-dilating operations," *Eur Arch Otorhinolaryngol*, vol. 260, pp. 57–61, 2003.

SELECTIVE ELECTRICAL SURFACE STIMULATION IN UNILATERAL VOCAL FOLD PARALYSIS (UVFP)

B. Schneider-Stickler¹, M. Leonhard¹, M. Krenn², W. Mayr²

¹ Dept. of Otorhinolaryngology

² Center for Medical Physics and Biomedical Engineering

Medical University of Vienna, Austria

berit.schneider-stickler@meduniwien.ac.at

matthias.leonhard@meduniwien.ac.at

matthias.krenn@meduniwien.ac.at

winfried.mayr@meduniwien.ac.at

Abstract:

Introduction: Selective electrical surface stimulation (SES) of the larynx is not yet routinely considered as therapy option in UVFP. Goal of this monocentric feasibility study was to provide systematic data on selective electrical stimulation of intrinsic laryngeal muscles in UVFP by using surface electrodes.

Patients and methods: Twenty-seven patients diagnosed with UVFP have been enrolled into the study. As external stimulator, a stimulette r2x by Schuhfried Medizintechnik has been used for SES. Sensitivity threshold, selective laryngeal response (vocal fold (VF) adduction in rest and shortening of the VF during phonation as confirmed in flexible laryngostroboscopy), and non-selective sensations (swallowing reflex, discomfort/pain, extrinsic laryngeal muscle activities) have been documented. A protocol with varying pulse widths and pulse amplitudes has been applied.

Results: Increased current intensity is required to make the stimulation perceivable by the patient when the applied pulse width decreases. Both ailing and healthy VFs were simultaneously stimulated for adduction in 70.4% of the cases at 100 and 250 msec with an average current intensity of about 9 mA. An audible change of frequency due to contracting twitching of VF was notable in 70.4% of the cases at impulse intensity of about 8.1 mA. In 21.6% of the cases no selective laryngeal stimulation with SES was achievable, as nonselective side effects or discomfort/pain limited further increase of pulse amplitude.

Conclusion: The SES should be considered as routine therapy option for patients with early UVFP, as the study could proof positive laryngeal response in almost three quarters of the patients.

Keywords: *electrotherapy, superficial electrotherapy, unilateral vocal cord paralysis, selective laryngeal stimulation, transcutaneous electrical nerve stimulation*

I. INTRODUCTION

UVFP treatment approaches generally aim to restore glottal closure for phonation and improve vocal function. So far, standard treatment in UVFP involves either surgical intervention, voice therapy or a combination of the two (1). Voice exercises in early UVFP alone may work, but it remains questionable as to whether there is a therapeutic effect beyond spontaneous regeneration (2). Surgical approaches are different techniques of injection laryngoplasty, external vocal fold medialization (thyroplasty type I), and reinnervation techniques. Selective electrical surface stimulation (SES) of the larynx is not yet routinely considered as standard therapy in UVFP.

Although SES has been widely applied for treating several types of voice disorders like muscle tension disorders (3, 4), benign focal fold lesion (5), and presbyphonia (6), only few papers reported on its effectiveness for treatment in UVFP patients (7-9). Ptok et Strack concluded after a 3-month therapy of 90 patients with UVFP (onset between 2 weeks and 6 months prior therapy) receiving either voice therapy or SES, that electrical stimulation-supported voice exercises are the method of choice for UVFP patients (8). Perez et al. applied SES in chronic UFVP with onset between 10 to 24 months before therapy (9). The authors did not focus in nerval regeneration but on improvement of voice quality with less irregularity of VF vibration and longer maximum phonation time.

SES in early UVFP focusses first of all on prevention of atrophy of the denervated muscle, the stimulation of the nerval regeneration process and the prevention of spontaneous activity like fibrillation (8).

Goal of this monocentric feasibility study was to provide systematic data on selective electrical stimulation of intrinsic laryngeal muscles in UVFP by using surface electrodes.

II. PATIENTS AND METHODS

The study has been approved by the Ethic Committee of the Medical University of Vienna (EK 2046/2017). So far, 27 patients diagnosed with UVFP have been enrolled into the study: 11 patients were female, and 16 patients were male. The mean age of the patients was $53.7y \pm 15.5y$ [(min=26y; max=75y)]. In 15 patients the right VF and in 12 patients the left VF was affected. In 74% the etiology of UVFP was iatrogenic after surgery, in 11% idiopathic, in 4% cancer, in 4% traumatic and in 7% of other origin. The mean time interval between onset of UVFP and SES examination was 5.7 ± 13.9 months (min=1 mo, max=73 mo).

A stimulette r2x by Schuhfried Medizintechnik has been used as external stimulator for SES. The electrodes were placed bilaterally between the upper and the lower part of the thyroid (fig.1). The test protocol comprised pulse widths between 1 to 500 msec with individually adjustable pulse intensity. For each pulse width following parameters were analyzed:

- sensitivity threshold
- VF response during rest and relaxed breathing
- VF response during phonation
- nonselective response (swallowing reflex, activation of platysma and strap muscles, coughing, pain).

The VF response has been examined by transnasal flexible laryngostroboscopy.

Fig.1. Test situation with surface electrodes positioned between upper and lower edge of the thyroid ala



III. RESULTS

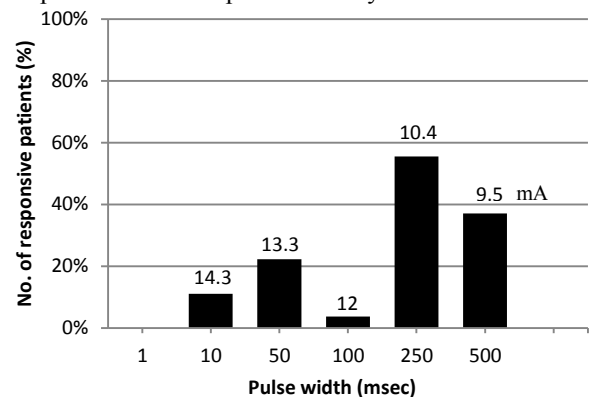
The values for sensitivity threshold are demonstrated in Table 1. As it is to be seen, an increased current intensity seems to be required to make the stimulation perceivable by the patient when the applied pulse width decreases.

Table 1. Sensitivity threshold in SES

Pulse width (msec)	No. of responsive patients	Mean (SD) in mA	Range (mA)
1	21/21	4.9 (1.8)	2.0-10.0
10	23/23	3.6 (2.2)	1.0-10.0
50	27/27	4.1 (3.1)	1.0-14.0
100	27/27	3.2 (1.8)	1.0-7.5
250	27/27	3.2 (1.9)	1.0-10.0
500	27/27	2.7 (1.4)	0.5-7.0

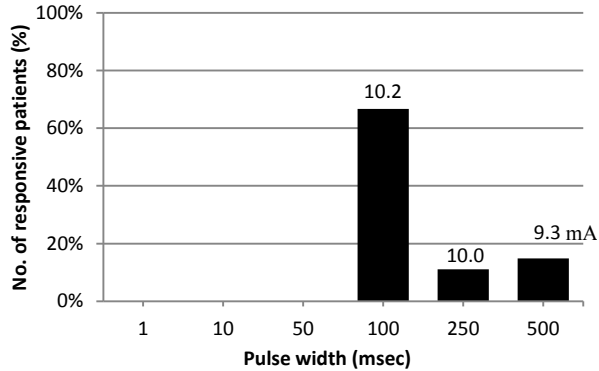
Concerning nonselective side effects, the pulse intensity of the different pulse widths, like eventual swallowing reflex (fig.2), cough reflex (fig. 3), platysma activity or strap muscle response (fig. 4) was evaluated before testing selective laryngeal stimulation effects.

Fig. 2. Swallowing reflex induced by SES depending on pulse widths and pulse intensity.



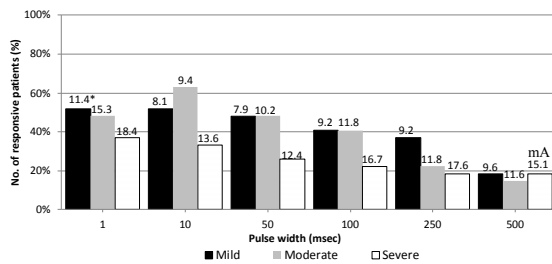
About 55% of the cases responded with a swallow reflex at a pulse width of 250 msec, at other pulse widths the risk of swallow reflex occurred much less frequently.

Fig. 3. Cough reflex induced by SES depending on pulse width and pulse intensity



Cough reflex was observed most often at a pulse width of 100msec as seen in about 65% of the cases when applied with an average current intensity > 10 mA. Swallowing and cough reflexes were always observed with current intensity higher than that was necessary to observe a VF response at rest and during phonation.

Fig. 4. Nonselective response of strap muscles depending on pulse width and pulse intensity



Nonselective muscle responses occurred more often during stimulation with shorter pulse widths- An adduction of the healthy and the ailing VF, as in this experimental setting both ailing and healthy vocal fold were simultaneously stimulated, has been taken as a proof of positive selective laryngeal stimulation. It was achieved in 70.4% of the evaluated cases (19/27) at 100 and 250 msec with an average current intensity of about 9 mA (table 2). This result was less reproducible with longer or shorter pulse widths. It was almost never achievable with PWs < 50 msec.

Table 2. Selective VF response (adduction) at rest under SES

Pulse width (msec)	No. of responsive patients	Mean (SD) in mA	Range (mA)
1	0/27	none	none
10	3/27	6.7 (0.6)	6-7
50	17/27	6.8 (2.0)	
100	19/27	9.1 (5.1)	4-27
250	19/27	8.8 (4.4)	4-25
500	15/27	9.2 (5.8)	3-28

Table 3. Selective VF response (twitching/shortening) during phonation under SES

Pulse width (msec)	No. of responsive patients	Mean (SD) in mA	Range (mA)
1	1/27	12	none
10	1/27	8	none
50	16/27	7.4 (2.8)	5-16
100	19/27	8.1 (1.9)	4-12
250	14/27	8.6 (2.6)	5-15
500	11/27	9.9 (6.4)	4-28

As seen in table 3, the optimum pulse width for selective

laryngeal stimulation during phonation as again 100 msec. With this pulse width again 70.4 % (19/27) patients responded with a noticeable frequency change by VF twitching. While pulse widths of 50 and 250 msec could also achieve evident selective laryngeal stimulation in more than half of the patients, even shorter or longer pulse width could achieve a positive response only in in few individual cases.

IV. DISCUSSION

This monocentric feasibility study was designed to examine the general possibility to use SES in patients with early UVFP.

The study results reveal that a selective laryngeal stimulation can be achieved in patients with early UVFP by using surface electrodes without any invasive procedure. Pulse widths of 100 and 250 msec could

induce selective laryngeal response by either adduction or twitching/shortening of VFs in more than 70 % of the patients tested without any considerably limiting and disturbing side effect like coughing or swallow reflex, or nonselective muscle response. At this point should be noted, that it is important to place the surface electrodes directly in front of the thyroid cartilages to avoid that the current stimulates the surrounding strap muscles and to reduce the swallow and cough reflex (10).

As we could show, the most effective stimulation parameters lie beneath the discomfort threshold and are compatible with the use of a common external stimulator for home-training.

As not only an adduction of the vocal fold but also a audible frequency change due to VF twitching could be achieved, it can be postulated that it seems to be possible to prevent atrophy of the denervated muscle until hopefully a regeneration process will start.

Gugatschka et al could show an increase in muscle volume in both an animal and human studies. A direct electrical stimulation with a stimulation electrode placed near the adduction branch of the recurrent laryngeal nerve could significantly increase the volume of the M. thyroarytenoideus in a model with elderly ovines (11). Recently, Feiner et al reported on SES in a small cohort of elderly women with individually varying results and the conclusion of the necessity of individual fitting (6). The authors presented the highest percentage of positive reactions if pulse widths of 50 msec and shorter have been applied. These results are in contrast to our studies, as in patients with early UFVP best positive response could be achieved with 100 and 250 msec. Our data are more in accordance with Ptok et al. who recommended the use of exponential currents with a 240 msec duration (8).

Taking into consideration the results of this study, an individual fitting of the SES is recommended in every patient with UVFP to ensure the selective effectiveness of the current parameters under consideration of nonselective side effects.

The long term effect of the SES regarding the influence on the neural regeneration process needs to be investigated in further studies.

V. CONCLUSION

The SES can stimulate selective laryngeal activity in a significant majority of patients with early UVFP. Thus, the SES should be considered as routine therapy for this patient group.

VI. ACKNOWLEDGEMENT

The authors would like to thank the company Med-El Medizintechnik GmbH for their support.

REFERENCES

- (1) Walton C, Carding P, Flanagan K. Perspectives on voice treatment for unilateral vocal fold paralysis. *Curr Opin Otolaryngol Head Neck Surg.* 2018 Jun;26(3):157-161.
- (2) Ptok M, Strack D. [Therapeutic effects of electrical stimulation therapy on vocal fold vibration irregularity]. *HNO.* 2009 Nov;57(11):1157-62.
- (3) Mansuri B, Torabinejad F, Jamshidi AA, Dabirmoghaddam P, Vasaghi-Gharamaleki B, Ghelichi L. Transcutaneous electrical nerve stimulation combined with voice therapy in women with muscle tension dysphonia. *J Voice.* 2018 Dec 7. pii: S0892-1997(18)30360-6.
- (4) Mansuri B, Torabinezhad F, Jamshidi AA, Dabirmoghaddam P, Vasaghi-Gharamaleki B, Ghelichi L. Application of high-frequency transcutaneous electrical nerve stimulation in muscle tension dysphonia patients with the pain complaint: The immediate effect. *J Voice.* 2019 May 8.
- (5) Siqueira LTD, Ribeiro VV, Moreira PAM, Brasolotto AG, de Jesus Guirro RR, Alves Silverio KC. Effects of transcutaneous electrical nervous stimulation (TENS) associated with vocal therapy on musculoskeletal pain of women with behavioral dysphonia: A randomized, placebo-controlled double-blind clinical trial. *J Commun Disord.* 2019 Jul 30;82:105923.
- (6) Feiner M, Gerstenberger C, Mayr W, Hortobagyi D, Gugatschka M. Exploring stimulation patterns for electrical stimulation of the larynx using surface electrodes. *Eur Arch Otorhinolaryngol.* 2019 Aug 14.
- (7) Ptok M, Strack D. [Therapeutic effects of electrical stimulation therapy on vocal fold vibration irregularity]. *HNO.* 2009 Nov;57(11):1157-62.
- (8) Ptok M, Strack D. Electrical stimulation-supported voice exercises are superior to voice exercise therapy alone in patients with unilateral recurrent laryngeal nerve paresis: results from a prospective, randomized clinical trial. *Muscle Nerve.* 2008 Aug;38(2):1005-11.
- (9) Garcia Perez A, Hernández López X, Valadez Jiménez VM, Minor Martínez A, Ysunza PA. Synchronous electrical stimulation of laryngeal muscles: an alternative for enhancing recovery of unilateral recurrent laryngeal nerve paralysis. *J Voice.* 2014 Jul;28(4):524.e1-7.
- (10) Ludlow CL, Humbert I, Saxon K, Poletto Ch, Sonies B, Crujido L. Effects of Surface Electrical Stimulation Both at Rest and During Swallowing in Chronic Pharyngeal Dysphagia. *Dysphagia.* 2007 Jan; 22(1): 1–10.
- (11) Gugatschka M, Jarvis JC, Perkins JD, Bubalo V, Wiederstein-Grasser I, Lanmüller H, Gerstenberger C, Karbiener M. Functional Electrical Stimulation Leads to Increased Volume of the Aged Thyroarytenoid Muscle.

Laryngoscope. 2018 Dec;128(12):2852-2857.

ULTRAHIGH RESOLUTION OPTICAL COHERENCE TOMOGRAPHY FOR DETECTING TISSUE ABNORMALITIES OF THE ORAL AND LARYNGEAL MUCOSA: A PRELIMINARY STUDY

Niels Møller Israelsen¹, Mikkel Jensen¹, Anders Overgård Jønsson², Mette Pedersen³

¹Department of Photonics Engineering, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark

²Faculty of Health, University of Copenhagen, Denmark ³Hon. Prof. IBC Cambridge, United Kingdom

¹nikr@fotonik.dtu.dk ²hbk777@alumni.ku.dk ³m.f.pedersen@dadlnet.dk

Abstract: Optical coherence tomography (OCT) is an imaging technology that provides cross-sectional images without biopsy of subsurface tissue structure at approximately 10 micrometer resolution to a depth of 1.5 mm using backscattered light from the tissue. OCT has shown promising in imaging normal upper airways and in various disorders. The use of OCT to image the upper airways during diagnosis and treatment of a vast array of disorders continues to develop along with innovative surgical techniques, of patients in anesthesia.

Monitoring the benign cellular and molecular events resulting in e.g. edema, in the clinic online is of great interest. There is a need to get information in the clinic about the normal histology of functioning mucosa compared with disorders causing edema especially at the inflammatory process level. The main interesting point is to find out whether it is related to infection, acid reflux, allergy or a fourth condition. Long-range OCT using Doppler OCT is providing useful clinical applications for diagnostic and therapeutic laryngeal procedures of the vocal folds (1). Till now a probe for oral examination of the mucosa with ultrahigh resolution OCT has been established.

Keywords: In vivo optical coherence tomography, upper airways, mucosa, OCT

I. INTRODUCTION

Optical coherence tomography (OCT) is an imaging technology that provides cross-sectional images without biopsy of subsurface tissue structure at approximately 10 micrometer resolution to a depth of more than 1.5 mm using backscattered light from the tissue. OCT has shown promising in imaging normal upper airways and in various disorders. The use of OCT to image the upper airways during diagnosis and treatment of a vast array of disorders continues to develop [1], along with innovative surgical techniques, of patients in anesthesia, eventually combined with confocal endomicroscopy [2].

A first full field highspeed and long-range OCT in vivo of the vocal folds was made recently [3]. Cross sectional images during phonation have also been

made [4]. In 1997 the epithelium, lamina propria and submucosa in the larynx were described with OCT in the larynx [5], and comparison of OCT pictures verified with histology based on biopsies have been made. Endonasal approach has been made to the larynx as well as automatic working adjustments to the working area [6].

Monitoring the benign cellular and molecular events resulting in e.g. edema, in the clinic online is of great interest. There is a need to get information in the clinic about the normal histology of functioning mucosa compared with disorders causing edema especially at the inflammatory process level. The main interesting thing is to find out whether it is related to infection, acid reflux, allergy ect.

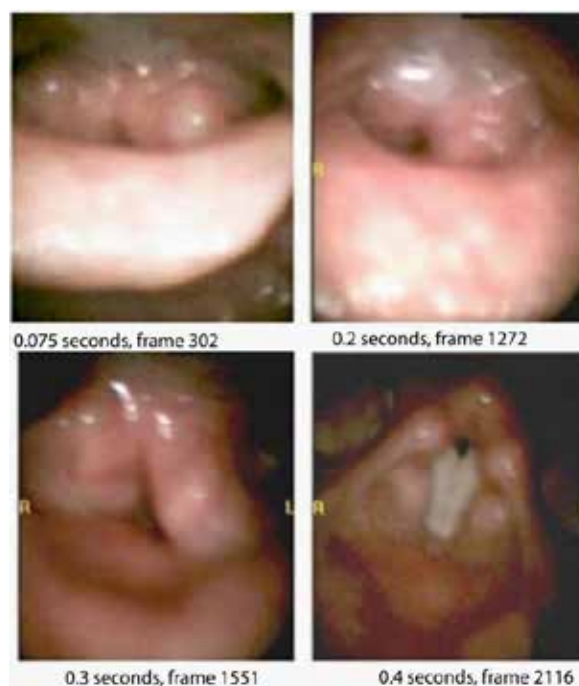


Fig 1. Frames taken from high-speed video set on recording 4000 pictures per second. The picture, at frame 1272 shows the appearance of mucus in the larynx, and shortly after (frame 1551) it is reduced. On frame 2116, the mucus has disappeared. The process took a total of 0.2 seconds [14].

Long-range OCT using Doppler OCT is providing useful clinical applications for diagnostic and therapeutic laryngeal procedures of the vocal folds. Till now a probe capable of oral examination of the mucosa with the ultrahigh resolution (UHR) OCT has been established in our laboratory to be combined with highspeed films of the upper airways. And we can show some preliminary pictures of the oral mucosa.

II. METHODS and III. RESULTS

The OCT system applied in this study is an ultrahigh resolutions system constructed in-house. The setup is based on a supercontinuum source from NKT Photonics providing a large near-infrared band centered at approximately 1.3 micrometers. The light is forwarded to a beam splitter after which a sample and a reference mirror are exposed. The returning combined light signal is relayed to a spectrometer (Wasatch Photonics) detecting interferometric signals in a wavelength range of 1.074-1478 micrometer. The spectra detected are resampled, dispersion compensated, and attenuation corrected. The system demonstrates an axial (depth) resolution below 3 micrometers, a lateral resolution of 6 micrometer, a sensitivity of 90 dB and a line rate of 76 kHz. The imaging depth and maximum field of view are 2 mm and 8 mm, respectively. In order to interface the lining oral mucosa, a handheld probe was a necessity. The in-house constructed probe is described in [7, 8]

For the larynx a probe has earlier been developed through the nose and orally also in laboratories for online use of OCT without ultrahigh resolution. The new aspect is a probe trans orally where ultrahigh resolution OCT can be combined with our highspeed films of 4000 pictures per second (Richard Wolf GmbH) which we have used clinically for nearly 10 years. The highspeed films are only showing the surface of the laryngeal structures, but in pathology a combination with ultrahigh resolution OCT will give more information about the mucosal function also in the larynx.

Ultrahigh resolution OCT images presented, show the oral lining mucosa in depth. B-scans show fine details of the transition from the epithelium to the lamina propria with information of blood vessel morphology. Three scans are presented to represent the information content provided by ultrahigh resolution OCT.

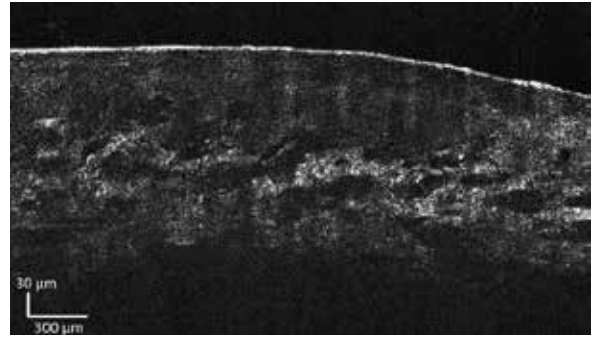


Fig 2. *In vivo optical coherence tomography of the oral mucosa of the inside of the lower lip. Scalebars for reference.*

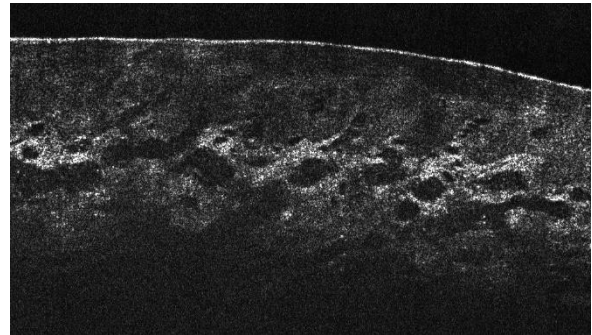


Fig 3. *In vivo optical coherence tomography of the oral mucosa of the inside of the lower lip. This area is rich in blood vessels.*

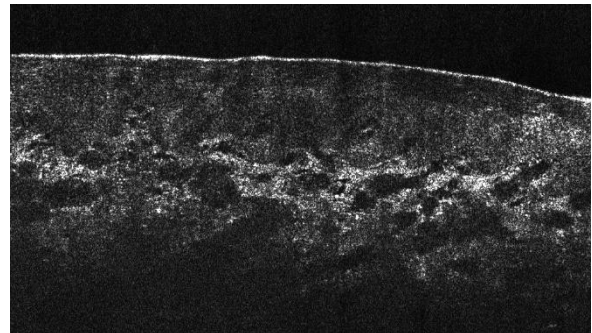


Fig 4. *In vivo optical coherence tomography of the oral mucosa of the inside of the lower lip.*

IV. DISCUSSION and V. CONCLUSION

It is possible to make quantitative assessment of the structure characteristics, e.g. fibrosis, atrophy, tissue inflammation, capillary loop density, vessel morphology as well as glands, and changes during treatment online with this setup. This means that the OCT method can document tissue changes with pharmacological and other medical interventions [9]. Another interesting aspect is development of the epithelium of the upper airways in childhood [10], especially of the vocal folds during puberty, where the hormonal cellular effect lacks understanding [11, 12, 13].

Documentation of pathology in benign disorders of the upper airway in the clinic is insufficient. There is a lack of differentiation possibilities for treatment documentation. The mucosa function is also of extreme interest to document pathology in singers [3,10]. Ultrahigh resolution OCT is a promising candidate for detecting tissue abnormalities of the upper airway mucosa in the clinic. Especially irregularities related to the junction between the epithelium and lamina propria and the thickness of epithelium is easily delineated by this significantly aid diagnosis also in other mucosal regions. Quantitative diagnoses are of interest with this ultrahigh resolution OCT setup in the arytenoid region behind the vocal fold for laryngopharyngeal reflux, and the vocal folds in cases where direct laryngoscopy is without indication. Evidential documentation is in demand in order to correlate online (often acute) laryngeal disabilities, with highspeed films [14]. Another interesting aspect is the development of mucosa especially on the vocal folds during puberty, an online approach with presentation of the throat mucosa will help us with diagnosis and treatment [15, 16].

REFERENCES

- [1] Pedersen M, Agersted A, Akram B, Mahmood S, Jønsson A.O, and Mahmood S.. "Optical Coherence Tomography in the Laryngeal Arytenoid Mucosa for Documentation of Pharmacological Treatments and Genetic aspects: a Protocol" *Advances in Cellular and Molecular Otolaryngology* 2016, 4, (1): 32246.doi:3402/acmo.v4.32246
- [2] Just T, Guder E, Witt G, Ovari A, Stüpnagel B v, Lankelau E, Prall F, Hüttmann,G, Pau H.W,"Confocal Endomicroscopy and Optical Coherence Tomography for Differentiation Between Low-Grade and High-Grade Lesions of the Larynx" in *Biomedical Optics in Otorhinolaryngology: Head and Neck surgery*. Eds.Springer New York, 2016, pp 479-490
- [3] Chen I, Sharma G.K, Badger C, Hong E, Chou L-d, Rangarajan S, Chang T.H, Cho K, Lee D, Goddard J.A, Chen Z, Wong B.J F " Quantitative motion analysis of optical coherence tomography images in vivo human folds" *Presentation, World Voice Consortium Conference Copenhagen* 2017
- [4] Coughlan C.A, Chou L-d, Jing J.C, Chen J.J, Rangarajan S, Chang T.H, Sharma G.K, Cho K, Donghoon L, Goddard J.A, Chen Z, Wong B.J.F "In vivo cross-sectional imaging of the phonating larynx using long-range Doppler optical coherence tomography" *Scientific reports* 2016, pp. 1-8: 6:22792 DOI:10.1038/srep22792
- [5] Sergeev A.M, Gelikonov G.V, Gelikonov F.I, Feldchtein R.V, Kuranov R.V, Gladkova N.D "In vivo endoscopic OCT imaging of precancer and cancer states of human mucosa," *OPTICS EXPRESS*, vol 1,13 1997, pp 434-440
- [6] Donner S, Bleeker S, Ripken T, Ptok M, Jungheim M, Krueger A "Automated working distance adjustment enables optical coherence tomography of the human larynx in awake patients", *J Med imaging* (Bellingham, Wash.) vol 2,2 2015, pp 026003
- [7] Israelsen NM, Maria M, Mogensen M, Bojesen S, Jensen M, Haedersdal M, Podoleanu A, Bang O "The value of ultrahigh resolution OCT in dermatology - delineating the dermo-epidermal junction, capillaries in the dermal papillae and vellus hairs", *Biomedical Optics Express*, 2018, 9(5), pp. 2240-2265, <https://doi.org/10.1002/jbio.201700348>
- [8] Mogensen M, Bojesen S, Israelsen NM, Maria M, Jensen M, Podoleanu A, Bang O, Haedersdal M, "Two optical coherence tomography systems detect topical gold nanoshells in hair follicles, sweat ducts and measure epidermis", *Journal of Biophotonics*, 2018, 11(9), <https://doi.org/10.1002/jbio.201700348>
- [9] Wei W, Choi WJ, Men S, Song S, Wang RK "Wide-field and long-ranging-depth optical coherence tomography microangiography of human oral mucosa (Conference Presentation)", *Proceedings SPIE* 10473, *Lasers in Dentistry XXIV*, 2018, 104730H
- [10] Gracia JA, Benboujja F, Beaudette K, Guo R, Boudoux C, Hartnick C.J "Using attenuation coefficients from optical coherence tomography as markers of vocal fold maturation", *Laryngoscope* 2016 E218-E223 doi.org/10.1002/lary.25765
- [11] Benboujja F, Hartnick C. "Clinical and surgical implication of intraoperative optical coherence tomography imaging for benign pediatric vocal fold lesions" *Int J Pediatr Otorhinolaryngol.* 2018, 114, pp 111-119. doi: 10.1016/j.ijporl.2018.08.036.
- [12] Nourmahnad A, Benboujja F, Hartnick C.J. "Using intraoperative optical coherence tomography to image pediatric unilateral vocal fold paralysis." *Int J Pediatr Otorhinolaryngol.* 2019, 121, pp 72-75. doi: 10.1016/j.ijporl.2019.02.039.
- [13] M. Pedersen, *Normal Development of Voice in Children*, 1st ed. Berlin-Heidelberg: Springer-Verlag, 2008.
- [14] Pedersen M, Eeg M, Laryngopharyngeal Reflux – A Randomized Clinical Controlled Trial, *Otolaryngology*, 2012. S1:004. doi: 10.4172/2161-119X.S1- 004. pp. 1-5.
- [15] Pedersen M, Mahmood S, Jønsson A, Mahmood MS, Akram BH, Agersted AA. Functional Examination of Voice, a Review. 2016. *Health Science Journal*, Vol.10 No.4:18. DOI: 10.4172/1791-809X.1000100420
- [16] Pham TT, Chen L, Heidari AE, Chen JJ, Zhukhovitskaya A, Li Y, Patel U, Chen Z, Wong B.J.F. "Computational analysis of six optical coherence tomography systems for vocal fold imaging: A

comparison study." *Lasers Surg Med.* 2019. doi:
10.1002/lsm.23060.

SESSION IX
KEYNOTE LECTURE

KEYNOTE LECTURE

DEVELOPING NEW SPEECH SIGNAL PROCESSING ALGORITHMS FOR BIOMEDICAL AND LIFE SCIENCES APPLICATIONS: PRINCIPLES, FINDINGS, CHALLENGES, AND A VIEW TO THE FUTURE

Athanasios Tsanas

Usher Institute of Population Health Sciences and Informatics
Medical School, University of Edinburgh and University of Oxford

Athanasios.Tsanas@ed.ac.uk

Summary: I will briefly outline the main physiological principles of voice production and describe how these link to the key concepts for developing speech signal processing algorithms to characterize speech and extract potentially useful information. I will demonstrate the applicability and differences of speech signal processing algorithmic concepts across different applications, in combination with state of the art

statistical machine learning techniques. Finally, I will touch on open questions, challenges, and upcoming problems as we develop robust, parsimonious, generalizable decision support tools mining speech signals across diverse biomedical and life sciences applications.

SESSION X
BIOMECHANICS/DEVICES

EXPERIMENTAL MODELLING OF GLOTTAL AREA DECLINATION RATE IN VOWEL AND RESONANCE TUBE PHONATION

J. Horáček¹, V. Radolf¹, V. Bula¹, A. M. Laukkanen²

¹ Institute of Thermomechanics of the Czech Academy of Sciences, Prague, Dolejškova 1402/5, 182 00 Prague, Czech Republic

² Speech and Voice Research Laboratory, Faculty of Social Sciences, Tampere University, Virta, Åkerlundinkatu 5, 33100 Tampere, Finland

jaromirh@it.cas.cz, radolf@it.cas.cz, bula@it.cas.cz, Anne-Maria.Laukkanen@tuni.fi

Abstract: The aim of this study was to investigate how the maximum area declination rate (MADR) corresponds to the maximum velocity of the vocal folds (VFs) just before collision. The measurements performed for phonation on vowel [u:] were compared with those obtained during phonation through a glass resonance tube with the distal end in air or submerged 10 cm in water.

The results show that the glottal area declination rate just before the glottal closure is substantially lower than the MADR, and also lower in phonation through a tube into air and into water than in vowel phonation. Thus, the results suggest the impact stress in tube therapy is lower than in vowel phonation.

Keywords: glottal closing, velocity, MADR, impact stress, voice therapy by using tubes.

I. INTRODUCTION

The maximum area declination rate (MADR) in the closing phase of the glottis during vocal folds' (VFs) vibration has been considered as a measure of the impact stress loading the VFs during collision [1] and, thus, an important quantity in studying vocal economy of phonation [2].

II. AIMS

The present study compared waveforms of glottal area variation, and glottal area declination rate, and MADR for phonation on vowel [u:] with those obtained during phonation through a glass resonance tube with the distal end in air or submerged 10 cm in water, simulating two widely used voice therapy methods [3, 4]. The aim was to investigate how well MADR corresponds to the maximum velocity of VFs just before collision.

III. METHODS

The experiments were performed on a physical model of phonation [5] using mean airflow rates $Q = 0.06 - 0.09$ L/s, which corresponded to mean subglottic

pressures $P_{sub} = 730 - 860$ Pa for simulated phonation with [u:]-vowel-shaped vocal tract (VT), $P_{sub} = 670 - 970$ Pa for phonation with [u:]-vowel-shaped VT attached to a glass resonance tube (length 27 cm, inner diameter 7.8 mm) ending in air, and $P_{sub} = 1.44 - 1.75$ kPa for phonation with [u:]-vowel-shaped VT attached to the same tube ending 10 cm in water. The fundamental phonation frequency of the VFs self-oscillation was in the range $F_0 = 90 - 93$ Hz. The values for P_{sub} , Q and F_0 were, thus, within physiologically relevant values for male voice production. The air was flowing from a model of the lungs to a trachea model and then through a 1:1 scaled three layer VFs' model to a vocal tract model [5]. The controlling parameter in measurement was the mean airflow rate Q measured by a float flowmeter (EMKO type DF3-09K5). The pressures, P_{sub} and P_{oral} , were registered with integrated pressure semiconductor sensors (NXP Freescale MPXV5010GC6U) and recorded by using the measurement system B&K Pulse. A high-speed CCD camera (NanoSense MkIII, maximum resolution 1280 x 1024 pixels) with a camera zoom lens (Nikon ED, AF NIKKOR, 70-300 mm, 1:4-5.6 D) were included in the measurement set up for investigating VFs vibration. The rate of 10 000 fps was used with the maximum possible resolution (548 x 104 pixels). For the purpose of this paper, the glottal area GA was evaluated in a middle cross-section of the VFs in the bandwidth of 21 pixels, centered at maximum amplitude of the glottal width, and where the edges of the glottis were vibrating approximately parallel creating a rectangular shape of the area, see Fig. 1.

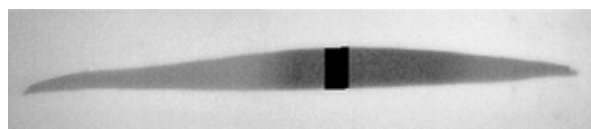


Fig. 1. Example of the glottal area region GA (the black part) evaluated in a middle cross-section of the self-oscillating VFs model.

IV. RESULTS

The waveforms of the measured glottal area for nine cycles of the VFs' vibration show that before each glottal closure the MADR value, given by the absolute value of the time derivative of the glottal area ($ldGA(t)/dt$), was followed by *lower values* of glottal area declination rate just before the glottal closure, the latter defined as the time instant where $GA=0$ and $dGA(t)/dt = 0$, see the waveforms example in Fig. 2.

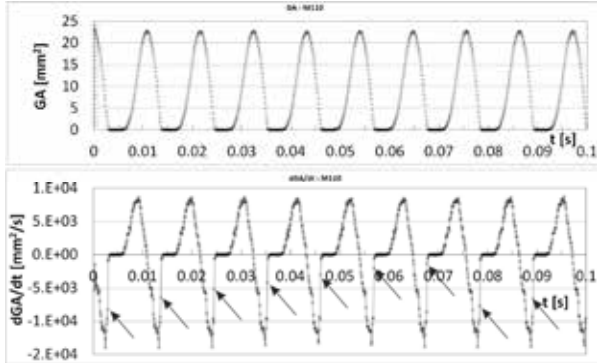


Fig. 2. Example of the glottal area waveform (top), and the area derivative waveform (bottom) measured for phonation on tube with the distal end in air ($Q=0.08$ L/s, $P_{sub}=864$ Pa, $F_0=93$ Hz, first formant frequency $F_1=115$ Hz, first subglottic formant frequency $F_{sub1}=725$ Hz).

The waveform for $dGA(t)/dt$ shows that in the first three cycles after MADR the area declination rate decreases before the VFs collision by ca 40%, 50% and 60% (compared to the MADR magnitudes), in the fourth cycle it decreases by about 70% and in the fifth cycle by ca 90%. The area declination rates before the VFs collision considered in this study are marked in Fig. 2 by arrows.

MADR and the last evaluable absolute values of the area derivative $ldGA(t)/dt|_{before}$, measured just before the glottal area closure, were first evaluated as averaged values from nine recorded periods of VFs vibration for each trial (see nine values of MADR followed by nine values of $ldGA(t)/dt|_{before}$ marked in Fig. 2 by the arrows). The results for each kind of phonation are shown in dependence on the flow rate in Figs. 3 and 4. As expected, MADR increases with the flow rate extensively and approximately linearly, while all values of $ldGA(t)/dt|_{before}$ are considerably lower and keeping similar level for all flow rates and kinds of phonation. Then these averaged values of MADR and $ldGA(t)/dt|_{before}$ were averaged again within the range of the studied air flow rates.

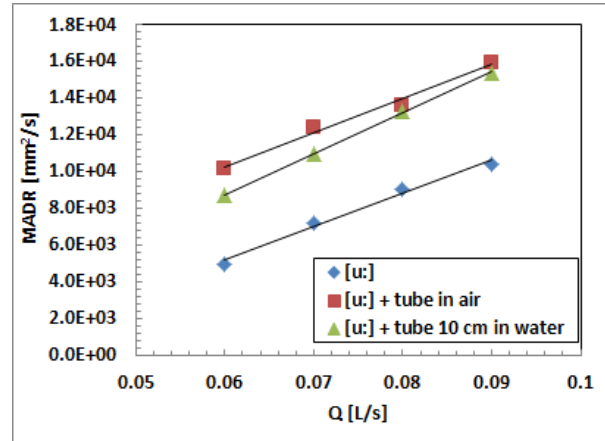


Fig. 3. Measured MADR as a function of the airflow rate for phonation on vowel [u:], [u:] prolonged by tube with distal end in air, and [u:] prolonged by tube with distal end 10 cm in water. The data were evaluated for all three ways of phonation as averaged values from 9 periods of VFs vibration within each flow rate setting.

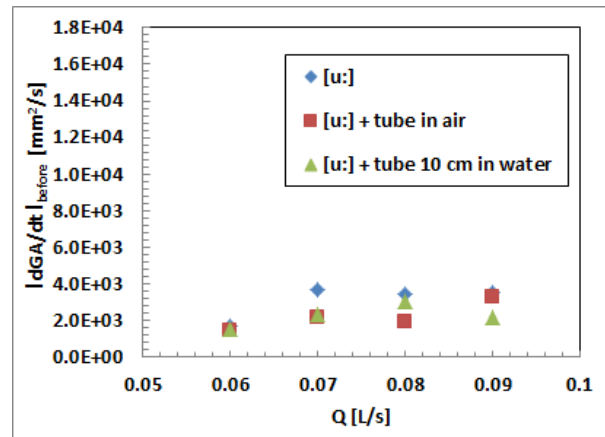


Fig. 4. Measured area derivative just before the glottal area closure as function of the airflow rate for phonation on vowel [u:], [u:] prolonged by tube in air, and [u:] prolonged by tube 10 cm in water. The data were evaluated for all three ways of phonation as averaged values from 9 periods of VFs vibration within each flow rate setting.

The averaged final results are summarized in Table 1. The highest average $MADR=1.3 \cdot 10^4$ mm^2/s was evaluated for phonation into the tube with the distal end in air. For phonation into the tube with the distal end in water MADR was slightly smaller, $1.21 \cdot 10^4$ mm^2/s , and the lowest $MADR=7.92 \cdot 10^3$ mm^2/s was found for vowel phonation. However, the average area derivative just before the glottal closure, $ldGA(t)/dt|_{before}$, was in all cases clearly smaller than MADR. In phonation through the tube into air

$ldGA(t)/dt|_{\text{before}}$ was only 17 % of MADR value, in phonation through the tube into water $ldGA(t)/dt|_{\text{before}}$ was ca 19 % of MADR, and in vowel phonation it was 39 % of MADR. Therefore, the highest average area derivative just before the glottal closure: $ldGA(t)/dt|_{\text{before}} \cong 3.1 \cdot 10^3 \text{ mm}^2/\text{s}$ was observed for vowel phonation, and the lower values were observed for phonation through the tube with distal end in air and in water, $ldGA(t)/dt|_{\text{before}} \cong 2.2\text{-}2.3 \cdot 10^3 \text{ mm}^2/\text{s}$.

Table 1. Averaged values of the MADR and the last evaluable absolute values of the area derivative before the glottal closure $ldGA/dt|_{\text{before}}$. Averaged values were obtained from the values shown in Fig. 2 by averaging data throughout the flow rates Q, R is the ratio of $ldGA/dt|_{\text{before}}$ to MADR and SD is the standard deviation of R.

Phonation	MADR [mm ² /s]	$ldGA/dt _{\text{before}}$ [mm ² /s]	R [%]	SD [%]
[u:]	7.92E+03	3.11E+03	39.3	8.4
[u:] + tube in air	1.30E+04	2.22E+03	17	2.9
[u:] + tube in water	1.21E+04	2.29E+03	18.9	3.9

V. DISCUSSION

Although the results of the glottal area variation were substantially influenced by low sampling frequency and low resolution of the high speed camera used, the results inevitably indicate that MADR is not a sufficient quantity for estimating the impact stress during glottal collision. Moreover, the results for the absolute values of the time derivative of the glottal area measured just before the glottal closure implicate that the impact stress for phonation into tubes is not higher than for phonation on vowel, but rather the opposite, which agrees with the results of previous impact stress measurements on the phonation model [6]. We can note that the place chosen for observation of the glottal area in the middle of the vocal folds model corresponds to the middle of the membranous part of human vocal folds, which is usually critical region for creating vocal fold nodules caused by impact stress, see e.g. [7, 8].

VI. CONCLUSION

The results suggest that the MADR is not a sufficient quantity for estimating impact stress during glottal collision and the glottal area declination rate just prior to VFs collision is lower in phonation through a tube

into air and into water than in vowel phonation and thus, the impact stress in tube therapy is lower than in vowel phonation.

ACKNOWLEDGEMENT

The study was supported by a grant from the Czech Science Foundation: No. 19-04477S “Modelling and measurements of fluid-structure-acoustic interactions in biomechanics of human voice production.”

Imaging by means of camera was supported by the European Regional Development Fund under Grant No. CZ.02.1.01/0.0/0.0/15_003/0000493 (Centre of Excellence for Nonlinear Dynamic Behaviour of Advanced Materials in Engineering).

REFERENCES

- [1] I.R. Titze, “Theoretical analysis of maximum flow declination rate versus maximum area declination rate in phonation,” *Journal of Speech, Language, and Hearing Research*, vol. 49, pp 439-447, 2006.
- [2] I.R. Titze and A.M. Laukkanen, “Can vocal economy in phonation be increased with an artificially lengthened vocal tract? A computer modeling study,” *Logopedics Phoniatrics Vocology*, vol. 32(4), pp 147-156, 2007.
- [3] A. Sovijärvi, “Nya metoder vid behandling av röstrubbningar,” [New methods for treating voice disorders], *Nordisk Tidskrift för Tale of Stemme*, vol. 3, pp. 121–131, 1969.
- [4] A.M. Laukkanen, A. Geneid, V. Bula, V. Radolf, J. Horáček, T. Ikavalko, T. Kukkonen, E. Kankare and J. Tyrmi, “How much loading does water resistance voice therapy impose on the vocal folds? An experimental human study,” *Journal of Voice*, In Press, Available online 22 November 2018, <https://doi.org/10.1016/j.jvoice.2018.10.011>.
- [5] J. Horáček, V. Radolf and A.M. Laukkanen, “Experimental and computational modeling of the effects of voice therapy using tubes,” *Journal of Speech, Language, and Hearing Research*, vol. 62, pp. 2227–2244, 2019.
- [6] J. Horáček, V. Radolf and A.M. Laukkanen, “Impact stress in water resistance voice therapy: A physical modeling study,” *Journal of Voice*, vol. 33(4), pp. 490-496, 2019.
- [7] K. Verdolini, C.A. Rosen, R.C. Branski, eds. “Vocal fold nodules (nodes, singer's nodes, screamer's nodes,” in *Classification Manual for Voice Disorders - I. Psychology Press*, 2014, pp. 37–40. ISBN 978-1-135-60020-4.
- [8] I.R. Titze, “Mechanical stress in phonation,” *Journal of Voice*, vol. 8(2), pp. 99-105, 1994.

DEVELOPMENT AND USE OF AN ANECHOIC SUBGLOTTAL TRACT FOR EXCISED LARYNX EXPERIMENTS

H. Lehoux¹, V. Hampala¹, J. G. Švec¹

¹Voice Research Lab, Department of Biophysics, Faculty of Science, Palacký University, Olomouc, Czech Republic
hugo.lehoux01@upol.cz, vit.hampala@gmail.com, jan.svec@upol.cz

Abstract: During phonation, both the supraglottal and subglottal tracts can have an influence on vocal fold vibrations. This paper proposes a method to remove the influence of the two tracts on the vocal fold vibrations which can be used in studies of inherent vibratory properties of the vocal folds. It uses excised larynges without a supraglottal tract and an anechoic (resonance-free) subglottal tract. The acoustic response of the developed anechoic subglottal tract was measured using a loudspeaker and a microphone placed at its open end. This response was then compared with the response of a “resonant” subglottal tract with adjustable length. Both subglottal tracts were then explored in excised larynx experiments, where the subglottal pressure, radiated sound, and electroglottograph (EGG) signal were recorded.

Keywords: excised larynx experiment, vibration interactions, subglottal pressure, anechoic subglottal tract, acoustic resonance

I. INTRODUCTION

The study of the inherent properties of the vocal folds has always faced the difficulty of accessing the larynx on living subjects. Some techniques, e.g. laryngoscopy or inverse filtering, can be used to overcome this difficulty; however, they do not allow isolating the vocal folds from the supraglottal and subglottal cavities. This can sometimes cause problems: Titze et al. [1], Wade et al. [2] and Zañartu et al. [3] reported occurrences of sudden pitch jumps and instabilities when the fundamental frequency of oscillation crossed the first supraglottal resonance frequency and these phenomena were observed also through mathematical simulations [4]. In addition, Zhang et al. [5] observed, using a physical model of the vocal folds, similar phonation instabilities caused by the coupling between the vocal folds and the subglottal tract.

This paper proposes a method to remove the influence of the supraglottal and subglottal tracts on the

vocal fold vibrations. While the influence of the supraglottal tract can be removed by using excised larynges, a subglottal space is always needed for air supply. Therefore, an anechoic (resonance-free) subglottal tract was developed and tested: its acoustic response was measured and compared to that of a “resonant” subglottal tract [6]. Both subglottal tracts were then explored in excised larynx experiments where the subglottal pressure, the radiated sound, and the electroglottograph (EGG) signal were measured. In those experiments, we investigated the differences in the shape of the subglottal pressure waveform, as well as the influence of the subglottal tract on the phonation instabilities.

II. METHODS

A novel anechoic subglottal tract was developed here. Its design was inspired by the work of Sondhi [7], who aimed to measure the glottal volume flow *in vivo*, without using inverse filtering methods. He extended the supraglottal tract with a straight tube terminated by a conical wedge made of sound absorbing material, greatly reducing the reflection of waves at this end and therefore removing the major acoustic resonances of the vocal tract. This technique allowed measuring the glottal waveform directly with an electret microphone placed inside the extension tube. Our developed subglottal tract, similarly, consists of a long (about 3.30 m) plastic tube with a conical wedge made of sound absorbing material at one end. The wedge prevents the reflection of the sound waves, which cancels the acoustic resonances. The subglottal tract is then seen by any source as an infinite, purely resistive waveguide.

In order to test the efficiency of our system, we compared it to a “resonant” adjustable subglottal tract previously developed by Hampala et al. [6]. It is composed of a straight glass tube of total length 0.55 m. At one end, a smaller outer diameter metal adapter can be fixed, to which a larynx can be attached. The other end is closed by a moveable piston, which can

adjust the effective length of the tract and therefore its resonance frequency.

Prior to the experiments, we measured the acoustic response of this system in order to identify the first resonance frequency and to mark piston positions corresponding to specific values. The first resonance frequency was found to be between 330 Hz (piston all the way down) to 800 Hz (piston all the way up), although these values come from measurements of the tract only, and they can vary slightly when a larynx specimen is attached. The piston was then removed and the anechoic tube was inserted instead, as shown in Figure 1. The system, comprising the 3.30 m plastic tube attached to the 0.55 m glass tube, will be referred to as the “anechoic subglottal tract” and the system comprising the 0.55 m glass tube with the piston attached will be referred to as the “resonant subglottal tract”.

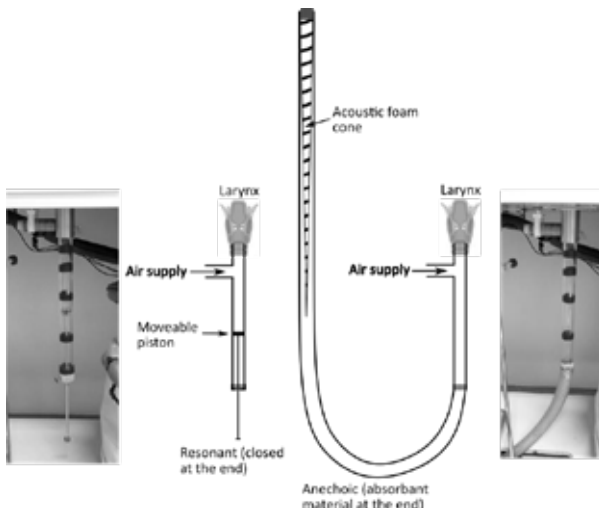


Figure 1: Drawing and photos of the two subglottal tracts: resonant (left) and anechoic (right).

The acoustic responses of both systems were measured using the following protocol:

- 1- A miniature loudspeaker (Ekulit LSF-23M/N/G) was attached, together with a small electret microphone (AV-JEFE, TCM141), to the open end of the system, at which the larynx is usually mounted. Plasterine was used to prevent air leaks.
- 2- Via Audacity software, the speaker generated one hundred impulses at the rate of one single impulse per second.
- 3- The microphone registered the system response from every impulse and the signal was recorded using Audacity software.

The responses of both the anechoic and resonant subglottal tracts were measured. Measurements of the

resonant subglottal tract were done with the piston set to six different positions, corresponding to specific resonance frequencies, and also with the anechoic subglottal tract. The recorded impulse responses were averaged in the time domain (to remove unwanted noise) and Fast Fourier Transforms (FFT) analysis was applied to the averaged time signals. The obtained FFT of the tracts responses were divided by the FFT of the loudspeaker response (without any tract), measured using the same protocol with the same microphone. This extra step was performed to remove the influence of the loudspeaker and microphone frequency response on the measurements. All the computations were done using custom Matlab scripts.

Following the measurement of the subglottal tract characteristics, excised larynx experiments were performed. The excised larynx experimental setup was as follows: an air pump generated a continuous airflow to the larynx, after passing through a heating and humidifying system, artificial lungs, and the subglottal tract. An excised larynx of a red deer was used. Here, the vocal folds were exposed by removing all the unnecessary tissues above them (epiglottis, ventricular folds, and parts of the thyroid cartilage). The larynx was inserted around the open end of the subglottal tract and tightened to prevent air leaks. EGG electrodes (Glottal Enterprise EG2-PC) were attached to the thyroid cartilage using small screws, and the subglottal pressure was measured with a 2.4 mm diameter pressure transducer placed in a hole through the dorsal ridge of the cricoid cartilage. A condenser microphone (MicW M416) placed approximately 5 cm above the vocal folds registered the radiated sound. The signals were sampled at 48 kHz using a DEWE-43 USB data acquisition system and recorded in the associated Dewesoft X2 software.

The first excised larynx experiment consisted of stable phonation through each subglottal tract. Metal pins were used to adduct the vocal folds; the flow was increased progressively up to approximately 400 ml/s, and was reduced back to 0 ml/s after a few seconds of stable phonation. The second experiment focused on phonation instabilities, and especially frequency jumps, caused by the subglottal tract coupled with the vocal folds. To investigate this phenomenon, the fundamental frequency (f_0) of oscillation was gradually varied by manually changing the elongation of the vocal folds. Several frequency increase and decrease cycles were performed, and the whole experiment was repeated three times with each subglottal tract. The resonance frequency of the resonant subglottal tract was set here to the lowest possible value of about 330 Hz.

III. RESULTS AND DISCUSSION

Figure 2 shows the frequency responses of the two subglottal tracts: the solid line represents the frequency response of the anechoic tract, and the other lines correspond to the resonant tract. For simplicity, the results from only two different positions of the piston are shown here (corresponding to 400 Hz and 800 Hz, approximately). From these responses, it is clear that the anechoic subglottal tract effectively removes the acoustic resonances. The small peak around 800 Hz was attributed to a damped resonance caused by a reflection of the air supply tube, as it was present in all measurements.

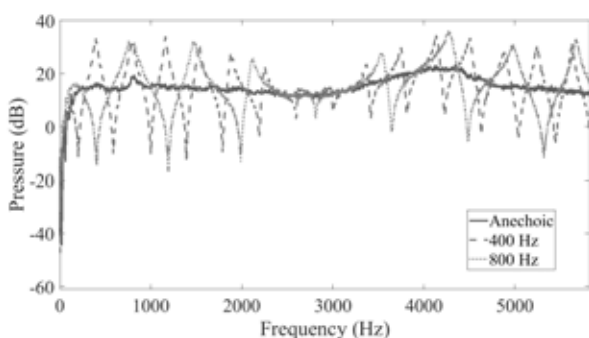


Figure 2: Acoustic responses of the subglottal tracts without a larynx. Solid line: anechoic subglottal tract; dashed line: resonant subglottal tract set to 400 Hz resonance frequency; dotted line: resonant subglottal tract set to 800 Hz resonance frequency.

For stable phonation, the radiated sound signal waveforms showed no apparent dissimilarity. The most significant difference between the two subglottal tracts was found in the shape of the subglottal pressure waveform. With the resonant subglottal tract, the subglottal pressure showed a complex waveform (Figure 3). The synchronized EGG signal was used to locate the timing of the closed and open phases. Visible fluctuations appeared during the closed phase due to subglottal acoustic resonances, similarly as observed previously in the experiments *in vivo* [8-10]. In contrast, in the case of the anechoic subglottal tract, the subglottal pressure waveform closely resembled an inversed theoretical signal of the voice source [7, 11-15]: it was decreasing when the vocal folds were decontacting/opening and increasing when they were contacting/closing, and stayed approximately constant during the contact/closed phase (see Figure 4).

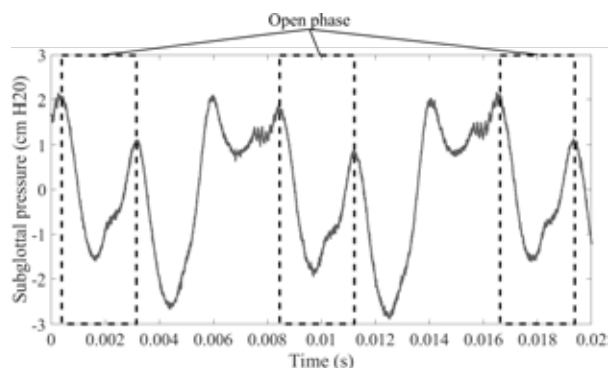


Figure 3: Subglottal pressure waveform obtained during an excited larynx experiment with the resonant subglottal tract set to 400 Hz resonance frequency. The open phase was identified using the EGG signal.

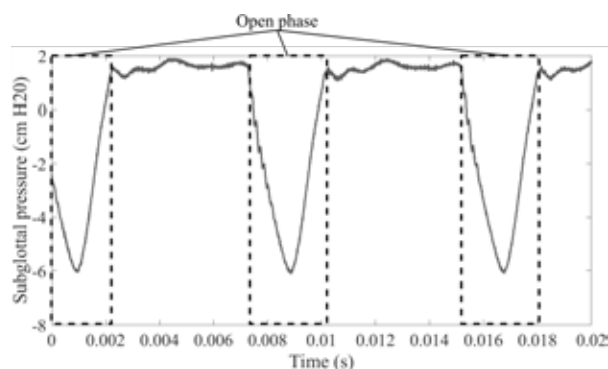


Figure 4: Subglottal pressure waveform obtained during an excited larynx experiment with the anechoic subglottal tract. Notice the almost steady subglottal pressure when the vocal folds were in contact/closed.

The second experiment revealed differences in vibratory regimes of the excited larynx between the anechoic and resonant subglottal tracts. To investigate the differences, f_0 s were extracted from the EGG signal, using Praat (version 6.1 [16]) software. Figure 5 shows a histogram representing the density of f_0 values per 4 Hz bins, for both subglottal tracts. The two vertical lines mark the first resonance frequency of the subglottal tract (325 Hz) and half of its value (162.5 Hz). These values were derived using the Long Term Average Spectrum (LTAS) method: the subglottal pressure signal was segmented into 4096 points windows, FFT analysis was applied to each window and averaged over the whole signal. Although the elongation was increased gradually, abrupt f_0 jumps occurred in both anechoic and resonant tracts. With the anechoic subglottal tract, the f_0 appeared to jump between approximately 150 and 175 Hz (see Figure 5a). No other significant instabilities were observed. In the resonant case, the “destination” f_0 of the jumps was dominantly in the proximity of the first resonance frequency of the tract, as shown in Figure 5b.

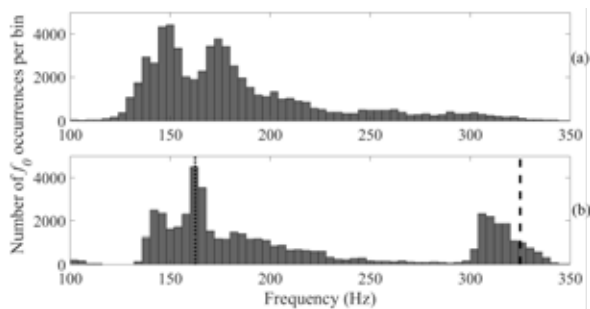


Figure 5: Histogram of fundamental frequencies (f_0) observed in the excised larynx with: (a) anechoic subglottal tract, and (b) resonant subglottal tract set to 325 Hz first resonance frequency. Dashed line: 325 Hz, dotted line: 162.5 Hz. Each bin has a width of 4 Hz.

These preliminary results thus confirm the influence of the subglottal tract on the subglottal pressure and on vocal fold vibrations. With the resonant subglottal tract, the f_0 oscillated dominantly between the first subglottal resonance frequency and half its value, suggesting an acoustic interaction with the subglottal tract. By its nature, the anechoic subglottal tract does not have acoustic resonances; therefore the occurring instabilities in that case can be related to the inherent vocal fold properties.

IV. CONCLUSION

An anechoic subglottal tract was developed, successfully suppressing its influence on vocal fold vibrations, and used in excised larynx experiments. Without the influence of the subglottal tract, the measured subglottal pressure was very similar to the predicted source signal. The developed tract can be used to study the inherent vibrational properties of the vocal folds, free of the acoustic interactions of the vocal folds with adjacent acoustic cavities.

V. ACKNOWLEDGMENT

This work was supported by the Czech Science Foundation (GA CR) project no. 19-04477S.

REFERENCES

[1] I. R. Titze, T. Riede, and P. Popolo, "Nonlinear source-filter coupling in phonation: Vocal exercises," *J Acoust Soc Am*, vol. 123, pp. 1902-1915, 2008.

[2] L. Wade, N. Hanna, J. Smith, and J. Wolfe, "The role of vocal tract and subglottal resonances in producing vocal instabilities," *J Acoust Soc Am*, vol. 141, p. 1546, Mar 2017.

[3] M. Zaňartu, 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 Am*, vol. 129, pp. 326-39, Jan 2011.

[4] I. R. Titze, "Nonlinear source-filter coupling in phonation: theory," *J Acoust Soc Am*, vol. 123, pp. 2733-49, May 2008.

[5] Z. Zhang, J. Neubauer, and D. A. Berry, "The influence of subglottal acoustics on laboratory models of phonation," *J Acoust Soc Am*, vol. 120, pp. 1558-1569, 2006.

[6] V. Hampala, J. Švec, D. Schovánek, and D. Mandát, "Utility Model No. 25585: Subglottal tract model (In Czech)," Czech republic Patent, 2013.

[7] M. M. Sondhi, "Measurement of the glottal waveform," *J Acoust Soc Am*, vol. 57, pp. 228-232, 1975.

[8] B. Cranen and L. Boves, "Pressure measurements during speech production using semiconductor miniature pressure transducers: Impact on models for speech production," *J Acoust Soc Am*, vol. 77, pp. 1543-1551, 1985.

[9] D. Miller and H. Schutte, "Characteristic patterns of sub-and supraglottal pressure variations within the glottal cycle," in *Transcr. XIIIth Symp. Care Prof. Voice*, 1984, pp. 70-75.

[10] H. Schutte and D. Miller, "Resonanzspieie der Gesangsstimme in ihren Beziehungen zu supra- und subglottalen Druckverluften: Konsequenzen fr die Stimmtheorie," *Folia phoniat*, vol. 40, pp. 65-73, 1988.

[11] R. L. Miller, "Nature of the vocal cord wave," *J Acoust Soc Am*, vol. 31, pp. 667-677, 1959.

[12] T. Murtola, P. Alku, J. Malinen, and A. Geneid, "Parameterization of a computational physical model for glottal flow using inverse filtering and high-speed videoendoscopy," *Speech Communication*, vol. 96, pp. 67-80, 2018.

[13] M. Rothenberg, "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing," *J Acoust Soc Am*, vol. 53, pp. 1632-1645, 1973.

[14] J. Sundberg, "Flow glottogram and subglottal pressure relationship in singers and untrained voices," *Journal of Voice*, vol. 32, pp. 23-31, 2018.

[15] J. Wolfe, D. Tze Wei Chu, J.-M. Chen, and J. Smith, "An Experimentally Measured Source-Filter Model: Glottal Flow, Vocal Tract Gain and Output Sound from a Physical Model," *Acoustics Australia*, vol. 44, pp. 187-191, April 01 2016.

[16] P. Boersma and D. Weenink. *Praat: Doing phonetics by computer*. Available: <http://www.fon.hum.uva.nl/praat/>

SHAKER: PRELIMINARY OBSERVATIONS OF A POTENTIAL DEVICE FOR VOICE TRAINING AND THERAPY

A.- M. Laukkanen¹, J. Horáček², V. Radolf²

¹ Speech and Voice Research Laboratory, Faculty of Social Sciences, Tampere University, Virta, Åkerlundinkatu 5, 33100 Tampere, Finland

² Institute of Thermomechanics of the Czech Academy of Sciences, Prague, Dolejškova 1402/5, 182 00 Prague, Czech Republic

Anne-Maria.Laukkanen@tuni.fi, jaromirh@it.cas.cz, radolf@it.cas.cz,

Abstract: A mechanic buzzer has originally been developed for clearing mucus from the lungs and trachea. A recent study tested the device on phonation and reported decreased vocal and laryngeal symptoms immediately after practicing on it [1].

This study presents measured 1) air pressure threshold and airflow threshold for buzzing during blowing and phonation into the buzzer using a physical model of voice production, and 2) oral pressure and electroglottogram (EGG) for two participants, also while blowing and phonating into the device. In both tests, a narrow mouthpiece out of two options was used, and the device was kept in horizontal and in upright positions. The buzzing roughly seems to correspond to water resistance therapy, i.e. phonation through a glass resonance tube submerged into water. Buzzer horizontally corresponds to 5cm submersion, and buzzer upright to 10 cm submersion of the tube in water.

Keywords: air pressure, airflow, EGG, larynx position, physical modelling

I. INTRODUCTION

A recent study tested phonation into a mechanic buzzer as a voice training and therapy method [1]. The buzzer (Shaker deluxe™) has originally been developed for clearing mucus from the lungs and trachea. Immediately after practicing with the buzzer, decreased vocal and laryngeal symptoms were reported [1]. It generates buzzing sound when the airflow sets a steel sphere inside a perforated hard plastic cover in vibration, which then shutters and opens the circular holes on the cover.

II. AIMS

As far as we know, the buzzer has not been formally tested for voice training use, and thus this study aimed to do so.

Test 1 measured air pressure threshold and airflow threshold for buzzing during blowing and for

phonation into the buzzer using a physical model of voice production [2].

Test 2 studied the device on two participants, similarly during blowing and during phonation. The narrower mouthpiece out of the two alternatives was used in both tests as the participants found it more comfortable to use.

III. METHODS

Test 1: A buzzer was connected to a physical voice production model consisting of three-layered silicon vocal folds (VFs), and a plexiglass vocal tract for vowel /u/, see Fig.1. The intermediate layer of the vocal folds was filled with water. The amount of water was used to regulate whether the VFs were abducted or adducted.

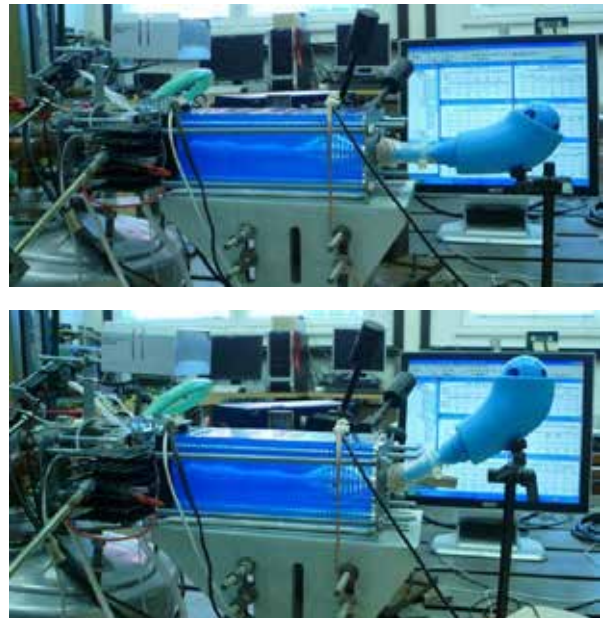


Fig. 1. Test 1, buzzer in horizontal (top) and in upright position (bottom) attached to a physical model of voice production [2].

The subglottic and oral pressures, P_{sub} and P_{oral} , were registered with integrated pressure semiconductor sensors (NXP Freescale MPXV5010GC6U) and the mean airflow rate Q was measured by a float flowmeter (EMKO type DF3-09K5).

First, the buzzing threshold pressures were measured during blowing into the shaker, i.e. letting airflow into it when the vocal folds were not involved (being in a highly abducted position due to VFs deflation caused by sucking out a portion of water from the model). For phonation, the vocal folds were slightly adducted to be in contact (by VFs inflation, filling the model with more water). Simultaneously by this maneuver (increasing the amount of water in VFs) the fundamental phonation frequency (F_0) was set to ca. 83 Hz, corresponding to a male low pitch voicing. The buzzer was connected to the model both in a horizontal and in an upright position (rotated up by an angle of ca 45 degrees, see Fig.1), i.e. when the circular hole under the vibrating sphere was in and in horizontal position compared to the edge surrounding the cover.

Test 2: A male and a female voluntary participant were recorded while (a) blowing into the buzzer and (b) phonating into it on sustained vowel [u] in different pitches and while producing pitch glides. Crescendos from buzzing threshold were also made. Subjective sensations during and immediately after buzzer use were registered with an open interview after buzzing. The buzzer was kept in two positions, see Fig. 2.

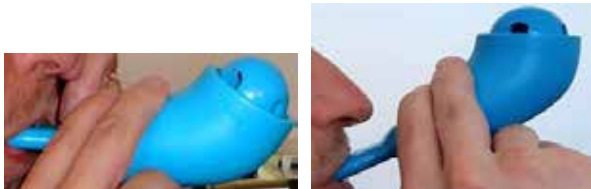


Fig. 2. Test 2, the buzzer used by the participants. The device was held both in horizontal (left) and in upright position (right) during blowing or phonating into it.

Acoustic signal was recorded with a sound level meter (B&K 2239) at 30 cm from the buzzer, electroglottographic (EGG) signal was registered with two electrode pairs for studying vocal fold contact area and vertical position of the larynx, VLP (EG-2, Glottal Enterprises). Three accelerometers (Delta Tron type 4519-002, weight 1.5 g) were attached to the skin for measuring vibrations on the side of the nose cartilage, and on the lower edge of the cricoid cartilage, and on the cheek. The oral pressure, P_{oral} , was registered with an integrated pressure semiconductor sensor (NXP Freescale MPXV5010GC6U) attached to a plastic tube (17 cm in length, 2.6 mm in inner diameter) that was held in the mouth corner.

All the signals (Test 1 and 2) were synchronously sampled and recorded by using the measurement system B&K Pulse (type 3560 C with Input/Output Controller Modules Type 7537A and 3109) controlled by a personal computer equipped by the SW PULSE LabShop (Version 10). The sampling frequency of the signals was 16.4 kHz.

IV. RESULTS AND DISCUSSION

Test 1, Model: The absolute flow rate threshold for barely detectable acoustic buzzing due to movement of the steel sphere inside the buzzer was 0.07 l/s for blowing (without phonation), and the corresponding P_{sub} was 0.92 kPa, buzzer being horizontally, see Table 1. For buzzer in the upright position, the buzzing threshold for flow increased to 0.08 l/s, and the corresponding P_{sub} increased to 1.89 kPa. This corresponds to the very general information given by the manufacturer stating that “the minimum air pressure required is 18 cmH₂O” [3].

Table 1. Buzzing threshold (BT) and phonation threshold (PT) for the mean P_{sub} , mean P_{oral} and mean flow rate Q measured for model. F_0 and F_b refer to fundamental frequency of phonation and frequency of buzzing.

Type of sound excitation	Buzzer position	P_{sub} [kPa]	P_{oral} [kPa]	Q [l/s]	F_0 [Hz]	F_b [Hz]	Note
Blowing (no phonation)	horizontal	0.92	0.7	0.07	0	20	BT
	upright	1.89	1.69	0.08	0	23	BT
Phonating	horizontal	0.25	0.09	0.02	83	0	PT
	upright	0.39	0.23	0.02	82	0	PT

The airflow threshold rate for clearly audible buzzing sound with phonation ($F_0 = 83$ Hz) was 0.015 l/s and the corresponding phonation threshold pressure P_{sub} was 0.25 kPa for buzzer being horizontally. For buzzer in the upright position, the corresponding threshold pressure increased to $P_{sub}=0.39$ kPa, while the airflow threshold 0.015 l/s was the same as for the horizontal position. The oral pressures P_{oral} were ca 200 Pa lower than P_{sub} at buzzing thresholds.

The buzzing threshold pressures (BTP) as well as the phonation threshold pressures (PTP) for the upright position of the buzzer were substantially higher than for the horizontal position, and the BTP values were considerably higher than the PTP values. Therefore, if the airflow rate increases from zero, the phonation starts first when the sphere moves only slightly from the circular hole and opens the airway output statically without vibration, and the clearly audible buzzing sound with phonation then appears for higher flow rate after crossing the BTP.

When the end of the shaker moves from the horizontal position slightly down, by eye not a recognizable change, the PTP goes relatively fast down. For a decrease of the shaker down by ca 5 mm, the PTP decreased from 1.66 kPa to $P_{sub}=1.22$ kPa.

It means that the phonation threshold in horizontal position is very sensitive to the buzzer's very small change downward.

Test 2, Humans: Table 2 summarizes measured air pressure values for mean oral pressure P_{oral} and peak-to-peak oral pressure $P_{oral\ ptp}$, fundamental frequencies F_0 and buzzing frequencies F_b for the participants. Fig. 3 illustrates results for P_{oral} , EGG waveform and VLP variation during phonation into the buzzer.

Table 2. Mean and peak-to-peak oral pressure (P_{oral} , $P_{oral\ ptp}$), fundamental phonation frequency (F_0) and frequency of buzzing (F_b) measured for two participants in blowing and in phonating into the buzzer during increasing the flow rate (noted as Inc. Q) or during constant phonation for 4 s (noted as phon. 4 s).

	Buzzer position	P_{oral} [kPa]	$P_{oral\ ptp}$ [kPa]	F_0 [Hz]	F_b [Hz]	Note
Male						
Blowing	horizontal	1.0 - 3.5	0.2 - 0.9	/	9-19	Incr. Q
Phonating	horizontal	0.70	0.20	93	12	phon. 4 s
		0.50	0.50	127	11	phon. 4 s
		0.60	0.30	139	13	phon. 4 s
	from upright to horizontal	2.2-0.9	0.4-0.4	93	12	phon. 4 s
Female						
Blowing	horizontal	0.75-2.0	0.07-0.87	/	9-18	Incr. Q
	upright	0.94-1.6	0.06-0.60	/	10-15	Incr. Q
Phonating	horizontal	0.60	0.70	158	11	phon. 4 s
	upright	1.0	0.60-1.0	150	13	phon. 4 s

In Table 2 we see that buzzing threshold (studied in P_{oral}) is higher with the buzzer in upright position compared to horizontal position. Buzzing frequency increases with flow rate (intensity of blowing and buzzing), varying between 9-19 Hz. There are differences in the values obtained between successive trials so no clear differences between blowing and phonation or between different F_0 values can be observed.

For blowing with the buzzer in horizontal position, the range of P_{oral} for buzzing measured in humans approximately matches the buzzing thresholds (BT) measured on model. In all cases of phonation, the oral pressure P_{oral} measured in humans was much higher than the phonation threshold (PT) measured on model.

Differences between the results obtained with the model and those obtained from the participants as well

as differences between human samples reflect the participants' difficulties in finding the exact buzzing thresholds (BT). Further causes for differences between model and humans may be 1) the structure of the model, which for example does not include the effect of the yielding walls in human vocal tract [4,5], and 2) differences in the exact horizontal and upright buzzer position in participants. The buzzer position determines the PT or BT.

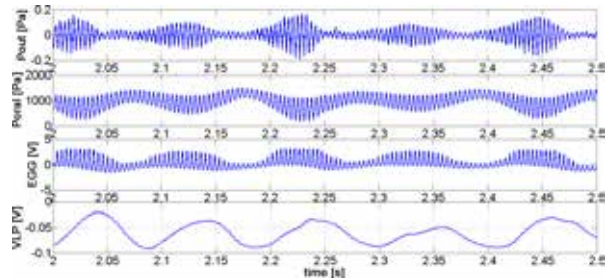


Fig. 3. (a) External sound pressure (top panel) obtained with microphone at 30 cm from the buzzer, (b) P_{oral} , (c) EGG, and (d) VLP (bottom panel) signals in time domain recorded for male voice. High pitch ($F_0=236$ Hz, $F_b=10$ Hz).

From Fig. 3 we see intense oscillation of P_{oral} which is reflected in variation of vocal fold contact and VLP during phonation into buzzer. In EGG waveform the contact phase is relatively short throughout the sample, suggesting that phonation is not pressed. The contact phase gets particularly short and the EGG signal amplitude becomes small (suggesting smaller contact area) at the moments when P_{oral} is at maximum. Results resemble those obtained for water resistance therapy [6].

Participants' subjective pros: Buzzer activates respiratory muscles, makes pressed phonation impossible, helps to keep a good upright head position (since the device does not buzz when the head is bent forward), makes it easier to glide to high pitches than on vowel, and phonation gets easier and louder after buzzing. *Cons:* May cause dryness in throat and tire adductors after extensive use, and in the long run may urge to excessive air consumption during phonation. Both participants were used to water resistance therapy and have been subjects in earlier experiments, e.g. in [4, 6]. They commented that buzzing feels easier than water resistance exercising.

V. CONCLUSION

A buzzer is potentially beneficial in voice training and therapy.

The air pressure and F_b values obtained with the buzzer in upright position roughly correspond to those obtained for exercises with a glass resonance tube

submerged 10 cm in water [3]. With the buzzer in horizontal position the P_{sub} threshold for buzzing is about half lower compared to the upright position.

ACKNOWLEDGEMENT

The study was supported by a grant from the Czech Science Foundation: No. 19-04477S “Modelling and measurements of fluid-structure-acoustic interactions in biomechanics of human voice production.”

REFERENCES

- [1] T. Lenharo Saters, V. Veis Ribeiro, L. T. Donalsonso Siqueira, B. Dantas Marotti, A. Ghedini Brasolotto, K.C. Alves Silverio, “The voiced oral high-frequency oscillation technique’s immediate effect on individuals with dysphonic and normal voices,” *Journal of Voice*, vol. 32(4), pp. 449–458, 2018.
- [2] J. Horáček, V. Radolf, V. Bula, J. Košina, “Experimental modelling of phonation using artificial models of human vocal folds and vocal tracts,” In: *Engineering Mechanics 2017* (Fuis, V. ed.). Brno: University of Technology, 2017, pp. 352-385.
- [3] <https://www.powerbreathe.com/shaker-deluxe>
- [4] J. Horáček, V. Radolf, A-M. Laukkanen, “Low frequency mechanical resonance of the vocal tract in vocal exercises that apply tubes,” *Biomedical Signal Processing and Control*, vol. 37, 2017, pp. 39-49.
- [5] J. Horáček, V. Radolf, A-M. Laukkanen, “Experimental and Computational Modeling of the Effects of Voice Therapy Using Tubes,” *Journal of Speech, Language, and Hearing Research*, vol. 62, 2019, pp. 2227–2244.
- [6] V. Radolf, A-M. Laukkanen, J. Horáček, D. Liu, “Air-pressure, vocal fold vibration and acoustic characteristics of phonation during vocal exercising. Part 1: Measurement in vivo,” *Engineering Mechanics*, vol. 21(1), 2014, pp. 53–59.

SESSION XI
SPEECH

BEATBOX SOUNDS RECOGNITION USING A SPEECH-DEDICATED HMM-GMM BASED SYSTEM

Solène Evain¹, Adrien Contesse^{2*}, Antoine Pinchaud^{*}, Didier Schwab¹, Benjamin Lecouteux¹,
and Nathalie Henrich Bernardoni³

¹Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

^{*}<http://www.vocalgrammatics.fr/>

²ÉSAD Amiens, De-sign-e Lab, 80080 Amiens, France

³Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

solene.evain@univ-grenoble-alpes.fr, AdrienContesse@gmail.com, APinchaud@gmail.com,
Didier.Schwab@imag.fr, Benjamin.Lecouteux@imag.fr, nathalie.henrich@gipsa-lab.fr

Abstract: Human beatboxing is a vocal art making use of speech organs to produce percussive sounds and imitate musical instruments. Beatbox sounds classification is a current challenge. We propose a beatbox sounds recognition system with an adaptation of the Kaldi toolbox, widely used for automatic speech recognition (ASR). Our corpus is composed of isolated sounds produced by two beatboxers and is composed of 80 different sounds. We focused on decoding with monophones acoustic models, trained with a HMM-GMM model. One type of transcription was used: a beatbox specific writing system named Vocal Grammaticics (VG) which uses concepts of articulatory phonetics.

Keywords: Human beatbox, automatic speech recognition, Kaldi, isolated sounds recognition

I - INTRODUCTION

Human beatboxing emerged in the '80s in the Bronx, a borough of New York City, and is associated with hip-hop culture. It consists in producing vocal percussions along with musical instruments imitations, such as trumpet or guitar. Beatbox sounds classification can be used for Music Information Retrieval as a request for searching different types of music [4] or for voice-controlled applications with a user-defined number of classes [3]. Good classification rates were obtained with an ACE-based system¹ on a limited range of classes, i.e. five main beatbox sounds *bass drum*, *open hi-hat*, *closed hi-hat*, *k-snare* and *p-snare* drums [7]. To the best of our knowledge, automatic recognition of beatbox sounds using a speech recognition system has only been ex-

plored by [5]. Their training database consists of isolated drum beatbox sounds (five classes *cymbal*, *hi-hat*, *kick*, *rimshot* and *snare*) and instrument imitations (8 classes). Performance was poor for instrument imitations (best recognition error rate of 41%), yet good performance was demonstrated for limited beatbox sounds classes (best recognition error rate of 9%).

In continuing this effort towards the development of an efficient and reliable automatic beatbox sounds recognition system, we aim to extend the number of sound classes and enable the recognition of subtle variants in beatbox sounds production. We consider human beatbox as a musical language composed of sound units that we shall call *boxemes* with reference to speech phonemes. This work was made with a view of creating an interactive artistic setup that would provide visual feedbacks during beatbox sounds production.

The paper is structured as follows. Section II presents the training database. The recognition system is presented in Section III. Different experiments are described in Section IV and their results given in Section V. Sections VI and VII provide a discussion and conclusion of the paper, along with future works.

II - MATERIALS

Our beatbox sounds corpus was recorded by two male beatboxers: a professional beatboxer (third author, stage name *Andro*) and an amateur one (second author). It is composed of 80 boxemes and could be considered as a large vocabulary corpus compared to previous corpora used in papers for classification. Isolated sounds only are considered here, rhythmic sequences being discarded in

¹Autonomous Classification Engine or ACE, developed for optimising music classification

first approach.

A articulatory-based pictographic writing system developed by second author and called *Vocal Grammaticics* [1] was used for annotation. In this latter, the glyphs are composed of two pieces of information: one about the speech organs that are used, and one about the manner the sounds are produced (plosive, fricative...). Fig. 1 illustrates this writing system in the case of a bilabial plosive sound with a morphological glyph representing two lips and a symbolic cross-shaped glyph representing plosion.



Figure 1: Representation of a bilabial plosive sound with *Vocal Grammaticics* pictographic writing system

Our Large Vocabulary Corpus was recorded with six microphones. Five of them were recording simultaneously and one was encapsulated (one or two hands cover the capsule of the microphone). The microphones differed in terms of specificities (e.g. condenser vs dynamic) and placement. Table 1 give the details of the microphones when table 2 is a recap chart of the composition of the corpus.

Microphone	Distance from the mouth	Specificity
Brauner VM1 (braun)	10 cm	condenser + pop filter
DPA 4006 (ambia)	50 cm	condenser ambient mic
DPA 4060 (tie)	10 cm	condenser
Shure SM58 (sm58p)	10 cm	dynamic
Shure SM58 (sm58l)	15 cm	dynamic
Shure beta 58 (beta)	1 cm	dynamic + encapsulated

Table 1: Recap chart of the different microphones

Large Voc. Corp.	
Beatboxers	Adrien (amateur), Andro (professional)
Date	2019
Num. of sounds	80
Num. of sounds per beatboxer	Adrien: 56/80 Andro: 80/80
Transcription	Vocal Grammaticics
Microphone	5 simultaneous + 1 encapsulated
Recording parameters	44100 Hz, 16 bits, mono, wav
Train	
Recording time	~92mn
Repetitions	6 or 2
Test	
Recording time	~114mn
Repetitions	7 (on average)

Table 2: Recap chart of the corpus

The training and testing of acoustic models was made with the Kaldi toolbox [6].

III - SPEECH RECOGNITION

Our approach is based on the assumption that human beatbox is structured like a musical language, using the speech organs to produce sound units that can be distinguished from each other and that each have a specific musical meaning for the beatboxer. In this context, a speech-dedicated recognition system could make it possible to automatically recognise beatbox productions. An ASR system can either be word-based or sub-word based [2]. Word-based systems require a model for each word in the vocabulary, trained with many repetitions of each word which are supposed to give the system a representation of the variability in speech production. Instead of cutting a sound into sub-words units and have a model per phoneme, a word-based system focuses on the sound as a whole and recognises the words it has been trained on only. This preliminary work focuses on isolated words recognition. Co-articulation or word boundary were discarded, yet keeping constraints on noise treatment, intra- and inter-speaker variability.

Mel Frequency Cepstral Coefficient (MFCC) acoustic features were extracted. They are based on human peripheral auditory system ([8]) and are widely used in ASR. Each beatbox sound was associated with a Hidden Markov Model (HMM).

IV - METHODS

Several systems were trained for the purpose of testing different recognition parameters. The influence of using different microphones with different placements and sensitivity was studied in order to know whether all the recordings could be used together for training a robust system.

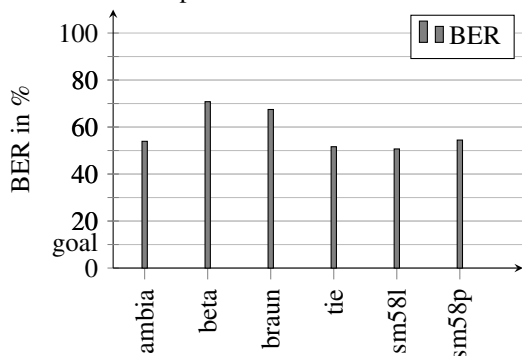
Finally, different parameters were tested: the increase of HMM states, the increase of MFCC, the addition of a pause phoneme in the lexicon and the increase of the silence probability. Some choices were based on [5]'s article.

V - RESULTS

One evaluation metric was used to evaluate the system. Based upon the Word Error Rate (WER), a *BER-Boxeme Error Rate* was calculated by adding the number of substitutions, insertions and deletions divided by the number of boxemes in the reference. The better the recognition, the smaller the BER value.

Graphs 4 and 5 give the BER for decoding performances. The "goal" line on horizontal axis represents our objective: obtaining a 10% BER or less, set to guarantee an interesting use of our system by the audience. Graph 4 shows the different performances in decoding for different types of microphones.

Graph 1: BER obtained with monophone acoustic models for the six microphones



The training sets are composed of recordings of the selected microphone to test.

As a result, DPA condenser microphones and Shure SM58 dynamic ones, either placed close or far from the beatboxer's mouth, provide similar performances. Worse recognition rates (with high BER) are found for recordings with encapsulated Shure beta 58 dynamic microphone and Brauner VM1 condenser microphone.

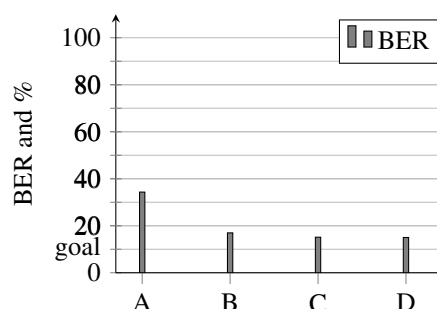
We then decided to vary the silence probability from 0.5 (default) to 0.9 with a step set to 0.1 and tested a 0.99 probability (as it has to be lower than 1). Our best

model was achieved with a 0.9 silence probability, which gave a 26,94% BER. When specifying on top a 'pause' phoneme before and after each boxeme in the lexicon, our best results are 16,97% BER. These are obtained with a 0.8 silence probability.

Graph 2 shows the evolution of the results with different parameters : higher silence probability, addition of a pause in the lexicon, 22 MFCC instead of default 13 and 5 HMM instead of default 3. The training set was made with recordings of non-encapsulated microphones. The test set is composed of recordings of the Shure SM58 'sm58p' microphone.

Graph 2: Evolution of BER with different parametrisations

A : default, B: 0.8 silence probability + pause, C: B + 22 MFCC, D: B + 5 HMM



Our best model is achieved with the C configuration and gives a 15.13% BER. B and D configurations are very close with a 16.97% BER and 16.24% BER respectively (see table 3 for details about the substitutions, insertions, deletions and correct boxeme rates).

In graph 3 and table 3, we observe that every change in parametrisation is beneficial for substitutions, insertions, deletions and correct boxeme rates. The most obvious benefit is for insertions rate which comes up to zero. The correct boxeme rate achieves 85% with the C configurations, meaning that 8.5 boxemes over 10 are well recognised.

	A	B	C	D
Substitutions	19.19%	12.73%	10.70%	12.36%
Insertions	9.41%	0.18%	0.18%	0%
Deletion	5.72%	4.06%	4.24%	3.87%
CBR	75.09%	83.21%	85.06%	83.76%

Table 3: Insertions, substitutions, deletions and Correct Boxeme Rate (CBR) for A B C D configurations
A : default, B: 0.8 silence probability + pause, C: B + 22 MFCC, D: B + 5 HMM

VI - DISCUSSION

As we can see from previous section, the efficacy of the different microphones is quite similar except for the Shure beta 58 and Brauner VM1 microphones which perform worse. We suppose it is because of the way we used them, and independent of the type of microphone. Indeed, the Shure beta 58 microphone is encapsulated and that use affects the performances of the microphone. As for the Brauner VM1 condenser microphone, we can observe it is performing worse than the other condenser microphone in our test (DPA 4060) and suppose it was placed too close from the mouth of the beatboxer. Finally, neither the number of MFCC nor the number of HMM states gives clear improvement. We suppose that having more HMM states could be interesting for complex sounds that are composed of two or more boxemes. This could be analysed in further studies.

VII - CONCLUSIONS AND PERSPECTIVES

Our system demonstrates the possibility of using a speech-dedicated recognition system to recognise beatbox sounds. Plus, we also demonstrate the possibility of recognising subtle variations of beatbox sounds as, for example, inhaled and exhaled sounds that are distinguished and not every time mixed up with one another.

So far, our best model was obtained with an increase of the silence probability (0.8 instead of 0.5), the silence phoneme "pause" being added in right and left contexts in the vocabulary and 22 MFCC. The best BER is 15.13%.

We could observe that the type of microphone used for recording does not seem to have any influence on the system. It depends more on their use (encapsulated or not). Putting aside the encapsulated microphone for training gives better results.

As for the different types of production, when mixed, they seem to badly degrade the performance. For now, regarding the substitutions, we can not conclude anything as the system seems to either mix up sounds that are quite similar to the ear or that have quite similar articulation, and sounds that are very different. We suppose that dividing the corpus depending on the sound length and adapting the number of HMM states could improve the system.

Dividing each sound in smaller chunks, as it is done for languages with phonemes or syllables is a perspective. Indeed, as the corpus vocabulary increases, the memory is more and more in demand with word-

based speech recognition. Having a boxeme-based model would decrease the number of models needed by the system and enable the treatment of coarticulation. Also, there are still rhythmic sequences and encapsulated sounds recognition to explore. Finally, it would be interesting to see if the difficult recognition of women and children voices in ASR is also a problem in beatbox sounds recognition.

VII - REFERENCES

- [1] A. Contesse and A. Pinchaud. *vocal grammatics*. Web page, www.vocalgrammatics.fr, Last consulted: 2019-08-29, Aug. 2019.
- [2] V. Gupta and M. Lennig. Large Vocabulary Isolated Word Recognition. In R. P. Ramachandran and R. J. Mammone, editors, *Modern Methods of Speech Processing*, The Springer International Series in Engineering and Computer Science, pages 213–230. Springer US, Boston, MA, 1995.
- [3] K. Hipke, M. Toomim, R. Fiebrink, and J. Fogarty. BeatBox: End-user Interactive Definition and Training of Recognizers for Percussive Vocalizations. pages 121–124, Como, Italy, May 2014. ACM.
- [4] A. Kapur, G. Tzanetakis, and M. Benning. Query-by-Beat-Boxing: Music Retrieval For The DJ. Barcelona, Spain, Jan. 2004.
- [5] B. Picart, S. Brognaux, and S. Dupont. Analysis and automatic recognition of Human BeatBox sounds: A comparative study. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4255–4259, Brisbane, QLD, Australia, 2015.
- [6] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely. The Kaldi Speech Recognition Toolkit. page 4, Hilton Waikoloa, Big Island, Hawaii, US, 2011.
- [7] E. Sinyor, C. McKay, R. Fiebrink, D. McEnnis, and I. Fujinaga. Beatbox classification using ACE. page 4, London, UK, 2005.
- [8] V. Tiwari. MFCC and its applications in speaker recognition. *International Journal on Emerging Technologies*, pages 19–22, 2010.

AM-FM DECOMPOSITION OF SPEECH SIGNAL: APPLICATIONS FOR SPEECH PRIVACY AND DIAGNOSIS

Petr Motlicek¹, Hynek Hermansky², Srikanth Madikeri¹, Amrutha Prasad¹, Sriram Ganapathy³
Idiap Research Institute, Martigny, Switzerland¹
The Johns Hopkins University, Baltimore, USA²
Indian Institute of Science Bangalore³

{petr.motlicek,srikanth.madikeri,amrutha.prasad}@idiap.ch, hynek@jhu.edu, sriramg@iisc.ac.in

Abstract: Although current trends in speech processing consider deep learning through data-driven technologies, many potential applications exhibit lack of training or development data. Therefore, considerably light signal processing techniques are still of interest. This paper describes an efficient technique for decomposing the AM and FM components of the speech signal, which is not based on frame-by-frame short-time analysis of the signal. Instead, we estimate all-pole models of frequency-localized Hilbert envelopes of large segments of speech signal at different frequencies. The technique on decomposition of speech signal into AM and FM components appears to be of interest in voice studies benefiting from alleviation of the message-bearing components of speech (e.g. security oriented applications such as speaker recognition, or speech diagnosis often relying on spectra averaging to discard the content of the speech). Similarly, discarding speaker information while preserving the message in the speech is of interest for privacy-oriented applications. Experimental results on automatic speech and speaker recognition tasks clearly show that the AM component preserves the content (message) of the speech, while the FM component carries the information related to the speaker.

Keywords: AM, FM, Linear prediction, Automatic speech recognition, Speaker recognition

I. INTRODUCTION

Dominant view of speech signal processing is still based on the linear model of speech production, where short segments of the signal (short enough so that the vocal tract does not significantly change within the segment) can be represented by short-time spectrum computed from these segments. The short-time spectrum consists of its spectral envelope (representing a linear filter emulating vocal tract transfer function at a given time instant) and its fine spectral structure. It is

widely accepted that the spectral envelope mainly represents the phonetic value of the speech segment (i.e. message) and the fine structure represents the spectrum of the excitation source. Many speech-oriented applications would benefit from being able to reliably separate contributions of the signal excitation and of the filtering.

Typical conventional techniques, such as linear prediction (LP) [7], are based on the linear modeling and apply frame-by-frame inverse filtering of speech using estimates of spectral envelopes of short speech segments. In this paper, we abandon the notion of the short-time spectrum of speech. Instead, we (along with work of Dudley 1940 [5]) see the speech as an audible signal generated by voice source (frequency modulated component FM), which is modulated by inaudible and mostly invisible movements of the vocal tract (amplitude modulated component AM). The movements of the vocal tract carry a bulk of the message in speech, while the voice source makes these tract movements audible, allowing for the message to be perceived by a listener.

The paper describes an efficient technique for decomposing the AM and FM speech components, not based on frame-by-frame short-time analysis. Instead, we estimate all-pole models of frequency-localized Hilbert envelopes of large speech segments at different frequencies. This is done by applying the LP technique to short segments of a cosine transformed speech signal. Since each segment of the cosine transformed signal represents the individual frequency component of the original signal, the resulting all-pole models yield the frequency-localized Hilbert envelopes of the signal. Inverse cosine transforms of their LP residuals then yield frequency-localized FM components of the voice source signal. Summing all frequency-local FM estimates yields the FM voice signal with its message alleviated. When the audible AM component of the speech signal is desired, the frequency-localized all-pole models of Hilbert envelopes are used to compute frequency-localized modulated noise

components, which are summed to yield the AM signal component carrying the speech message.

II. AM-FM DECOMPOSITION

The concept of AM-FM decomposition is presented through frequency domain linear prediction (FDLP) - an efficient technique for autoregressive modelling of temporal envelopes of the signal [8]. FDLP proposes to model the speech in critical bands as a modulated signal with the AM component obtained using Hilbert envelope estimate and the FM component obtained from the Hilbert carrier. The sub-band temporal envelopes can then be estimated using FDLP. Unlike traditional temporal domain LP representing the envelope of the power spectrum of the signal [7], FDLP particularly exploits the prediction power of slowly varying long-term AM envelopes of speech signals in critical sub-bands. The final FDLP model provides smoothed, minimum phase representation of temporal rather than spectral envelopes.

The duality between time and frequency domains suggests that the power of autoregressive models can be applied equally well to discrete spectral representations of the signal instead of time-domain signal samples. Interestingly for FDLP, it has been analytically shown that the squared magnitude response of the all-pole filter approximates the Hilbert envelope of the signal. At the same time it is known that the quadrature version of a real input signal and its Hilbert transform are identical for many modulated signals, known in practice. We can therefore presume that the Hilbert envelope approximates squared AM envelope of the signal. Thus, FDLP estimates the AM envelope of the signal and the FDLP residual contains the FM component of the signal. Acoustic signals in sub-bands are modulated signals and hence, FDLP can be used for AM-FM decomposition of sub-band signals.

III. DETAILED ANALYSIS

Source-filter linear model of speech production: Our current view of speech is dominated by the concept of the linear model of speech production (Chiba and Kajiyama 1942) [6], where the stationary source signal is filtered by the stationary filter. It assumes no interaction between the two components of this model (hence “linear”). This model is the basis for the LP speech analysis.

Carrier nature of speech: Before Chiba and Kajiyama, Homer Dudley (Dudley 1940) [4] published his concept of speech, where he suggested that for the human communication by speech, nature evolved a technique which is conceptually identical to the (then

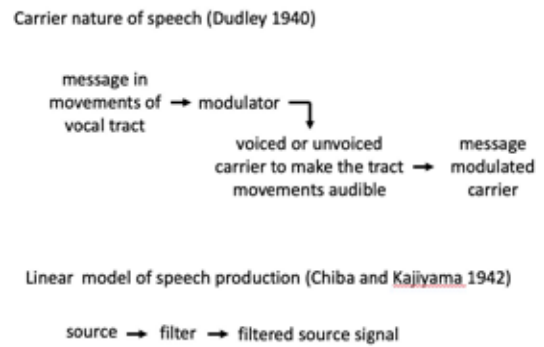


Fig. 1: Dudley's concept of speech [5] as a modulated carrier signal and the linear model of speech production.

dominant) AM radio communication. In his concept, messages are carried in signal changes, reflected in slow movements of vocal tract. The movements are made audible by using them for modulating the audible voice carrier. The current paper follows this concept in the form of the FDLP.

Estimating components of models of speech: In deriving the speech messages, we are primarily interested in the vocal tract movements, i.e., in the modulation function. On the other hand, in many applications of voice technologies such as a speaker recognition, or voice pathologies, it is the carrier, which is of interest.

Conventional method of the carrier extraction is inverse filtering, where estimated spectral envelopes of short speech segments are used for design filters, which are then used for whitening the respective short segments of speech signals. A typical example of this technique is the LPC inverse filtering [7]. This in effect yields the modulating function, which is sampled at the frame-rate of the short-time analysis. Since the assumptions of stationarity and linearity are easily violated, an accurate estimation of the individual components of this model can be difficult [9].

We are following the original Dudley's concept, where estimated temporal envelopes of spectral trajectories of speech signals at different frequencies are used for alleviating message components in respective frequency bands. Estimating the modulating function was originally done by analog low-pass filtering of spectral energies in different frequency bands [4]. Here, we show that the concept of the all-pole modeling employed in the LP analysis can be successfully adopted for the estimation of spectral energy trajectories in different frequency bands.

IV. FDLP

The concept of the FDLP for modeling short segments of Hilbert envelopes was investigated in [11] and extended by modeling of Hilbert envelopes in narrow frequency bands in [12,8].

In FDLP, the LP prediction is applied to the cosine transform of the speech signal $s(t)$, $t \in \langle 0, T \rangle$. One way to compute the cosine transform $q(t)$, $t \in \langle 0, T \rangle$ of a signal $s(t)$ is through the Fourier transform of the signal $S_{sym}(t)$, $t \in \langle 0, 2T \rangle$, which is the even symmetrized $s(t)$, i.e., $q(\omega) = F[S_{sym}(t)]$. The $q(\omega)$ is a function of frequency and is real and even symmetric.

Being after the cosine transform in frequency domain allows for a selection of the frequency range to be further processed. The signal

$$q_w(\omega) = q(\omega)w(\omega), \text{ where window } w(\omega_0) = \begin{cases} w_{\omega_0} & \omega = -\Delta\omega \leq \omega \leq \Delta\omega \\ 0 & \text{otherwise.} \end{cases}$$

ω_0 indicates the center of the frequency band to be processed. The Fourier transform of $q_w(\omega)$, which is still real and causal, obeys the Krammers-Kroening relation $F[q_w(\omega)] = \{s_{\omega_0}(t) + H[s_{\omega_0}(t)]\}$.

The signal in a given frequency band, centered at ω_0 , $s_{\omega_0}(t)$, now stands in place of the real part of the Fourier transform and its Hilbert transform takes place of its imaginary part. The instantaneous energy in the signal in a given frequency band (Hilbert envelope) $H_{\omega_0}(t) = s_{\omega_0}(t)^2 + H[s_{\omega_0}(t)]^2$ is an equivalent of the power spectrum $P(\omega)$ in the time-domain LP.

The autoregressive model computed from the cosine transform of the signal $q(\omega)$ obeys the equation

$$E_{\omega_0} = \frac{1}{2T} \int_{-T}^T \frac{H_{\omega_0}(t)}{H_{\omega_0}^{\wedge}(t)} dt,$$

where $H_{\omega_0}^{\wedge}(t)$ is the all pole autoregressive model of the Hilbert envelope $H_{\omega_0}(t)$ and E_{ω_0} is the error of the model fit in the frequency band centered at ω_0 over the time interval T . The form of the error equation implies a good fit of the spectrum of the autoregressive model $H_{\omega_0}^{\wedge}(t)$ to the peaks of the Hilbert envelope $H_{\omega_0}(t)$. Center of the frequency window $w(\omega_0)$ is typically gradually moved through the whole frequency range of the signal to be processed.

Re-synthesis from the FDLP: The FDLP model can be used to construct inverse filter for whitening the segment of the cosine transform. Whitened segment is inverse cosine filtered to represent the whitened signal in the respective band. Adding whitened signals from all frequency bands yields the carrier signal. Modulating white noise in the frequency bands by the estimated FDLP Hilbert envelopes yields whispered-like speech with the original speech message.

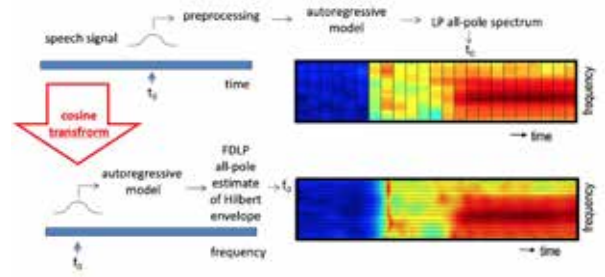


Fig. 2: The upper part of the picture shows the conventional LP as used in estimation of short-time spectral envelopes of short segments of speech centered at different times t_0 . The lower part shows the process of estimation of Hilbert envelopes in different frequency bands of speech signal centered at f_0 .

V. APPLICATIONS

The technique on decomposition of speech signal into AM and FM components appears to be of interest in voice studies, which would benefit from alleviation of the message-bearing components of speech (e.g. security oriented applications such as speaker recognition, or speech diagnosis often relying on spectra averaging to discard the content of the speech). In this paper, we empirically show that AM and FM components of the speech signal carry different types of information, AM related to the content and FM related to the speaker information, respectively.

V. EXPERIMENTS

We apply the AM-FM decomposition proposed in [10]. FDLP approach described in Section IV uses a simple window on top of cosine transformed (1000 ms long) speech segment to select a particular frequency band. Unlike previous, the following experiments apply slightly different FDLP version, available freely at Github¹. First, the input speech is decomposed into 32 critically-sampled frequency sub-bands by using a conventional quadrature mirror filter (QMF) bank. FDLP is then applied on each sub-band to model the sub-band temporal envelopes (AM components). The LP residual represents the FM in the sub-band signal. These steps are reversed at the synthesis side, to reconstruct the signal back from QMF sub-band components.

Two sets of experiments are performed: automatic speech recognition (ASR), and speaker verification (SV) deployed on (i) original (fullband) speech, (ii), the speech reconstructed only from the AM sub-band components (i.e. envelope extracted using FDLP), and

¹ github.com/iiscleap/SignalAnalysisUsingAm-FM

Tab. 1: ASR and SV results measured in terms of word error rate (EER) and equal error rate (EER), respectively, on Librispeech corpus.

	ASR system	SV system
	WER [%]	EER [%]
Original speech	10.1	14.7
AM-only	14.9	26.5
FM-only	53.9	25

(iii) the speech reconstructed only from the FM sub-band components (i.e. carrier part alone). Subjective listening tests clearly show that the AM-only reconstructed signal sounds whispered. With the carrier part alone, the synthesized signal sounds message less.

Dataset and tool: For ASR and SV experiments, we use Librispeech corpus [3] which consists of read speech from audio books. We employ 100 hours for training (train-clean-100) and 5.4 hours for testing (test-clean). Kaldi toolkit [2] is used for building both ASR and SV.

ASR: the system is built around a conventional HMM-GMM framework. We use standard Kaldi (tri4) recipe comprising MFCC features projected by LDA+MLLT [1]. Roughly ~3.5K triphones and ~40k Gaussians are used to build HMM-GMM.

SV: Gaussian Mixture Models (GMMs) with 32 components are trained for each speaker in test set. Each GMM is built with the expectation-maximization algorithm to maximize the likelihood of the data [13]. Only 10s of speech data were used for both GMM development and testing. Cross-pair trials for SV experiments were generated and trials comparing the same audio are excluded. T-norm is applied on the test scores.

VI. DISCUSSIONS AND CONCLUSIONS

The paper discusses employment of AM-FM decomposition to efficiently alleviate message bearing components from the speech. The technology is demonstrated on ASR and SV tasks. As can be seen from Tab. 1, the speech signal reconstructed from AM components yields WER~14.9%, close to the performance of the original signal (WER~10.1%) on the standard ASR task. On the other side, the speech reconstructed from FM-only components largely increases WER (~53.9%). In the case of SV task, the obtained results are less obvious. Original speech still

provides the best performance (EER~14.7%) as the SV engine also exploits the content to model the speaker. Nevertheless, the speech signal reconstructed from FM-only components still outperform AM-only speech (EER~25%) which clearly indicates that the speaker related information is preserved by the Hilbert carrier. FDLP technique described in this paper, allowing to decompose the speech into AM and FM components, operates on large segments of signal at different frequencies. Empirically obtained results on automatic speech and speaker recognition tasks confirm our assumptions (determined by subjective listening) that the AM-FM decomposition can reliably separate the content and speaker related information from speech, which can be applied in various speech-oriented tasks.

REFERENCES

- [1] S. Rath, et al, "Improved feature processing for Deep Neural Networks," *Proc. of Interspeech 2013*.
- [2] D. Povey, A. Ghoshal, et al., "The Kaldi Speech Recognition Toolkit," in *Proc. of IEEE ASRU*, 2011.
- [3] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *IEEE ICASSP*, 2015.
- [4] H. W. Dudley, "The vocoder," *Bell Labs Rec.*, vol. 18, pp. 122-126, 1939.
- [5] Dudley, H. (1940). The carrier nature of speech. *Bell System Technical Journal*, 19(4), pp. 495-515.
- [6] T. Chiba, M. Kajiyama (1958), "The vowel: Its nature and structure" (Vol. 652), Tokyo: Phn. society of Japan.
- [7] J. Makhoul (1975), "Linear prediction: A tutorial review," *Proceedings of the IEEE*, 63(4), pp. 561-580.
- [8] M. Athineos, D. Ellis, "Frequency-domain linear prediction for temporal features", *Proc. IEEE ASRU Workshop*, pp. 261-266, December 2003.
- [9] Alku, P. (1992). "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech communication*, 11(2-3), pp. 109-118.
- [10] S. Ganapathy, P. Motlicek and H. Hermansky, "Autoregressive Models Of Amplitude Modulations In Audio Compression", *IEEE Transactions on Audio, Speech and Language Processing*, August 2010.
- [11] R. Kumaresan, "An inverse signal approach to computing the envelope of a real valued signal", *IEEE Signal Proc. Letters*, vol. 5, no. 10, pp. 256-259, 1998.
- [12] J. Herre, J. D. Johnston, (1996, November), "Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS)," *In Audio Engineering Society Convention 101*.
- [13] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech communication* 17.1-2 (1995): pp. 91-108.

THE TOMATIS ELECTRONIC EAR EFFECTS ON SIMPLE VOCALIZATIONS

M. Prenassi¹, W. Coppola², G. Ramponi³, T. Agostini², S. Marceglia^{1,3}

¹ IRCCS Ca' Granda ospedale Maggiore policlinico Milano, Milan, Italy

² Dipartimento Scienze della Vita, Università degli studi di Trieste, Trieste, Italy

³ Dipartimento di Ingegneria e Architettura, Università degli studi di Trieste, Trieste, Italy

m.prenassi@gmail.com, walter.coppola@alice.it, ramponi@units.it, agostini@units.it, smarceglia@units.it

Abstract: The Tomatis electronic ear is a device that could modify the natural audio feedback between the emitted voice and the ears of a talking or singing individual. Our aim was to test if the device causes quantifiable vocal variations having the subjects repeat sustained vowel sounds (i.e. /a/, /i/, /u/) with different frequency filters applied by the device. The subjects are 19 native adult Italian speakers (8 females) testing 4 different filtering methodologies: unfiltered feedback (control), low pass filter at 4 kHz, high pass filter at 4 kHz and a high pass filter at 8 kHz. All subjects quantifiably modified their vocalization in response to the varying methodologies for at least one letter of each filtering method: 81.29% of the sessions of all subjects were significantly different in fundamental frequency from the control ($p < 0.05$, Kruskal-Wallis test). Among subjects, the variation trend was significant only for the fundamental frequency of the letter /u/ of a particular subgroups categorized by mean fundamental frequency. This initial work shows that the vocal variations caused by the Tomatis device are quantifiable but subject specific, laying the groundwork to test new parameters to find common trends of configurations.

Keywords: Tomatis, Electronic Ear, audio feedback, audio stimulation

I. INTRODUCTION

Alfred Tomatis was a French scientist, founder of Audio-Psycho-Phonology, an auditory rehabilitation methodology that stimulates the ear modifying the auditory input. This stimulation is delivered through a device called Electronic Ear. This device is based on a series of amplifiers, filters, and electronic controls, which receives the sound, emitted by a source, processes it and sends it back to the subject through a special headset. Tomatis's theory of listening is the product of a series of rigorous neurophysiological studies, based on the phylogenetic and ontogenetic analysis of the development of the nervous system [1][2][3]. It was fundamental to highlight the common origin and the consequent structuring of the organs

responsible for vocal emission (for example, V cranial pair for the musculature of the mandible and for the muscle of the hammer, VII cranial pair for the upper part of the larynx, for facial muscles and for the muscle of the stirrup), thus evidencing the very close correspondence between listening and voice production.

The conclusions reached by Tomatis are as follows:

“the voice can only contain the frequencies that the ear can hear (the larynx emits only the harmonics that the ear can hear)” and “if one modifies the hearing, the voice is unconsciously and immediately modified”[4][5][6]. In 1957, the theory was experimentally corroborated by a team led by Raoul Husson in the Functional Physiology laboratory at the Sorbonne in Paris [7]. After this experiment, fewer than a dozen offshoot and the related training systems have been developed based on this effect, with mild claims of effectiveness [8]. Only a fraction of these studies used the voice of the subject as auditory input. Our aim, in this preliminary work, is to test a new model of the Electronic Ear and the vocal variations that it causes on subjects emitting simple sounds (i.e. single sustained vowels) that are modified and fed back to them through special earpieces. This experiment was chosen to test the effectiveness of the device at a fundamental level, as a first step to map the actual capabilities of the device and of the method.

II. METHODS

In total, 19 native Italian speakers, 8 females and 11 males without speech impairments were recruited. The experimental setup included a microphone (Shure BETA 58A, Beyerdynamic TG V56c), the Tomatis system (Brain-Activator MBL), and an external recording device (M-AUDIO Fast Track Pro, sampling at 44100 samples/s, and a recording computer). The subjects were standing in a pre-marked position with their back and head touching a wall. The microphone was placed in a fixed position.

This preliminary experiment consisted of 4 segments, each composed of three sessions. In the first task of each segment, the subjects had to repeat 20 times the

vocalization of /a/ described as the corresponding vocal sound in Italian; this is called “a” session. After this, they waited 60 seconds and repeated the task with the letters /i/ and /u/ (“i” session and “u” session). The three sessions together define a segment.

In the first segment (NF), the Tomatis system acted as a straightforward audio loop, without any deliberate signal manipulation (no frequency filtering and no delay). NF is used as a control. The following three segments included a frequency filtering of the voice, first a low pass with -3dB cut-off frequency at 4 kHz (LP4K), then a high pass with the same cut-off frequency (-3dB at 4 kHz, named HP4K), and finally a high pass at 8 kHz (also -3dB cut-off, named HP8K). The total 4 segments of 3 sessions (NF, LP4K, HP4K and HP8K, for the sessions /a/, /i/, /u/) resulted in 80 vocalizations per subject and letter. The vocalizations were segmented and analyzed with the software PRAAT (<http://www.fon.hum.uva.nl/praat/>, downloaded June 2019) and a built-in algorithm [9] to extract the fundamental frequency (F0) and the first two formants (F1 and F2 respectively). If the algorithm was not capable of detecting a formant within a predetermined interval IF_i (1) the vocalization was discarded.

$$IF_i = MFV_i \cdot (1 \pm 0.5) \quad (1)$$

MFV_i: mean vowel formant (based on data from [10]);

i = 1: F1;

i = 2: F2.

Table 1. Percentage of sessions, ordered by letter, fundamental frequency and formants, statistically different from the control group (Kruskal-Wallis, $p < 0.05$). MeanV: mean change by letter for all the methods. MeanM: mean change by method for all the letters.

		LP4K %	HP4K %	HP8K %	MeanV %
F0	/a/	68.42	73.68	84.21	75.44
	/i/	78.95	84.21	84.21	82.46
	/u/	94.74	89.47	73.68	85.96
MeanM % F0		80.70	82.46	80.70	81.29
F1	/a/	42.11	42.11	57.89	47.37
	/i/	36.84	47.37	57.89	47.37
	/u/	52.63	68.42	57.89	59.65
MeanM % F1		43.86	52.63	57.89	51.46
F2	/a/	42.11	42.11	47.37	43.86
	/i/	63.16	47.37	68.42	59.65
	/u/	57.89	63.16	52.63	57.89
MeanM % F2		54.39	50.88	56.14	53.80

III. RESULTS

All subjects completed the experiment, resulting in 228 sessions (76 segments), but we had to discard some poor quality samples; overall, we analyzed 18 vocalizations for each session, for a grand total of 4104 vocalizations. Among the segments, regarding F0, 81.29 % of the sessions of all subjects were significantly different from the control NF ($p < 0.05$, Kruskal-Wallis test) as shown in Table 1. The first and second formants were less influenced by the Tomatis loop than F0, with 51.46% of sessions significantly different from NF for F1 and 53.80% of sessions significantly different from F2 (Table I). Still, all subjects responded to some extent to the feedback, modifying their vocalization formants in response to a method for at least one letter.

In Fig. 1 are shown the vocalizations of four different subjects of the letter /i/, even if some sessions are visually different from the control group (Fig. 3, C and D).

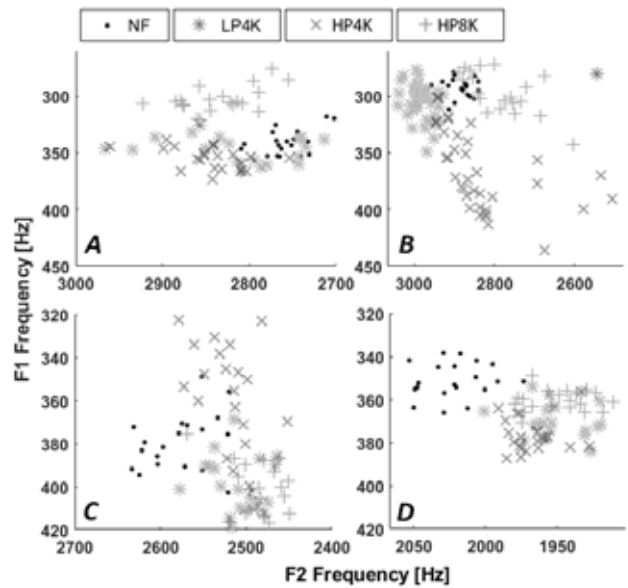


Fig. 1. First and second formants for the letter /i/ of four subjects, divided by methodology. A: subject 1, high F0 group, B: subject 7, high F0 group; C: subject 13, low F0 group, D: subject 19, low F0 group.

To assess the presence of underlying trends, further analysis were conducted.

Means and standard deviations were calculated for each session. Subjects were divided into two groups: high-F0 and low-F0 (the threshold was the total average of F0). The groups were composed of 8 female and 2 male subjects for the high-F0 category (in total 10 subjects) and the remaining 9 male subjects for the

low-F0 category. This classification was performed because group-specific filtering effects were observed in the preliminary analysis.

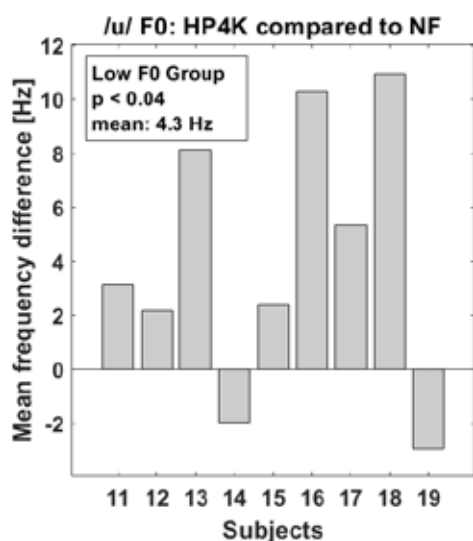


Fig. 2: Difference for the F0 of letter /u/ between the HP4K method and control (NF), regarding the low F0 group.

In the low-F0 group there was a significant increase by 4.3 Hz ($p < 0.04$, Wilcoxon signed rank test, Fig. 2) in the fundamental frequency (F0) of the letter /u/ of the HP4K session. For the same vowel, between HP4K and LP4K, F2 increased by 68.23 Hz ($p < 0.006$, Wilcoxon signed rank test) in the high-F0 group, and increased for the letter /a/ by 33.63 Hz ($p < 0.02$, Wilcoxon signed rank test) for the low-F0 group, as shown in Fig. 3. The standard deviation analysis had a statistical significance only in the F2 formant for vowels /a/ and /i/ in the high-F0 group: it decreased from HP8K to HP4K and increased from HP4K or HP8K to LP4K.

In the low-F0 group, the standard deviation had statistical significance only for the letter /i/: it increased in F1 for HP8K compared to controls, and for LP4K compared to controls. It decreased in F2 for HP8K compared to HP4K. Fig. 4 shows the average values of F0 for all the subjects in the high-F0 and low-F0 groups; all the letters showed a distinct increase between control NF and the methods, except for /a/ in the lower F0 group. However, these results are statistically different only for the letter /u/ in the lower F0 group between NF and HP4K ($p < 0.05$, Wilcoxon signed rank test).

IV. DISCUSSION

As shown in the results section a clearly audible and statistically relevant response is evoked by the filtered Tomatis audio loop, even if a unified trend response

among subjects is not clearly delineated. The first results we present in this paper indicate some ability of the Tomatis system to modify these covariances and as a consequence to act on general voice quality. Indeed, experiments in which the relationship among F0 and the F1 and F2 formants are synthetically modified are reported in the open literature (e.g., [11]). It was shown that the perceived quality of a voice depends on the covariance of the formants, which should correspond

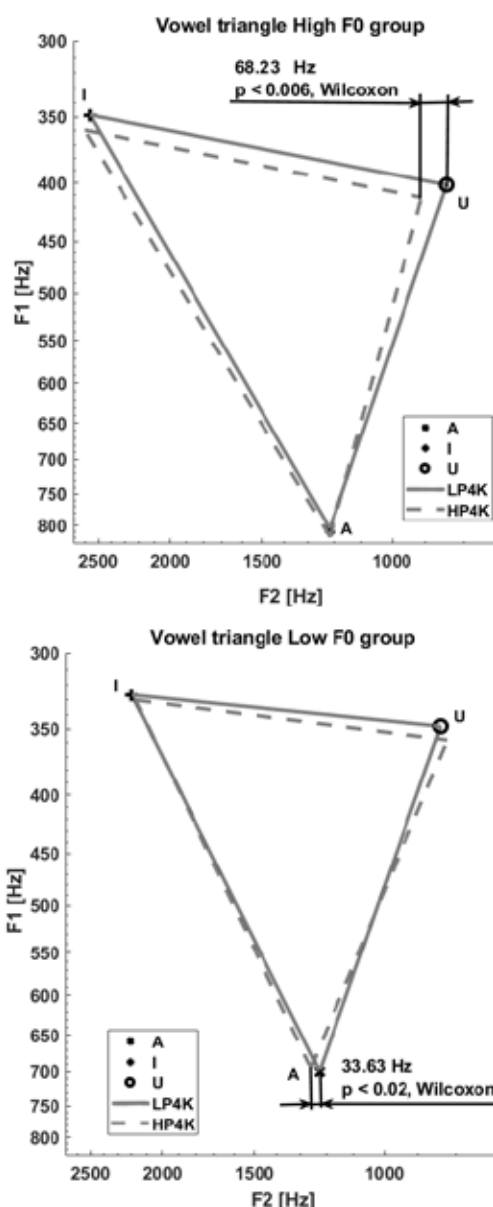


Fig. 3: First and second formant for the all-subject average vowel triangle of the letter /a/, /i/, /u/ for the LP4K and HP4K methods. At the top the high F0 group and at the bottom the low F0 group methods with the significant F2 drifts for the letter /u/ (top) and /a/ (bottom).

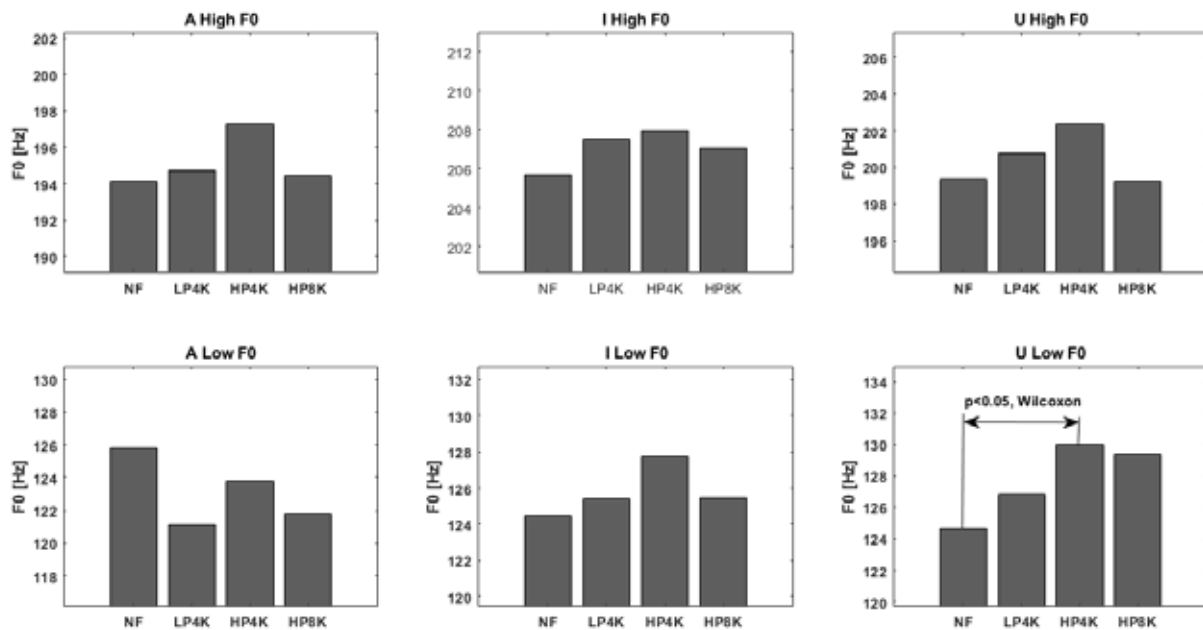


Fig. 4: Average values of F0 for all the subjects of the high-F0 and low-F0 groups. In U, Low F0, the differences between NF and HP4K are statistically significant ($p < 0.005$, Wilcoxon signed rank test)

to an internalized representation of human voice. Even the perception of emotions of the speaker depends on the formants' properties [12]. Even if not statistically relevant, it is also worth of notice the precise pattern followed by the all-subject average of F0 shown in Fig. 4. The HP4K method is able to elicit bigger changes in the fundamental frequency compared to the controls.

V. CONCLUSION

In this preliminary experiment, we showed that the vocal changes elicited in the subjects by the Electronic Ear are quantifiable. The specific effects depend on the type of vocalization and on the class of the subject (high-F0 or low-F0). A clear trend in the fundamental frequency was detected only for the /i/ and /u/ vowels. The standard deviation analysis suggests that a central (4 kHz cut-off frequency) high pass voice filter tends to increase the sound variation. This initial work may be useful to understand the capabilities of the Electronic Ear and the Tomatis methodology.

REFERENCES

- [1] P. Sollier, *Listening for Wellness. An Introduction to the Tomatis Method*. Walnut Creek, CA: Mozart Center Press, 2005.
- [2] W. Coppola, *Itinerari sonori. Corpo, linguaggio ed espressione nell'opera audio-psico-fonologica di Alfred Tomatis*. Milan: Mimesis, 2015
- [3] C. Campo, *L'orecchio e i suoni fonti di energia. Il metodo Tomatis*, Riza Scienze, 74, 1993.
- [4] A. Tomatis, *L'oreille et le langage*, Paris: Édition du Seuil, 1963.
- [5] A. Tomatis, *Vers l'écoute humaine, Tome I, Qu'est-ce que l'écoute humaine? Tome II, Qu'est-ce que l'oreille humaine?*, Paris: ESF, 1974.
- [6] A. Tomatis, *L'oreille et la voix*, Paris: Laffont, 1987.
- [7] R. Husson, "Physiologie de la Phonation", *Practica Oto-Rhino-Laryngologica*, 1957, vol. 50.2, pp. 123-139.
- [8] B.M. Thompson and S.R. Andrews, "An historical commentary on the physiological effects of music: Tomatis, Mozart and neuropsychology", *Integrative Physiological and Behavioral Science*, 2000, vol. 35, pp. 174.
- [9] D. G. Childers, *Modern spectrum analysis*. 1st ed., IEEE Computer Society Press, 1978, pp. 252-255.
- [10] P. Ladefoged, and K. Johnson, *A course in phonetics*, 6th ed., USA: Wadsworth, 2011.
- [11] P.F. Assmann, and T.M. Nearey. "Relationship between fundamental and formant frequencies in voice preference," *The Journal of the Acoustical Society of America*, 2007, 122.2 EL35-EL43.
- [12] W. Xixin, L. Songxiang, C. Yuewen, L.Xu, Y. Jianwei, D. Dongyang, M. Xi, H. Shoukang, W. Zhiyong, L. Xunying, M. Helen, *ICASSP 2019 - IEEE International Conference on Acoustics, speech and Signal Processing (ICASSP)*, 2019.

SESSION XII
VOICE VS OTHER PHYSIOLOGICAL
SIGNALS/DISEASES

THE INTELLIGIBILITY OF POLISH SPEECH IN CLEFT LIP AND PALATE CHILDREN

Wiktor Gonet¹, Edyta Zomkowska², Maria Hortis-Dzierzbicka³

¹Department of Phonetics and Phonology, Maria Curie-Skłodowska University, Plac Marii-Curie Skłodowskiej 4a, 20-031 Lublin, Poland

^{2,3}Department of Otorhinolaryngology and Head and Neck Diseases, Faculty of Medical Sciences, University of Warmia and Mazury, Warszawska 30, 10-082 Olsztyn, Poland

¹wiktor.gonet@gmail.com, ^{2,3}dr_ahortis@poczta.onet.pl

Abstract: Assessment of the correctness of pronunciation by children after cleft lip and palate surgery produces only raw results that should be viewed against the phonemic pattern of that language. For example, Polish has three places of articulation of affricates, whose articulation requires greater precision than that of the single affricate in English. If we grant 1 point for the correct pronunciation of an affricate in each place of articulation, the maximum score to be won by Poles will equal 3, and those produced by the English, 1. A reverse protocol can be used, giving 1 point to an English child, and 0,333 points to a Polish child for the correct pronunciation of each affricate. The choice of these protocols has not been studied in depth.

To remedy this situation, a new approach is emerging, following the assumptions worked out by Jakielski (1998) for stutterers [1]. The author of the present paper has studied the problems English students have in the pronunciation of certain Polish speech sounds [2], [3], [4], [12]. In this context, too, the more members a given subclass has, the more difficult each one is to pronounce.

Keywords: speech intelligibility, cross-language comparisons, phoneme patterns, clustering possibilities.

I. INTRODUCTION

The assessment of speech intelligibility in patients who have undergone surgical fusion of the split lip and palate is currently carried out using various verbal tests in which the patient describes the picture or reads or repeats after the instructor sentences containing appropriately selected language material. The result of this test is to determine the correctness of pronunciation on a point scale. However, when comparing surgery results for speakers of different languages, this method of assessment is not fair for speakers whose native phonetic system has a comparatively large number of consonants and their

clusters, e.g. Polish, while English is an example of a language with a less complex sound pattern. The more complex phonemic structure in Polish makes correct pronunciation more difficult because of the necessity to articulate larger numbers of phonemes with greater precision.

In order to correct this situation, a method should be developed in which the assessment of the pronunciation of patients speaking a language with a complex phonetic system will be offset, with regard to a less difficult sound, by the use of an appropriate conversion factor which would treat one of three places of articulation as more difficult than one of one.

The rudiments of the comparison method have been proposed by [9], [10], [11] and several other authors. The role of phonetic factors as determinants of stuttering has also been investigated. These authors have shown that a number of linguistic factors increase the likelihood that a speaker who stutters will experience difficulty. Early work examined the sentence position, the phone a word starts with, word length and word type. Later studies have also looked at the relationship of utterance length, sentence structure, clause structure and phrase structure with stuttering

In order to implement the present project that relates the structural complexities described above to precision of articulation in children following CLP surgery. it is necessary (1) to gather information on the phonetic structure of selected languages, (2) to develop a method for calculating the degree of phonetic difficulty for each natural class within the languages studied.

The assumption underlying the proposal offered here is an empirical observation made in the field of teaching foreign language phonetics, that the more sounds occur in a natural class (vowels, semi-vowels, laterals, nasals, plosives, fricatives and affricates), the narrower is the acceptable variability margin and, therefore, more precision is required in articulating each sound.

In this paper we are referring to the acquisition of native pronunciation to replace pathological speech. However, the difficulty in comparing the articulation of foreign sounds by a foreign learner will also vary depending upon the relations obtaining between individual sounds and larger sound structures.

Thus, for an English person, to learn each of the Polish affricates is a formidable task reached by only few percent of the learners. To learn the English affricate constitutes an incomparably easier task.

II. METHODS

The method used in this paper comprises a comparative analysis of the phonological subcomponents of the languages whose sounds we are comparing and the assessment of the articulation of each phoneme.

It needs to be emphasized that the present study is performed at the phonemic level. Yet certain allophones should also be considered in a future study, as the pronunciation of some of them makes pronunciation much more difficult than that of the principal variant.

In this paper, we also focus our attention on the paradigmatic view of the sound pattern, leaving clustering possibilities for a further part of the study.

The counting of the scores for phonemes will be done in the following way. The number of phonemes in a given sound class will be viewed against the background of average numbers of such sounds found in other languages. Any resulting "surplus" will increase the complexity of a given class.

The resulting score for the sum of all the natural classes will also show the overall complexity of the language's sound pattern.

III. RESULTS

Here are the results of a comparison of English and Polish natural classes:

1. Semi-vowels: English 2 and Polish 2.
2. Laterals: English 1, Polish 1.
3. Nasals: English 3, Polish 4 (because of the palatal nasal).
4. Plosives: English 3, Polish 3.
5. Fricatives: 5 in both languages.
6. Affricates: English 1, Polish 3.

Consider now Tab. 1 that shows the Complexity Index (CI) scores for English (Column 2), Polish (Column 3) and a general reference language (Column 4). Columns 5 and 6 show the scores of the values 2

and 3 relativized with regard to Column 4 in such a way that if the individual language score is larger than the Reference Score, then this surplus is written as a positive number in Column 6. If the language score and the Reference Score have the same values, Column 6 counts it as 0, and if the language score is a number smaller than that in the Reference Score, then the difference is shown as a negative number e.g. Row 6).

1	2	3	4	5	6
#	English	Polish	Reference Score	English Score	Polish Score
1	2	2	2	0	0
2	1	1	1	0	0
3	3	4	3	0	1
4	3	3	3	0	0
5	5	5	4	1	1
6	1	3	2	-1	1
SUM	15	18	15	0	3

Tab. 1. The consonant CI in English and Polish.

Thus the incorrect articulation of the English nasal /n/ and the affricate /tʃ/ have scored, respectively, 3-3=0, 1-2=-1 points. Similar sounds in Polish, i.e. /n/ and /tʃ/, will score 4-3=1 point and 3-2=1 point.

Hence, when studying the results of a speech therapy examination, to assure a correct comparison this study with other languages, the examiner should use appropriate values from Tab. 1 to raise or lower the obtained scores to bring a comparison into the realm of objectivized experimentation.

The comparisons are made separately for each natural class (1-6 in Tab.1).

IV. DISCUSSION

Despite the fact that a large number of books and articles comparing languages have been published since the beginning of Comparative Linguistics and throughout the later periods, and those publications have dealt with various levels of grammatical analysis including phonology and phonetics, almost none looked at the sound systems of two or more languages in order to carry out qualitative comparisons, especially those that made statements concerning the easy-difficult axis.

The present paper has arisen at the intersection of several motivational axes along the easy-difficult assessment axis, such as

- evaluation of the quality of pathological speech in children after CLP,
- assessing the progress in the acquisition of foreign language pronunciations,
- stuttering therapy carried out in various languages,

V. CONCLUSION

To conclude this paper, one should emphasize the fact that there is a lacuna of publications that would look at the phonetic structure of languages with an aim of a cross-language comparison for medical and pedagogical purposes.

Thinking of future developments of this research, the paper presents only one aspect of the comparison of sound patterns of languages, focusing on consonants. What remains to be studied in future in successive parts of the project is to work out the principles of comparing vowels and – at the prosodic level – the clustering possibilities, as well as a study of the order of the acquisition of the phonetic system of languages.

Studying the complexity of the articulatory movements of individual sounds, as well as the complexity of natural sound classes in a language, and the clustering possibilities of consonants – all these issues can account for the varying difficulty of acquiring languages that differ in phonological complexity [6], [7], [8].

REFERENCES

- [1] Jakielski K. J. Motor organization in the acquisition of consonant clusters. University of Texas at Austin: Ann Arbor Michigan; 1998. UMI Dissertation services. PhD thesis. Available through quotations in other papers.
- [2] P. Ladefoged and I. Maddieson, I. 1996. *The Sounds of the World's Languages*. Oxford: Blackwell.
- [3] Gonet, W., Szpyra-Kozłowska, J. i Świąciński, R. (2010) *Clashes with Ashes*. W: Waniek-Klimeczak, E. (Red.) *Issues W: Accents of English 2: Variability and Norm*. Cambridge: Cambridge Scholars' Publishing, 213-231.
- [4] Gonet, W., R. Gubrynowicz (1995). *The Sound Pattern of Polish*. In: *First SixMonthly Report; BABEL - A Multi-Language Database, COPERNICUS Project 1304*. Commission of the European Community, Reading University, September 1995, 11-21.
- [5] Gonet, W. i Trochymiuk, A. (2007) *Typologia procesów fonologicznych opracowana na podstawie materiału zawartego w bazie danych COPERNICUS 1304- BABEL PL*. W: T. Woźniak i A. Domagała (eds) *Język, interakcja, zaburzenia mowy. Metodologia badań* Lublin: Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej, 185-211.
- [6] Locke, J. 2008. Cost and Complexity: Selection for Speech and language. *Journal of Theoretical Biology* 251:640-652.
- [7] Maddieson, I. 2006. Correlating Phonological Complexity: Data and Validation. *Linguistic Typology* 10:106-123.
- [8] Pellegrino, F, Marsico, E, Chitoran, I, and Coupé, C eds. 2009. *Approaches to Phonological Complexity*. Berlin - New-York: Mouton De Gruyter.
- [9]. Howell, P., Au-Yeung, J., Yaruss, S., Eldridgge, K. (2006). *Cloin Linguist Phon*. 2006 Nov; 20(9): 703-716.
- [10]. Weiss A. L, Jakielski K. J. Phonetic complexity measurement and prediction of children's disfluencies: A preliminary study. In: Maassen B, Hulstijn W, Kent R, Peters HFM, van Lieshout PHMM, editors. *The 4th Nijmegen Conferences on Speech Motor Control and Stuttering*; Nijmegen: Uitgeverij Vantilt; 2001.
- [11]. Throneburg NR, Yairi E, Paden EP. The relation between phonological difficulty and the occurrence of disfluencies in the early stage of stuttering. *Journal of Speech and Hearing Research*. 1994;37:504–509.
- [12] *Explorations in the Acoustics of English Sounds* (2016). Maria Curie-Skłodowska University Press, 366 p.

THE EFFECT OF PHYSICAL ACTIVITY ON THE PHONETIC CHARACTERISTICS OF SPEECH

V. V. Evdokimova¹, E. A. Zakharchenko¹

¹ Saint Petersburg State University/Department of Phonetics, Saint-Petersburg, Russia
postmaster@phonetics.pu.ru
evgenya.97@mail.ru

Abstract: The characteristics of phonation in different physiological states is an important task, which is engaged in both medical and phonetic studies. This paper examines the impact of physical stress on speech from the phonetic point of view. The main objective of this study is to identify the phonetic features of speech in the state of physical fatigue and before it. The speech material includes 8 speakers (4 male and 4 female) in two states - before physical exercises and after. In both states speakers were asked to read texts. A three-step analysis (auditory, acoustic and statistical) was conducted. The results of the acoustic and auditory analysis showed dependency between acoustic parameters and speech under physical stress in terms of fundamental frequency, as well as the duration of the phonemes. This study also takes into account the intensity of breath as an important component of speech after a physical task. After physical activity, a greater number and variety of speech disruptions appear in the speakers' speech, breathing becomes more frequent and more intense, more pauses appear, the duration of the sounds changes, and such phenomena as noisy final vowels, amplitude jumps and voice tremor appear.

Keywords: speech after physical exercises, phonetics, phonation

I. INTRODUCTION

The presence of physical activity causes changes in the processes of speech production and, as a result, leads to changes in the speaker's speech. These changes can be observed in the calculation of acoustic and perceptual signs. That is, there is a definite connection between the state of physical fatigue and the resulting speech behavior [1].

The physiological and behavioral changes that result from a combination of physical activity and the process of speech production can depend on many factors. These factors include the age of the speaker, the level of his physical fitness, the peculiarity of physical exercise, the time of its completion, the degree of fatigue of the speaker, etc [2-6].

Physical activity affects the speech in different ways. During a physical activity the human body comes into a state of fatigue, which affects all life systems. The

speech production system is also under the influence of the load, changes appear during the generation of the air flow and during the work of the speech organs as resonators. Thus, physical activity influences all the processes of speech generation, which affect the human voice and the resulting speech signal.

The effects of physical activity on speech can be considered from the point of view of two different aspects:

- physiological aspect;
- phonetic aspect.

From a physiological point of view, stress affects a person's heartbeat, lung ventilation, and respiration [7]. The heartbeat, like breathing, becomes more frequent. All these changes can be considered from the phonetic side, the state of fatigue will be manifested in insufficient attention to the text to be read, the complexity of articulation, changes in pausing. As a result it will lead to changes in the acoustic and perceptual characteristics of the speech signal, to failures when reading the text.

The physical load can be created by natural and artificial methods. An artificial way involves intentionally providing physical exercises that the participants of the experiment perform before reading the material.

Many studies have long paid attention to how the processes of speech production and physical activity interact [9]. Most often, in these works a person's speech is recorded against the background of physical exercise, that is, they contain a criterion for the simultaneity of tasks [1].

Acoustic parameters are detected at the stage of acoustic analysis of the material, although often they are "barely perceptible". When studying a person's voice in different states, the following acoustic parameters are most often counted. The most studied parameter is the fundamental frequency of oscillations of the vocal cords, which depends on their length, thickness and voltage. Changes in the pitch of the voice are provided by the muscular apparatus of the larynx, from which it follows that the value of pitch can provide information about the functional and anatomical state of the larynx

at the moment. Intensity of the voice, jitter and shimmer are usually taken into account [10]. These phenomena are always presented in a person's speech, but their meanings can change with the physical load a person receives.

More complex parameters can be used to assess voice quality: signal to noise ratio, ration of different harmonics of voice, the ratio between the amplitude of the first harmonic and the amplitude of the second harmonic; harmonic richness factor - the ratio between the sum of the harmonics amplitudes and the amplitude of the frequency response component [1]. These parameters are sensitive to changes in the functioning of the vocal folds, and not to changes in higher areas of the vocal tract, that is, they form a deeper acoustic voice analysis.

In a state of physical fatigue, physiological breathing becomes more frequent, it becomes deeper and more intense. This leads to impaired speech breathing, and, as a consequence, to changes in the processes of speech production [11]. Speech in a state of physical fatigue can be characterized by both an increase in the duration of the statement, and its decrease. It depends on the individual characteristics of the speaker and his respiratory strategy. The first strategy is related to the speaker's attempt to pronounce the largest possible audio segment before the start of the next respiratory cycle. The second strategy is that the speaker prefers to take more frequent inhalations and exhalations, which reduces the duration of the statements. Such respiratory difficulties are connected with the fact that the speaker is in a state of physical fatigue and has difficulty breathing after exercise [12].

In a study by Keith W. Godin, John H. L. Hansen, the average frequency of the frequency response for each sound in two states was calculated [11]. It was found that the nasal consonants are most susceptible to changes.

Our paper examines the impact of physical stress on speech from the phonetic point of view. The main objective of this study is to identify the phonetic features of speech in the state of physical fatigue and before it.

II. METHODS

In this paper, we recorded the speech material and analyzed its phonetics properties. The records include 8 speakers aged between 18-26 (4 male and 4 female) in two states - before physical exercises and after it. Physical activity for the announcers was jogging for 10-15 minutes and squats. Immediately after receiving such a load, the announcers read the text in a state of physical fatigue.

In both states, speakers were asked to read texts, including sentences, text, and a poem. The material for reading was different for the two states, so as not to

cause addiction. Thus, 16 records were obtained for 8 speakers in two states. The duration of the audio file before exercise is an average of 2 minutes, after exercise - 3 minutes.

Then the segmentation of the speech was made on different speech units: phonemes, intonation units and pauses. All the speech failures were considered in detail. A three-step analysis (auditory, acoustic and statistical) was conducted.

It is important to note that by the breath period in this work we mean the audible stretch of breathing of any duration, which is due to the need to determine how much time the stray non-speech breathing takes away from the speaker. The level of speech failures is an indication of all speech errors and deviations that the speakers made when reading the text. After processing the collected material, its analysis was carried out. The first step is an auditory analysis to identify speech malfunctions and various phonetic phenomena in the speech of speakers in a state of physical fatigue. Acoustic analysis confirmed some assumptions regarding the change in the voice of the speakers. In the Praat program and using scripts for this program, the following were calculated:

- values of the fundamental
- values of jitter and shimmer;
- the values of the amplitude and temporal characteristics of breathing, pauses and phonetic units. Next, a statistical analysis of the data was carried out. For two speaker states, the mean value, standard deviation, minimum and maximum values for the following acoustic parameters were calculated:
- duration of different types of sounds;
- duration of pauses;
- duration of breath periods;
- intensity of breath periods;
- duration of pauses.

III. RESULTS

After auditory analysis, it was possible to immediately notice such phenomena as:

- the appearance of rapid and intensified inspirations and exhalations;
- higher voice announcers;
- higher rate of speech;
- the appearance of various speech malfunctions.

Acoustic and statistical analyzes were produced to confirm the first three phenomena. Speech failures were considered precisely in the framework of auditory analysis.

Two types of speech failures were found in the recordings: speech failure in violation of a language norm (error or deviation); speech failure as an independent phenomenon (false start).

When reading the text in the normal state, the speakers realized 18 speech malfunctions: 4 errors, 9 deviations from the language norm and 5 false starts, as an independent phenomenon. Auditory analysis of the recordings of speakers after physical activity made it possible to detect some phonetic. The main causes of such phenomena are the increased and increased respiration of the speakers, which greatly affects the speech signal. Under the influence of breathing, the noise of the final vowels is observed in the words, and this happens throughout the entire signal until breathing is restored. Also such phenomena as amplitude jumps and voice tremor appear frequently in the material. The following tables 1-6 and fig. 1-2 show the variation of acoustics parameters for speech before and after the physical load.

Table 1. Mean pitch before and after physical load

Speaker	Pitch (Hz)	
	Before exercises	After exercises
D1	216	228
D2	245	258
D3	199	207
D4	211	219
D5	112	119
D6	116	128
D7	105	109
D8	112	118

Table 2. Jitter before and after physical load

Speaker	Jitter (local %)	
	Before exercises	After exercises
D1	1.863	1.904
D2	1.832	1.764
D3	2.057	2.028
D4	1.888	2.025
D5	2.257	2.432
D6	3.130	2.817
D7	2.782	2.570
D8	2.358	2.482

Table 3. Shimmer before and after physical load

Speaker	Shimmer (local %)	
	Before exercises	After exercises
D1	6.884	7.786
D2	8.298	8.882
D3	8.646	10.216
D4	7.746	9.168
D5	8.847	9.577
D6	9.127	9.790
D7	11.052	10.955
D8	10.219	10.666

Table 4. Mean duration of stressed vowels

State vowel phoneme	Mean duration of stressed vowels (sec)	
	Before exercises	After exercises
/a/	0,094	0,089
/e/	0,079	0,091
/i/	0,083	0,068
/i/	0,070	0,068
/o/	0,082	0,076
/u/	0,077	0,084

Table 5. Mean duration of unstressed vowels

State vowel phoneme	Mean duration of unstressed vowels (sec)	
	Before exercises	After exercises
/a/	0,053	0,054
/e/	0,064	0,058
/i/	0,050	0,047
/i/	0,051	0,046
/o/	0,057	0,056
/u/	0,056	0,059

Table 6. Pause duration in speech after physical exercises

Speaker	Number of pauses	Pause duration (sec)				
		mean	St.dev	Min	Max	range
D1	86	0,708	0,540	0,052	2,758	2,706
D2	74	0,640	0,332	0,100	1,419	1,319
D3	58	0,855	0,833	0,076	3,286	3,210
D4	76	0,673	0,474	0,054	1,982	1,929
D5	51	0,599	0,436	0,047	2,121	2,075
D6	76	0,863	0,564	0,107	3,047	2,941
D7	96	0,521	0,463	0,054	2,062	2,008
D8	79	0,799	0,478	0,072	2,019	1,947

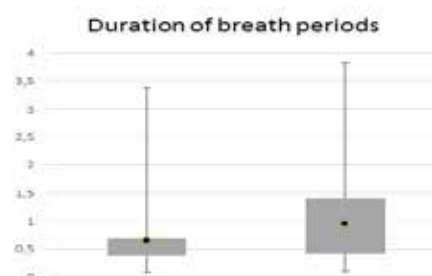


Figure 1. Duration of breath periods before and after physical stress (sec)

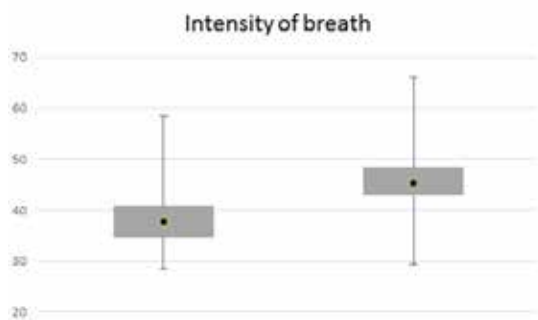


Figure 2. Intensity of breath periods (dB) before and after physical stress

IV. DISCUSSION

It is obvious that the human voice after exercises changes, because this is proved at all stages of the analysis of the material. After physical activity, a greater number and variety of speech disruptions appear in the speakers' speech, breathing becomes more frequent and more intense, more pauses appear, the duration of the sounds changes, and such phenomena as noisy final vowels, amplitude jumps and voice tremor appear. The results of this work can add to the knowledge about the difference between speech characteristics in a state of physical fatigue, voice fatigue and in terms of speech pathologies. The practical value of the work lies in the fact that the knowledge gained can be used in the development of systems for monitoring the human condition, where rapid and intensive breathing can mean a sharp deterioration in physical condition.

V. CONCLUSION

Based on the results of the study, the following conclusions can be drawn: A voice in a state of physical fatigue changes, as this is proved at all stages of the analysis of the material. After physical exertion, frequent inspirations and exhalations appear in speech, which have a longer duration and intensity, and, as a result, the speech breathing of the speakers gets lost. Pitch increases in a state of physical fatigue. Presumably, we can talk about the lengthening of stressed vowels /e/ and /i/, however, to identify trends it is necessary to increase the amount of analyzed material. After physical exertion, the announcers demonstrate a loss of attention, which manifests itself when reading the text. A significant number of hesitations and false starts appear in their speech. Lost speech breathing contributes to disruption of syntagmatic division and a change in Pitch movement. Increased and intensified breathing leads to noises, amplitude jumps and voice tension.

REFERENCES

- [1] K. W. Godin, "Analysis of the effects of physical task stress on the speech signal", *J. Acoust. Soc. Am.* 130. — 2011. — P. 3992-3998.
- [2] L. V. Bondarko, *Phonetics of Russian modern language*. Saint-Petersburg State University, 1998 (in Russian).
- [3] S. V. Kodzasov, and O. F. Krivnova, *General Phonetics*. Moscow, 2001.
- [4] G. / Fant, *Acoustic Theory of Speech Production*. Netherlands: Mouton, 1960.
- [5] J. L. Flanagan, "Source-system interaction in the vocal tract," *Ann. N.Y. Acad.Sci.* vol. 155, pp. 9-17, 1968.
- [6] J. L. Flanagan, *Speech Analysis, Synthesis, and Perception*. New York: Springer, 1972.
- [7] B. Johannes, P. Wittels, R. Enne, G. Eisinger, C. A. Castro, J. L. Thomas, A. B. Adler, and R. Gerzer, "Non-linear function model of voice pitch dependency on physical and mental load", *European Journal of Applied Physiology*, 101, 267-276
- [8] B. Schuller, F. Friedmann, Fl. Eyben, The Munich Biovoice Corpus: "Effects of Physical Exercising, Heart Rate, and Skin Conductance on Human Speech Production", LREC 2014.
- [9] S. P. A. Sanjay A. J. H. L. Hansen, "Detection of speech under physical stress: model development, sensor selection, and feature fusion", *INTERSPEECH-2008*, 817-820.
- [10] Fl. Hoenig, A. Batliner, El. Noeth, S. Schnieder, J. Krajewski, "Acoustic-prosodic characteristics of sleepy speech - Between performance and interpretation", *Proceedings of the International Conference on Speech Prosody*. — 2014. — P 864-868.
- [11] K. W. Godin, J. H. L. Hansen, "Physical task stress and speaker variability in voice quality", *EURASIP J. Audio, Speech and Music Processing* 2015. — TX, USA, 2015. — P. 29.
- [12] K. W. Godin, T. Hasan, J. H. L. Hansen, "Glottal Waveform Analysis of Physical Task Stress Speech", *INTERSPEECH-2012, Wed-SS6-15*. — Portland, OR23, 2012. — P. 1-4.

CHANGES TO VOICE PRODUCTION CAUSED BY LONG-TERM HEARING LOSS (HL)

I.V. Kastyro¹, A.N. Kovalenko¹, V.I. Torshin¹, E.S. Doroginskaya¹, N.A. Kamanina¹

¹ Department of Physiology, Peoples' Friendship University of Russia (RUDN University), Moscow, Russia
ikastyro@gmail.com; rugolospro@gmail.com; vtorshin@mail.ru; lenucca83@list.ru; ychenick@mail.ru

Abstract: A new procedure of vowel acoustic space (VAS) transformation for the purpose of characterization of vowel production in individuals with HL was developed.

Recordings of sustained Russian cardinal vowels /a/, /i/, /u/ of 5 women (4 with HL and 1 without HL) and 5 men (3 with HL and 2 without HL) were acoustically analyzed. For each participant, two first formants of each vowel were measured and log-transformed ($\log F_1$, $\log F_2$). VAS was transformed into right triangles, their /u/ corners were moved to the origin, and their legs were aligned with axes.

VAS was almost symmetrical, equal and have a maximum size in the control group consisted of subjects with no HL while these of hearing-impaired group tended to have reduced size and to stretch along one axis.

Our study showed that a new VAS normalization approach can distinguish at least three groups of people with HL. There are those with vowel triangles stretched along $\log F_1$ -axis, with vowel triangles stretched along $\log F_2$ -axis, and with symmetrical vowel triangles. Causes of the VAS differences require further investigation.

Keywords: Hearing loss, vowel acoustic space, acoustic analysis

I. INTRODUCTION

Study of VAS is an important noninvasive tool for assessing speech development and disorders. VAS size is an acoustic metric which proved to be effective in the quantification of articulatory function in adults with psychological and neurological disorders [1, 2], and children [3].

VAS area is affected not only by articulatory subsystem problems but also by hearing impairment since hearing provides the necessary feedback for speech production control [4]. The vowel space in individuals with hearing impairment is often described as reduced [5].

Although VAS triangles area is well established and widely used index for vowel acoustic research,

some studies demonstrated its inability to differentiate some cases of speech disorders. To improve differentiation, some new acoustic metrics were developed. Although researchers mainly focused on dysarthric speakers [6, 7].

Despite the considerable amount of literature concerning vowel space, little research effort has been spent on the speech of Russian adult subjects with HL.

The aim of this paper is to characterize VAS in hearing-impaired Russian adults using newly developed acoustic metrics.

II. METHODS

Subjects and voice sampling: Ten native Russian speakers aged between 20 and 40 years (5 men, 5 women) completed the informed consent process and participated in the current study. The experimental group was composed of seven participants (4 women and 3 men, 30.7 ± 6.3 years) diagnosed with profound sensorineural HL with no concomitant disorders. All of them except one man reported of having no more than 1-year experience using a single-sided hearing aid. The control group was composed of three participants (1 woman and 2 men, 26.3 ± 6.5 years) without hearing complaints. The unaided hearing threshold levels of the subjects were measured with a clinical audiometer (AA220, Interacoustics/Denmark) in a sound-treated booth at pure-tone threshold levels of 500; 1,000; 2,000; and 4,000 Hz. The hearing thresholds for both ears at all frequencies were averaged to represent the average hearing threshold for each subject.

All subjects were asked to sustain prolonged Russian cardinal vowels /a/, /i/, and /u/ at a comfortable pitch in their typical speaking voice mode at an individually comfortable voice intensity level.

The voice signals were recorded using a condenser cardioid microphone with flat frequency response from 40 Hz to 10 kHz at mouth-to-microphone distance 30 cm (Behringer C-1, Behringer/Germany). The signals were digitally sampled and recorded using an IBM personal computer compatible sound adapter at the rate of 44.1 kHz.

The F_1 – F_2 formants pairs for each vowel were computed from medial 3 seconds of the steady-state of

vowel production using PRAAT[8]. Then, formants values were log-transformed and applied for VAS triangles construction and transformation.

VAS triangles transformation: The following procedure was used to transform VAS triangles.

First, the origin of the formants coordinate system was translated to the /u/ corner of a VAS triangle:

$$\begin{aligned} \log F_1 a'' &= \log F_1 a - \log F_1 u, \\ \log F_2 a'' &= \log F_2 a - \log F_2 u, \\ \log F_1 i'' &= \log F_1 i - \log F_1 u, \\ \log F_2 i'' &= \log F_2 i - \log F_2 u, \\ \log F_1 u'' &= 0; \log F_2 u'' = 0. \end{aligned} \tag{1}$$

Second, axes were rotated to align the logF₂-axis with the /u/-/i/ side of a VAS triangle:

$$\begin{aligned} \log F_1 a' &= \log F_1 a'' \cos \varphi - \log F_2 a'' \sin \varphi, \\ \log F_1 i' &= 0, \\ \log F_2 i' &= \log F_1 i'' \sin \varphi - \log F_2 i'' \cos \varphi, \end{aligned} \tag{2}$$

where φ is the angle between the logF₂-axis and the /u/-/i/ side of a VAS triangle, and

$$\begin{aligned} \cos \varphi &= \frac{\log F_2 i''}{\sqrt{(\log F_1 i'')^2 + (\log F_2 i'')^2}}; \\ \sin \varphi &= \frac{\log F_1 i''}{\sqrt{(\log F_1 i'')^2 + (\log F_2 i'')^2}} \end{aligned}$$

The final step was to place /a/ corner of a VAS triangle on the logF₁-axis:

$$\log F_2 a' = 0. \tag{3}$$

All three steps depicted in Fig. 1

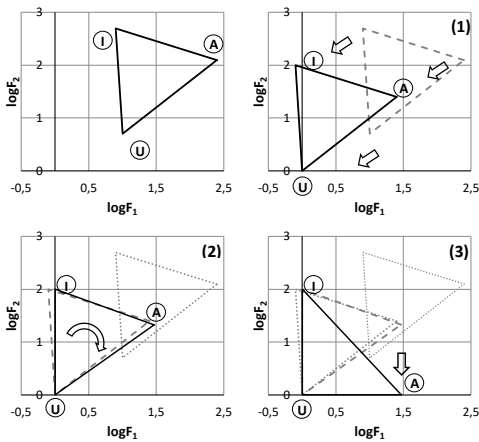


Figure 1. Schematized representation of the three steps of the VAS triangles transformation.

The transformation described above preserves areas of original VAS triangles and makes two other metrics – Euclidian distance between /u/ and /i/ corners, and Euclidian distance between /a/ corner and /u/-/i/ side – more prominent. Now, the first of them is determined with respect to the new coordinate system as logF₂i'. The second is represented by logF₁a'. A VAS triangle area is simply obtained by multiplication of these values. logF₂i'/logF₁a' is a measure of a VAS triangle symmetry.

III. RESULTS

Transformed VAS triangles are shown in Fig. 2 compared to literature reference data.

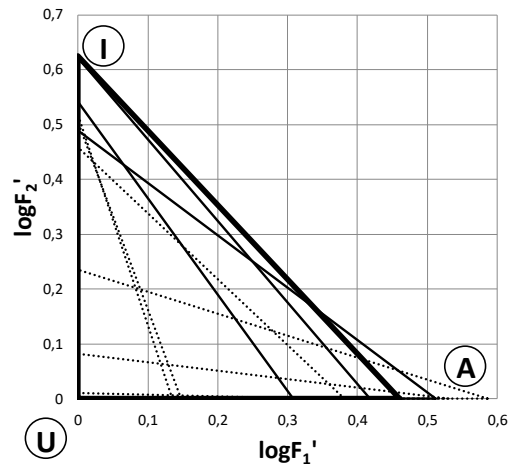


Figure 2. Transformed VAS triangles. Dotted lines – subjects with HL, solid lines – subjects without HL, thick line – literature reference [9].

Two women with severe HL have so small areas of VAS triangles that they are almost invisible on the present scale. One man who has been using hearing aids for a long time exhibits vowel space area almost as big as in healthy subjects. The rest of the experimental group shows smaller triangles stretched along one of the axes.

Control group VAS triangles are comparable with the reference one. One man from the control group shows somewhat reduced vowel space as well as a minor HL.

Average HL was plotted against VAS. Linear regression analysis revealed a strong linear relationship as demonstrated in Fig. 3.

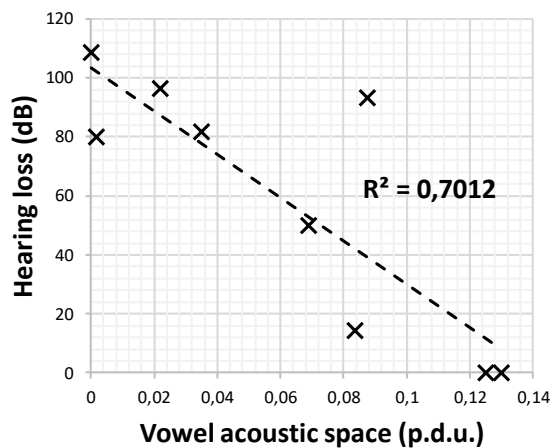


Figure 3. Relationship between HL and VAS.

Table 1 summarize data on hearing impairment and vowel space metrics.

Table 1. Hearing loss and vowel space metrics in subjects.

	HL (dB)	VAS	$\log F_{2i}'/\log F_{1a}'$
Experimental group			
E_f_1	96.3	0.022	0.16
E_f_2	50.0	0.069	0.40
E_f_3	108.8	0.000	35.29
E_f_4	80.0	0.002	0.03
E_m_1	81.9	0.035	3.85
E_m_2	n/a	0.037	3.44
E_m_3	93.1	0.087	1.21
Control group			
C_f_1	0.0	0.130	1.49
C_m_1	0.0	0.125	0.96
C_m_2	14.4	0.084	1.76
Reference[9]	0.0	0.144	1.36

IV. DISCUSSION

There were only a few studies providing an evaluation of VAS in hearing-impaired adults [10, 11]. Nevertheless, most papers both on adults and children with HL report decrease of VAS area[5, 10, 11], as also confirmed by this research. Log-transformed VAS area proved to be susceptible to a slight decrease of auditory acuity in a man with no hearing complaints. Also, VAS size metric differentiated a man with long experience of using hearing aids from the rest of the experimental group.

Other researchers didn't report any noticeable intragroup differences in HL subjects. Our VAS transformation procedure highlighted that HL subjects, probably, were not a homogeneous group. Further study is required to clarify these findings.

V. CONCLUSION

This study confirms that hearing impairment is positively correlated with a VAS decrease.

The transformation procedure revealed that the VAS didn't decrease uniformly in subject with HL but tended to diminish to a greater extent along one coordinate axis or another.

Causes for this observation couldn't be clearly determined in the present work due to small groups sizes.

The VAS in healthy people is larger and more symmetrical.

ACKNOWLEDGEMENTS

The publication has been prepared with the support of "RUDN University Program 5-100".

REFERENCES

- [1] S. Scherer, L. Morency, J. Gratch, and J. Pestian, "Reduced vowel space is a robust indicator of psychological distress: A cross-corpus analysis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4789–4793.
- [2] J. A. Whitfield and A. M. Goberman, "Articulatory-acoustic vowel space: Application to clear speech in individuals with Parkinson's disease," *J. Commun. Disord.*, vol. 51, pp. 19–28, 2014.
- [3] H. K. Vorperian and R. D. Kent, "Vowel Acoustic Space Development in Children: A Synthesis of Acoustic and Anatomic Data," *J. Speech Lang. Hear. Res.*, vol. 50, no. 6, p. 1510, Dec. 2007.
- [4] M. Lejska, "Voice field measurements—a new method of examination: the influence of hearing on the human voice," *J. Voice*, vol. 18, no. 2, pp. 209–215, Jun. 2004.
- [5] J. Verhoeven, O. Hide, S. De Maeyer, S. Gillis, and S. Gillis, "Hearing impairment and vowel production. A comparison between normally hearing, hearing-aided and cochlear implanted Dutch children," *J. Commun. Disord.*, vol. 59, pp. 24–39, 2016.
- [6] S. Sapir, L. O. Ramig, J. L. Spielman, and C. Fox, "Acoustic metrics of vowel articulation in Parkinson's disease: Vowel space area (VSA) vs. Vowel articulation index (VAI)," *Model. Anal. vocal Emiss. Biomed. Appl.*, vol. 9, pp. 173–175, 2011.

- [7] N. Roy, S. L. Nissen, C. Dromey, and S. Sapir, "Articulatory changes in muscle tension dysphonia: Evidence of vowel space expansion following manual circumlaryngeal therapy," *J. Commun. Disord.*, vol. 42, no. 2, pp. 124–135, 2009.
- [8] P. Boersma and V. van Heuven, "Speak and unSpeak with Praat," *Glott Int.*, vol. 5, no. 9–10, pp. 341–347, 2001.
- [9] M. F. Derkach, R. Y. Gumetskii, B. M. Guba, and M. E. Chaban, *Dinamicheskie spektry rechevykh signalov (The dynamic spectrograms of speech signals)*. Lvov: Vysshaya Shkola, 1983.
- [10] K. Nicolaidis and A. Sfakiannaki, "An acoustic analysis of vowels produced by greek speakers with hearing impairment," *ICPhS*, no. August, pp. 1969–1972, 2007.
- [11] E.-A. Choi and C.-J. Seong, "The Articulation Characteristics of the Profound Hearing-Impaired Adults' Korean Monophthongs: with Reference to the F1, F2 of Acoustic Vowel Space," *Phonetics Speech Sci.*, vol. 2, no. 4, pp. 229–238, 2010.

COMPARISON OF IMMEDIATE EFFECTS OF VOCAL BREATHING EXERCISES AND PHYSICAL EXERCISES ON HEART RATE VARIABILITY IN HEALTHY STUDENTS

A.N. Kovalenko¹, I.V. Kastyro¹, V.I. Torshin¹, Y.S. Guhschina¹, E.S. Doroginskaya¹, N.A. Kamanina¹

¹ Department of Physiology, Peoples' Friendship University of Russia (RUDN University), Moscow, Russia
rugolospro@gmail.com; ikastyro@gmail.com; vtorshin@mail.ru; gushchina@mail.ru; lenucca83@list.ru; ychenick@mail.ru

Abstract: To improve voice, voice therapists use breathing exercises meant for changing breathing patterns in a person. It is well known that there is heart rate variability (HRV) in synchrony with respiration called respiratory sinus arrhythmia (RSA). Therefore, breathing exercises may influence not only voice but also cardiovascular system.

Our study was aimed to detect any influence of vocal breathing exercises HRV and to compare it to that of physical exercises using short-term ECG recording.

3-minutes ECG records were made before and after squat-stands and breathing exercises for 13 healthy students (8 women, 5 men) aged from 19 to 21. As the most relevant parameters, rMSSD, LF, HF, LF/HF were extracted from ECG records.

We found two types of cardiovascular response based on rMSSD. The first type of response is associated with post-squat-stand rMSSD decrease and post-breathing rMSSD increase. The second type of response is manifested in the opposite effects. Unfortunately, obtained data did not let to uncover causes these phenomena.

Keywords: Breathing exercises, heart rate variability, vocal exercises

I. INTRODUCTION

Heart rate (HR) in humans fluctuates in phase with the respiration cycle. Usually, HR increases during inspiration and decreases during expiration. This phenomenon is referred to as respiratory sinus arrhythmia (RSA) [1].

Changes in respiratory frequency and depth of ventilation that commonly occur during different behavioural tasks (e.g., mental arithmetic, bicycle ergometry, exposure to anxiety-inducing stimuli) can significantly affect RSA levels [2].

Voice therapy and vocal training usually involve an alteration to breathing pattern of a subject to improve phonatory output [3]–[5].

Hence, vocal breathing exercises can produce a complex of integrated respiratory and cardiovascular responses.

In our study, we tried to uncover the impact of breathing exercises on the cardiovascular system in comparison to simple physical exercises applying short-term ECG.

II. METHODS

Subjects: Thirteen healthy young adult (8 female and 5 male 19.8 ± 0.6 years, body mass index: 22.9 ± 5.2 kg/m²) subjects were enrolled for this study. All subjects had a clear history of cardiovascular, respiratory and cerebrovascular diseases and were not taking any form of medication, and had abstained from exercise, caffeine and alcoholic beverages for at least 12 h before data collection.

We randomly divided participants into Group1 (3 female and 3 male 19.8 ± 0.4 years, body mass index: 23.1 ± 3.9 kg/m²) and Group2 (5 female and 2 male 19.9 ± 0.7 years, body mass index: 22.6 ± 6.4 kg/m²).

The experimental design is summarised in Fig. 1.

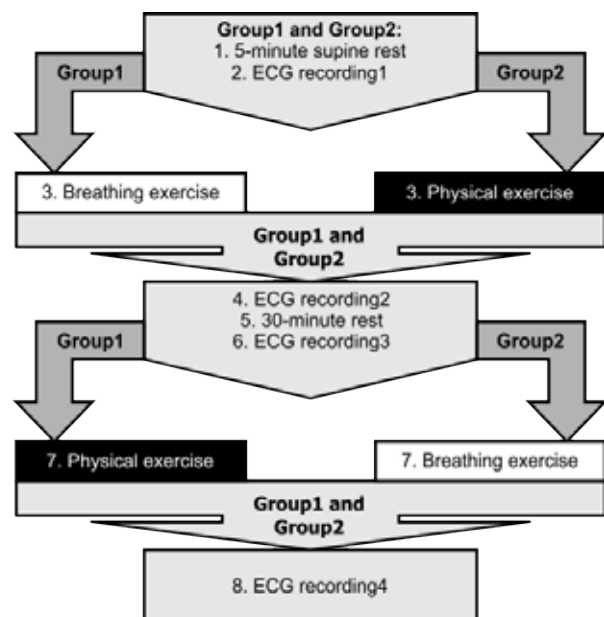


Figure 1. Flow chart of the experimental design.

Two groups were formed to find out if the sequence of exercises made a difference.

EGG recordings: 3-minutes ECG records were obtained in the supine position by using three-lead Biopac MP30B-CE and analysed by Kubios HRV 2.1 software. As the most relevant parameters, we extracted rMSSD, LF and HF from ECG records [6], [7]. Statistical analyses were performed using JASP version 0.10.2.

Physical exercises: Squat-stand manoeuvres (SSM) were used as a physical load. The squat-stand cycle was performed at frequency 0.17 Hz (3 s squatting, followed by 3 s standing) for 3 min.

Breathing exercises: A periodic loud pronunciation of /s/ sound were employed as a breathing exercise. It was performed in the same manner: 0.17 Hz (3 s breathing out on loud /s/, followed by 3 s of natural breathing) for 3 min.

III. RESULTS

A. Intergroup comparison before an experiment

We compared HRV data of the groups before our experiment using Mann-Whitney U-test and detected no significant difference ($p > 0.05$). The comparison suggested that groups were equal concerning their cardiovascular properties. The data are presented in Table 1.

Table 1. Preexperimental intergroup HRV comparison.

rMSSD1 (ms)	LF1 (n.u.)	HF1 (n.u.)	LF1/HF1
Group1			
47.2±22.2	50.9±16.3	49.1±16.3	1.2±0.7
Group2			
36.2±15.7	62.6±14.1	37.4±14.1	2.0±1.2

B. Intergroup comparison before exercises

Table 2. Intergroup HRV comparison before breathing exercises.

sp-rMSSD (ms)	sp-LF (n.u.)	sp-HF (n.u.)	sp- LF/HF
Group1			
47.2±22.2	50.9±16.3	49.1±16.3	1.2±0.7
Group2			
39.9±13.4	52.0±13.5	48.0±13.5	1.3±0.8

Also, intergroup comparison of HRV before a breathing exercise and physical exercise was

performed using Mann-Whitney U-test. The data could be found in Table 2 and Table 3.

No significant differences were found. Therefore we decided to combine the data of both groups according to the type of exercise.

Table 3. Intergroup HRV comparison before physical exercises.

ex-rMSSD (ms)	ex-LF (n.u.)	ex-HF (n.u.)	ex- LF/HF
Group1			
48.8±16	56.9±24.9	43.1±24.9	2.8±3.7
Group2			
36.2±15.7	62.6±14.1	37.4±14.1	2.0±1.2

B. Before and after exercises HRV

For every subject, differences of rMSSD, LF, HF and LF/HF before and after exercises were calculated. The values are shown in Table 4. Differences for HF are not presented since they are equal to LF differences with the sign reversed.

Table 4. HRV metrics differences.

Breathing exercise			
Gender	d(sp- rMSSD)	d(sp-LF), n.u.	d(sp- LF/HF)
f	-18,5	-3,1	-0,1
m	-16,4	-10,7	-0,2
m	-13,3	11,9	0,4
f	-5,9	-5,5	-0,3
f	-4,1	4,7	0,3
f	-1,6	14,7	1,0
f	1,5	-3,6	-0,5
f	1,9	-8,2	-0,6
f	9,3	-10,1	-0,2
m	11,7	6,0	0,2
f	15,6	-4,3	-0,2
m	15,8	-10,4	-0,8
m	22,3	17,4	1,3
Physical exercise			
Gender	d(ex- rMSSD)	d(ex-LF), n.u.	d(ex- LF/HF)
f	4,0	-25,1	-0,7
m	2,7	-25,9	-0,7
m	0,1	-31,2	-1,4
f	2,5	-6,0	-1,2
f	-0,2	-6,6	-0,5
f	-7,8	11,8	0,4

f	-7,0	-1,5	-0,2
f	-9,2	-3,6	-0,4
f	-0,5	26,5	0,7
m	-4,8	7,3	0,7
f	-5,0	28,0	1,2
m	-15,5	-11,7	-0,9
m	-15,4	-45,5	-9,2

Linear regression revealed a moderate relationship between rMSSD differences for breathing $d(\text{sp-rMSSD})$ and physical $d(\text{ex-rMSSD})$ exercises (Fig. 2).

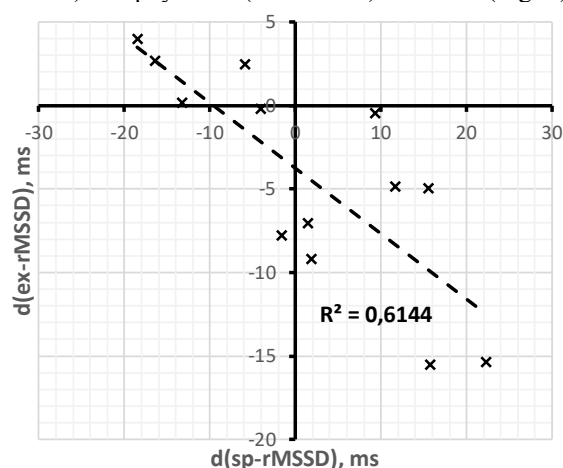


Figure 2. Relationship between rMSSD difference for a physical exercise $d(\text{ex-rMSSD})$ and rMSSD difference for a breathing exercise $d(\text{sp-rMSSD})$.

This finding mostly means that subjects with a post-squat-stand increase of rMSSD demonstrated a post-breathing decrease of rMSSD and vice versa.

IV. DISCUSSION

As can be seen from Table 4, both physical and respiratory exercises lead to a diminution of LF and LF/HF in most participants (9 out of 13). LF/HF ratio is usually used to assess the sympathovagal balance, and its decrease means augmentation of parasympathetic dominance[8]. However, spectral measures are highly sensitive to technical errors within RR data and ways of their correction [9] and should be taken with caution.

The rMSSD reflects the beat-to-beat variance in HR and is the primary time-domain measure used to estimate the vagally mediated changes reflected in HRV [10]. Researchers have proposed it as reliable metrics for short-term ECG [11]. The negative difference between the pre and post-exercise rMSSD results is due to less variability across time in R-R intervals following a bout of exercise, since

parasympathetic modulation is reduced. The positive differences mean the opposite effect.

Taking into consideration the above statements, Fig. 2 is likely to be explained as follows. Breathing exercise induces a raising of cardiac parasympathetic modulation in those subjects who show sympathetic activation after physical exercise and vice versa.

V. CONCLUSION

The sequence of exercises did not affect HRV metrics in healthy students.

No significant differences were revealed between men and women.

Two types of cardiovascular response based on rMSSD were found.

First, participants who showed post-squat-stand rMSSD decrease tend to have post-breathing rMSSD increase.

Second, a reverse effect was found in the rest of the participants.

These findings call for further investigation to ascertain whether they are reproducible.

ACKNOWLEDGEMENTS

The publication has been prepared with the support of "RUDN University Program 5-100".

REFERENCES

- [1] P. Grossman and E. W. Taylor, "Toward understanding respiratory sinus arrhythmia: Relations to cardiac vagal tone, evolution and biobehavioral functions," *Biol. Psychol.*, vol. 74, no. 2, pp. 263–285, 2007.
- [2] F. H. Wilhelm, P. Grossman, and M. A. Coyle, "Improving estimation of cardiac vagal tone during spontaneous breathing using a paced breathing calibration," *Biomed. Sci. Instrum.*, vol. 40, no. January 2014, pp. 317–24, 2004.
- [3] S. L. Schneider and R. T. Sataloff, "Voice Therapy for the Professional Voice," *Otolaryngol. Clin. North Am.*, vol. 40, no. 5, pp. 1133–1149, Oct. 2007.
- [4] R. B. Goldenberg, "Singing Lessons for Respiratory Health: A Literature Review," *J. Voice*, vol. 32, no. 1, pp. 85–94, Jan. 2018.
- [5] A. P. Mendes, W. S. Brown, C. Sapienza, and H. B. Rothman, "Effects of Vocal Training on Respiratory Kinematics during Singing Tasks," *Folia Phoniatr. Logop.*, vol. 58, no. 5, pp. 363–377, 2006.
- [6] F. Shaffer and J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms," *Front. Public Heal.*, vol. 5, no. September, pp. 1–17, 2017.
- [7] J. L. Elghozi and C. Julien, "Sympathetic control of short-term heart rate variability and its

pharmacological modulation,” *Fundam. Clin. Pharmacol.*, vol. 21, no. 4, pp. 337–347, 2007.

[8] A. Malliani, M. Pagani, F. Lombardi, and S. Cerutti, “Cardiovascular neural regulation explored in the frequency domain,” *Circulation*, vol. 84, no. 2, pp. 482–492, 1991.

[9] D. Nunan, G. R. H. Sandercock, and D. A. Brodie, “A quantitative systematic review of normal values for short-term heart rate variability in healthy adults,” *PACE - Pacing Clin. Electrophysiol.*, vol. 33,

no. 11, pp. 1407–1417, 2010.

[10] F. Shaffer, R. McCraty, and C. L. Zerr, “A healthy heart is not a metronome: an integrative review of the heart’s anatomy and heart rate variability,” *Front. Psychol.*, vol. 5, no. September, pp. 1–19, 2014.

[11] M. R. Esco and A. A. Flatt, “Ultra-short-term heart rate variability indexes at rest and post-exercise in athletes: Evaluating the agreement with accepted recommendations,” *J. Sport. Sci. Med.*, vol. 13, no. 3, pp. 535–541, 2014.

AN ALGORITHM FOR DETECTING THE ONSET OF LINGUISTIC SEGMENTS IN CONTINUOUS ELECTROENCEPHALOGRAM SIGNALS

C. Tonatiuh Hernández-del-Toro¹, Carlos A. Reyes-García²

Computer Science Department, INAOE, Puebla, México

¹tonahdz@inaoep.mx, ²kargaxxi@inaoep.mx.

Abstract: A Brain Computer Interface based on imagined words can decode the word a subject is thinking on through brain signals to control an external device. In order to build a fully asynchronous Brain Computer Interface based on imagined words in electroencephalogram signals as source, we need to solve the problem of detecting the onset of the imagined words. Although there has been some research in this field, the problem has not been fully solved. In this paper we present an approach to solve this problem by using values from statistics, information theory and chaos theory as features to correctly identify the onset of imagined words in a continuous signal. On detecting the onsets of imagined words, the highest True Positive Rate achieved by our approach was obtained using features based on the generalized Hurst exponent, this True Positive Rate was 0.69 and 0.77 with a timing error tolerance region of 3 and 4 seconds respectively.

Keywords: Imagined words, Onset Detection, Continuous Signal.

I. INTRODUCTION

A Brain Computer Interface (BCI) based on imagined speech (hearing self-voice internally without any muscular movement), can decode the syllable, vowel or word a subject is thinking on through brain signals in order to control an external device, this technology can be used to improve life conditions of persons with disabilities and to improve our natural human abilities in many technological fields.

Imagined speech as electrophysiological source has shown very promising results in recent years, however, there are still some problems that are not fully solved [1]. Imagined speech can be exploited in many ways (syllable, vowel or word). Among all these, we will focus in imagined words.

In the task of building a fully asynchronous BCI that takes electroencephalogram (EEG) signals as input and uses imagined words as electrophysiological source, many problems need to be solved, among them, is the problem of detecting the onset of

imagined words in a continuous signal, this is, to correctly identify when the user starts to imagine a word. This requires to correctly classify between mental linguistic activity and idle states. This problem, if solved, can lead to a BCI based on imagined words that is activated in the exact moment the user desires it to.

To the best of our knowledge, there has not been an approach to solve the problem of identifying the onset of imagined words in continuous EEG signals. However, in [2] they tried to identify the onset of high pitch sound imagery production.

The present work describes an algorithm to detect the onset of imagined words in a continuous EEG signal using common statistical values, the Shannon entropy and the generalized Hurst exponent as features. The dataset used in this work is described in [3], which has recordings of 27 subjects imagining 5 different words in Spanish needed to control a pc pointer.

The algorithm proposed uses linguistic and nonlinguistic segments to train a classifier for latter sequential evaluation on a continuous signal to predict where the onsets of imagined words are. The measure of performance is the True Positive Rate (TPR). This is, the number of onsets that are correctly identified divided by the total number of onsets.

For the feature set based on statistical values, our algorithm achieves an average TPR of 0.65 and 0.73 with a timing error tolerance region (TETR) of 3 and 4 seconds respectively for the detection of the onsets.

For the features based on the generalized Hurst exponent, our algorithm achieves an average TPR of 0.69 and 0.77 with a TETR of 3 and 4 seconds respectively for the detection of the onsets.

II. METHODS

A. Common Average Reference

The Common Average Reference (CAR) method, is a technique used in digital signal processing to obtain a higher signal to noise ratio in signals. It consists on subtracting from each sample the

information that is present in all channels, it is defined as

$$V_i^{CAR} = V_i - \frac{1}{n} \sum_{j=1}^n V_j^{ER}, \quad (1)$$

where V_j^{ER} is the potential between the i^{th} electrode and the reference, and n is the number of electrodes.

B. Statistical values

Eighth values are used as features, four of them are statistical values commonly known: Mean (μ), Max, Min, Sum. And four of them are values not commonly known: Skewness (μ_3), Kurtosis (μ_4), Shannon entropy $S(x)$ and generalized Hurst exponent $H(q)$.

Skewness: Is a measure of how asymmetrical a distribution around the mean is. For a set $X = \{x_1, x_2, \dots, x_n\}$, skewness is defined as:

$$\mu_3 = \frac{\mu_3}{\sigma^3} = \frac{E[(X - \mu)^3]}{(E[(X - \mu)^2])^{3/2}}. \quad (2)$$

Kurtosis: Is a measure of the shape of a distribution curve around the mean. For a set $X = \{x_1, x_2, \dots, x_n\}$, kurtosis is defined as:

$$\mu_4 = \frac{\mu_4}{\sigma^4} = \frac{E[(X - \mu)^4]}{(E[(X - \mu)^2])^{4/2}}. \quad (3)$$

Skewness and kurtosis also can be defined as the 3th and 4th standardized moments.

Shannon Entropy: In information theory, the entropy $S(x)$ is the average rate at which information is produced by a stochastic source of data. The higher the Shannon entropy, the bigger the information is given by a new value in the process. It is also defined as the number of necessary bits to encode the information in the process. For a time series $X = \{x_1, x_2, \dots, x_n\}$, entropy is defined as:

$$S(X) = -\sum_{i=1}^N p(x_i) \log_2(p(x_i)), \quad (4)$$

where $p(x_i)$ is the probability of getting x_i .

Generalized Hurst Exponent: The Generalized Hurst Exponent $H(q)$ [4, 5, 6], is used in time series analysis and fractal analysis as a measure of scaling properties by the q^{th} order moments of the distribution of the increments. For a time series $X = \{x_1, x_2, \dots, x_n\}$, the generalized Hurst exponent can be obtained from the relations in eq. (5) and eq. (6)

$$K_q(\tau) \sim \left(\frac{\tau}{\nu}\right)^{qH(q)}, \quad (5)$$

$$K_q(\tau) = \frac{\langle |X(t+\tau) - X(t)|^q \rangle}{\langle |X(t)|^q \rangle}, \quad (6)$$

Where $K_q(\tau)$ is given from $X(t)$, with $t = \nu, 2\nu, \dots, k\nu, T$ (scale observation period T , and time resolution ν).

For $q = 1$, the generalized Hurst exponent is closely related with the original Hurst exponent, which measures how chaotic or unpredictable is a series. Original Hurst exponent is related with fractal dimensions and has been in the study of seizures in the temporal lobe in animals with EEG [7].

III. RESULTS

The objective of this experiment is to identify the onset of imagined words in continuous EEG signals. The dataset of signals consists of recordings of 27 subjects described on [3]. Each element (subject) has 5 signals recordings, each of them containing 33 repetitions of one of 5 imagined words in Spanish: “arriba”, “abajo”, “izquierda”, “derecha” and “seleccionar”. Which mean: up, down, left, right and select respectively. The process is performed in two stages:

Training Stage: First, from each subject, 5 Folds are made with the 5 signals by extracting one of them and using it as test signal, and the other four signals are kept for the training stage. This procedure is made 5 times in order to obtain a cross validation scheme (Each fold contains 4 signals for training and 1 signal for testing).

In each of the signals, we know *a priori* the markers of the beginning and ending of each imagined word, these markers are going to help us to collect a training corpus and to compare the predicted onsets of the classifiers with the real onsets to calculate a TPR.

Then, for each fold, from the train signals, two types of segments are selected:

- **Imagined words segments:** From the markers of each repetition of the word, 2 instances are obtained. The first instance is obtained by taking a window of 128 samples right to the marker of beginning of each imagined word. The second segment is obtained by taking a window of 128 samples left to the marker of the ending of each imagined word. From the 33 repetitions, the first and the last are discarded, thus obtaining from the 4 training signals, 248 instances labeled as imagined words.
- **Idle states segments:** Similarly to the imagined word segments, from the markers of each repetition of the word, 2 non-linguistic segments are obtained. The first segment is obtained by taking 128 samples left to the marker of the beginning of each imagined word. The second segment is obtained by taking 128 samples right to the marker of the ending of each imagined word. From the 33 repetitions, the first and the last are also discarded, thus obtaining from the 4 training signals 248 instances labeled as idle states.

This process yields 496 instances for each subject, 248 are imagined words and 248 are idle states. This scheme is better illustrated in fig. 1.

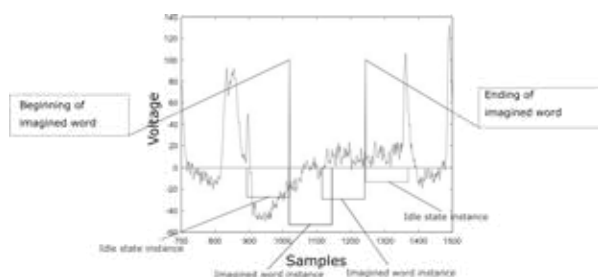


Fig. 1. Example of extraction of segments

In each of the 496 instances, noise reduction is performed by applying the CAR method described before and then, two feature sets are obtained. The first feature set is built by calculating 8 values per channel: Mean, Max, Min, Kurtosis, Skewness, Sum of all the samples per channel, Entropy and generalized Hurst exponent with $q = 1$. This feature extraction method yields 112 features per instance: 8 features per each channel (14). The second feature set is built by extracting the generalized Hurst exponent for $q = 1, 2, 3, 4, 5$ of each channel. This feature

extraction method yields 70 features per instance. Two Random Forest classifiers are separately trained for each feature set.

Testing Stage: With the remaining signal of the fold in each feature set, we test our classifiers by following the steps below.

1. The test signal is segmented into windows of 1 second (128 samples), each window is taken sequentially and with no overlap.
2. Then, each 1 second window is also preprocessed with CAR and the feature extraction method of each feature set is respectively calculated as we did with the training samples, creating with this an instance for each window.
3. Then the instances are evaluated with the Random Forest classifier of each feature set to define if it is an imagined word or an idle state.
4. For each classified window, a value is given from the classification: 0 if it is an idle state and 1 if it is an imagined word. The results are appended sequentially into a vector. With this vector, we calculate the onset and ending of each imagined word by unifying all 1's into imagined words and all 0's into idle states, with this, we calculate the predicted onset in time as $n \cdot 128$, being n the position of the first 1 from an imagined word in the vector of sequential classifications.
5. The predicted onsets of imagined words are compared with the true onsets to see how far the classification was. This is made by setting up a TETR [3] of 3 seconds (384 samples) and 4 seconds (512 samples) to verify if the onset is inside the TETR window in order to consider it as a True Positive.

This procedure is repeated for the 5 folds and the TPR is averaged.

Results of the onsets identifications

The predicted onsets are compared with the real onsets to measure the TPR of onsets, this is, the number of onsets that are correctly identified with a TETR as suggested on [2] for 3 and 4 seconds, then divided by the total number of onsets. Figs. 2 and 3 show the results obtained for each subject on both feature sets.

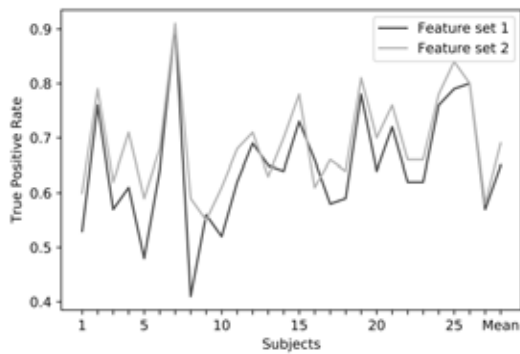


Fig. 2. True Positive rate obtained in both feature sets with a TETR of 3 seconds

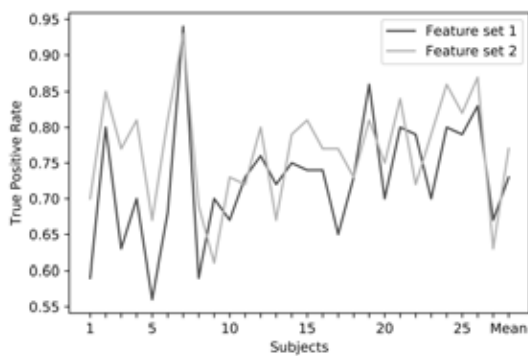


Fig. 3. True Positive rate obtained in both feature sets with a TETR of 4 seconds

IV. DISCUSSION

From figs. 2 and 3 we can see that in the TETR of 3 seconds, the average TPR on detecting the onsets was 0.65 for the first feature extraction method and 0.69 for the second one. In the TETR of 4 seconds, the average TPR on detecting the onsets was 0.73 for the first feature extraction method and 0.77 for the second one. Between the two feature extraction methods the highest TPR is obtained by using the second one based only in generalized Hurst exponent, which suggests that the use of this features to identify the onsets of imagined words on continuous signals is more accurate. The TETR is centered in the real onset which shows us that the error of detecting the onset is in the worst case, half of the TETR (1.5, 2 seconds respectively). The average span of an imagined word in the corpus is 256 samples (2 seconds). Thus this result looks promising for the future research work. Nevertheless, there is still needed to implement another measure that penalizes the false positives detected by the model. In [2] the True False Positive Rate is introduced which could give a better visualization of this results, thus this new measure will be implemented as future work.

V. CONCLUSION

In this paper, two feature extraction methods were calculated to test the possibility of detecting the onset of imagined words in EEG signals using a dataset of 27 subjects that imagined 5 different words. The results showed that in the first feature extraction method based on statistical values, the TPR obtained was 0.65 and 0.73 for TETRs of 3 and 4 seconds respectively. On the second feature set based only in generalized Hurst exponent, the TPR on detecting the onsets of imagined words was 0.69 and 0.77 for TETRs of 3 and 4 seconds respectively.

REFERENCES

- [1] Y. Song and F. Sepulveda, "An online self-paced brain-computer interface onset detection based on sound-production imagery applied to real-life scenarios", *2017 5th International Winter Conference on Brain-Computer Interface (BCI)*, pages 46–49, Jan 2017.
- [2] Y. Song and F. Sepulveda. "A novel onset detection technique for brain-computer interfaces using sound-production relate cognitive tasks in simulated-online system", *Journal of Neural Engineering*, vol 14, no. 1, 2017.
- [3] Alejandro A. Torres-García, Carlos A. Reyes-García, Luis Villaseñor-Pineda, and Gregorio García-Aguilar. "Implementing a fuzzy inference system in a multi-objective eeg channel selection model for imagined speech classification", *Expert Systems with Applications*, vol. 59, pp. 1–12, 2016.
- [4] T. Di Matteo, T. Aste, and M.M. Dacorogna, "Scaling behaviors in differently developed markets", *Physica A: Statistical Mechanics and its Applications*, vol. 324, pp. 183–188, 2003.
- [5] T. Di Matteo, "Multi-scaling in finance", *Quantitative Finance*, vol. 7 no. 1, pp. 21–36, 2007.
- [6] T. Di Matteo, T. Aste, and M.M. Dacorogna, "Long-term memories of developed and emerging markets: Using the scaling analysis to characterize their stage of development", *Journal of Banking and Finance*, vol. 29, no. 4, pp. 827–851, 2005.
- [7] Claudia Lizbeth Martínez-González, Alexander Balankin, Tessy López, Joaquín Manjarrez-Marmolejo, and Efraín José Martínez-Ortiz, "Evaluation of dynamic scaling of growing interfaces in eeg fluctuations of seizures in animal model of temporal lobe epilepsy", *Computers in Biology and Medicine*, vol. 88, no. 1, pp. 41–49, 2017.

DISCRIMINATION BETWEEN CHILDREN AND ADULT FACES USING BODY AND HEAD RATIO AND GEOMETRIC FEATURES

C.A. Reyes-García¹, E. Morales-Vargas¹, H. Peregrina-Barreto¹, C. Manfredi²

¹División de ciencias computacionales, Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, México

²Department of Information Engineering, Università degli Studi di Firenze, Firenze, Italy
{emoralesv, kargaxxi, hperegrina} @ccc.inaoep.mx, claudia.manfredi@unifi.it

Abstract: The classification between a child and an adult is an important task in the medical and security field. It consists in the discrimination between children and adult faces in image sequences. Image processing techniques were used to identify regions of interest and computational intelligence was used for classification. First, in the input image, a body detector was used to measure the size of the human body in the image presented. Later, a face detector projects a bounding box around the face. Then, a facial model is fitted in the face of the person to improve the measure of the heads size. Finally, a normalization step is used to ensure size invariance in the image. Using the body size, head size and facial landmarks, geometric features were extracted to perform the classification. Experimental results show accuracy rates of 86%.

Keywords: Classification, image processing

I. INTRODUCTION

Object detection is a widely used technique of image processing, it consists in defining the boundaries of certain objects with a semantic meaning (a car in a street or a tumor in a medical image for example). Basically, the main purpose of object detection is to put a marker or bounding box in the location of a specific object when it is detected. It's suggested that the object recognition be made in real time to expand the applications of the methods. Then, to detect an object in an image, the features used to feed a classifier need to be robust and easy to calculate [1]. In Computer Science, discrimination between children and adult faces in images is a problem not commonly tackled, it consists in the identification and boundary defining of children and adult faces using computational algorithms. Classification between children and adults is useful for safety, medical or surveillance purposes [2] and is often performed in a very controlled environment (insignificant camera rotations or illumination changes, absence of other objects in the background or absence of facial occlusions) [3]. On the other hand, the problem being a variant of pedestrian detection, initial steps to tackle this problem are performed using algorithms for human body detection in images, which can be useful for surveillance purposes. Although it's a different

problem, some approaches aim to detect children from adults using age classification problems which is probably not the best suitable option using closed circuit television cameras [4]. Although, some works obtained good results in closed environments the problem here relies in the environment where the human face to be discriminated is at, not in the whole image, to allow the algorithms to extract facial features as wrinkles in an easy way. To attack the problem in this kind of environments where no subtle facial features can be extracted, we proposed a methodology for differentiating between children and adults in outdoor scenes.

The proposed method aims to extract features based on body and head sizes avoiding subtle features, such as the relation between the body and face or geometric features, to perform the differentiation between children and adult faces. Also, the effect of fitting a facial model to the initial head recognition was studied to improve the measures and their impact in the recognition rates. The results suggest that it's possible to improve actual methods based on body and head ratio with acceptable accuracy rates when the human body and face are present in an image.

II. METHODS

A methodology for children and adult face discrimination that is divided into three steps was proposed: (A) identification of the region of interest where the human body and their related face are detected. Next (B) the feature extraction step is performed, and finally (C) a classification model is applied. The method initiates in the identification of the regions of interest, which are the face and body of a human. For this purpose, a detector using Aggregate Channel Features [5], [6] trained with the INRIA person dataset was used [6]. For face recognition Viola-Jones algorithm was used [7]. Later geometrical features of detected face and body were extracted, the geometrical features are representing the distance between facial region of interest. Geometrical features were extracted obtaining the distance between the region of interest and normalizing them using the distance between the eyes.

A normalization step was studied to know its effects in the classification rates. Finally, an SVM was used to perform the classification step. Fig. 1 depicts the proposed methodology for children and adult faces discrimination.

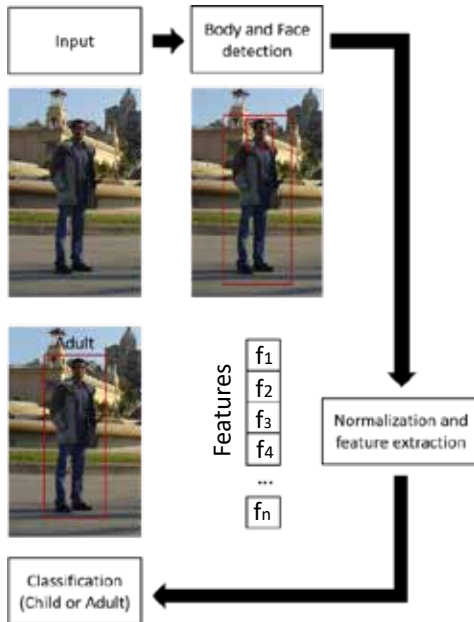


Fig. 1. Proposed methodology for children and adult discrimination using image processing and computational intelligence.

A. Body and head detection

The proposed method focuses on discrimination between children and adult faces. Some studies show that not only the proportion of the face can provide enough information for this purpose. The proportion of the face can provide enough information to discriminate between children and adults because the human head reaches its full size until a person becomes a teenager. For this reason, it's focused first on the identification of the body in the image because it occupies the largest region in the picture. The body identification also serves to reduce the area of search of the face to decrease the number of false positives. In other words, first it's focus on the identification of the body and within the detection area a face that is known to be present is looked for, if there were several detections, we select the best rated or the one that is most likely to be a face.

For the identification of human bodies in an image we use a detector trained using the INRA database with Aggregate Chanel Features. The detection algorithm provides the area where a human body is in the image, using this data, the Viola-Jones algorithm was used to detect all the faces present in the image. Viola-Jones

extracts Haar-like characteristics to identify patterns in the image using a set of templates to match them into distinguishable facial features. An example of the templates used to compute Haar-like features are depicted in Fig. 2. Extracted characteristics pass through a series of weak classifiers trained with the Adaboost algorithm. The main characteristic of the decision algorithm is that all the weak classifiers must agree that a face is being detected to further reduce the detection of false positives.

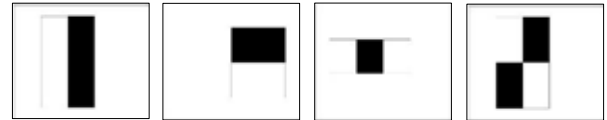


Fig. 2. Templates used to extract the Haar-like features used to detect faces in the image.

B. Feature extraction

Second step to perform the classification of children and adult faces consists in the extraction of features to feed a model of classification. Many features to represent the ratio between the size of the body and the size of the head were explored, the rationale behind this decision is due to the fact that studies remark that head size is bigger at younger ages; extraction of head size as it cannot be a standalone indicator because size in terms of pixels is a relative measure of sizes, it does not take into account the closeness of the object. For this reason, the relation of the body and the head is calculated using Eq 1. where $hsize$ represents the size of the detected bounding head in pixels and $bsize$ is the measured size of the body.

$$r = \frac{hsize}{bsize} \quad (1)$$

After obtaining the bounding box surrounding a face, using the area of where a face is located as an initialization step, a set of 68 facial landmarks were fitted into the face using Intraface models [5], [8]. Fig. 3a depicts an initial facial model before its fitting and 3b show the facial model fitted.



Fig. 3. Matching between a facial model and an image. (a) shows the initialization and (b) the fitted model.

Considering that algorithms for children and adult faces only use the size of the detected bounding box provided by a face detection algorithm to obtain the features and aiming to obtain a better measure, the fitted facial model was used to improve the measure of not only the head but also the facial regions of interest such as: eyebrows, mouth, or nose, allowing the extraction of distance between the different facial regions of interest to improve classification rates. Then, in addition to the ration of head and body, 31 geometric features describing sizes of facial distinctive areas were extracted. The features were obtained computing either the distance or the angle between two points of the obtained facial landmarks. All the features were normalized respective to the distance obtained with points 40 and 43 of Fig. 4. Fig. 4 shows an enumeration of the 68 facial landmark points and Table 1 displays the coordinate pairs to compute the distance or angle.

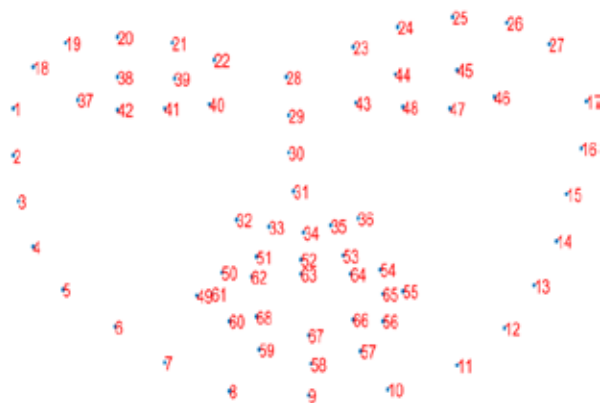


Fig. 4. Enumeration of the facial landmarks obtained using Intraface [5].

Table 1. Coordinate pairs to features computing.

Description	Pairs
Eyebrow (angles)	{20,28,42}, {25,28,47}
Eyebrow stretching	{19,37}, {26,46}
Eyes stretching	{37,38}, {38,39}, {39,40} {40,41}, {41,42}, {42,47} {43,44}, {44,45}, {45,46} {46,47}, {47,48}, {48,43}
Eyes opening	{38,42}, {39,41}, {44,48} {45,47}
Mouth stretching	{49,52}, {52,55}, {55,58}, {58,49}
Mouth openness	{63,67}
Facial sizes	{40,9}, {43,9}, {43,34}, {40,34}, {58,9}, {34,52}

C. Classification

The final stage of the methodology consists in the classification model, for this purpose, an SVM classifier with a linear kernel was used. For validation purposes a k-fold cross validation with 10 folds was used.

III. EXPERIMENTS AND RESULTS

For training and validation purposes, images taken from the INRIA database were used [7]. The INRIA database contains 1,805 64x128 images of humans cropped from a large set of photos. People in the images are usually standing but in any orientation. From the INRIA database a set of 145 images were taken. Images where a human or a child is standing with their face are visible were taken. 74 images of adults and 71 images of children were used for training and validation of the model.

Here the question of how the results of differentiating between children and adult faces can be improved is addressed. Two experiments were performed to analyze if the results can be improved: first, it was explore if adding geometrical features to the model could lead to a better representation and second, it was explore if fitting facial landmarks could improve the measure of head size, leading to better discrimination (BLHR). Figure 5 depicts how facial landmarks can lead to a better understanding of facial sizes and relations rather than the Viola-Jones algorithm output.



Fig.5. Facial landmarks fitted into an image using Intraface [8]. The initial adjustment was made using the located face with the Viola-Jones algorithm [7].

We performed experiments with (BNLHR) and without facial landmarks alignment aiming to mitigate the rotation and occlusions that can cause head sizes to be measured incorrectly. The normalization method consists in a heuristic based on affine transformations, first, a rotation aligns the face using the canthus of the eyes and the facial landmarks are normalized in the range [0,1] [9]. Fig. 6 depicts the normalization process. Finally, experiments adding the geometric features were performed (+GF). Results are shown in Table 2.

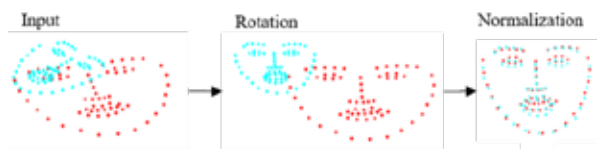


Fig. 6. Graphical representation of the normalization of the facial landmarks to reduce rotation and size noise.

Table 2. Results obtained with the proposed methodology.

Method	Accuracy
BHR [10]	61%
BHR+GF	78%
BLHR	64%
BLHR+GF	86.4%
BNLHR	74.6%
BNLHR+GF	84.7%

IV. DISCUSSION

Although there are not many works tackling the problem of the discrimination between children and adult faces, current solutions are focused in controlled environments which decreases the number of applications to use this kind of models. The proposed model archived a higher accuracy with the reference models [2], [4] fitting a set of 68 facial landmarks to the output of the recognition model to improving the measurement of the face and the classification rates. On the other hand, a limitation of the model is that the body and the face must be present in the image, a possible way to tackle this limitation may be to use an algorithm to recognize heads instead of a face recognition algorithm.

V. CONCLUSIONS

In this work a methodology for the classification of children and adult faces based on facial geometric features was presented. The proposed model not only use the relation between the head and body, but it also uses geometric features describing the normalized sizes of different facial regions of interest. For this approach the obtained accuracy rates increased from 61% to 78% which is seen as an improvement. Trying to alleviate the noise introduced by image size, a normalization was included to the methodology but without getting a significant improvement including or excluding the geometrical features. Improving the measure of the head was using a model fitted that uses the output of a facial recognition algorithm got the best improvement in terms of accuracy of our model and with the addition of the geometrical features better results were obtained.

In conclusion, the results suggest that it's possible to improve discrimination between children and adult faces measuring the head by using facial landmarks instead of the bounding boxes provided by a facial recognition algorithm.

VI. ACKNOWLEDGMENTS

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 AMEXCID, supported by México and Italy governments.

REFERENCES

- [1] K. Lee, C. Lee, S. A. Kim, and Y. H. Kim, "Fast object detection based on color histograms and local binary patterns," IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON, pp. 1–4, 2012.
- [2] M. B. Hans Weda, "AUTOMATIC CHILDREN DETECTION IN DIGITAL IMAGES," pp. 1687–1690, 2007.
- [3] W.-B. Horng, C.-P. Lee, and C.-W. Chen, "Classification of Age Groups Based on Facial Features," Tamkang J. Sci. Eng., vol. 4, no. 3, pp. 183–191, 2001.
- [4] O. F. Ince, I. F. Ince, J. S. Park, J. K. Song, and B. W. Yoon, "Child and adult classification using biometric features based on video analytics," ICIC Express Lett. Part B Appl., vol. 8, no. 5, pp. 819–825, 2017.
- [5] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Aggregate channel features for multi-view face detection," IJCB 2014 - 2014 IEEE/IAPR Int. Jt. Conf. Biometrics, 2014.
- [6] T. Surasak, I. Takahiro, C. H. Cheng, C. E. Wang, and P. Y. Sheng, "Histogram of oriented gradients for human detection in video," Proc. 2018 5th Int. Conf. Bus. Ind. Res. Smart Technol. Next Gener. Information, Eng. Bus. Soc. Sci. ICBIR 2018, pp. 172–176, 2018.
- [7] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," Surf. Sci., vol. 394, no. 1–3, 1997.
- [8] F. De La Torre, W. S. Chu, X. Xiong, F. Vicente, X. Ding, and J. Cohn, "IntraFace," 2015 11th IEEE Int. Conf. Work. Autom. Face Gesture Recognition, FG 2015, no. July, 2015.
- [9] O.-E. F. Morales-Vargas E., Reyes-García C.A., Peregrina-Barreto H., "Facial expression recognition with fuzzy explainable models," Model. Anal. Vocal Emiss. Biomed. Appl., 2017.
- [10] O. F. Ince, J. S. Park, J. Song, and B. W. Yoon, "Child and Adult Classification Using Ratio of Head and Body Heights in Images," Int. J. Comput. Commun. Eng., vol. 3, no. 2, pp. 120–122, 2014.

SONIFICATION TECHNIQUES APPLIED TO EEG SIGNALS OF NONMOTOR GENERALIZED ONSET EPILEPTIC SEIZURES

L. Frassinetti¹, R. Guerrini^{2,3}, C. Barba², F. Melani², F. Piras², C. Manfredi¹

¹ Department of Information Engineering, Università degli Studi di Firenze, Firenze, Italy

² Paediatric Neurology Unit, Neuroscience Department, Children's Hospital A. Meyer, Università degli Studi di Firenze, Firenze, Italy

³ IRCCS Stella Maris Foundation, Pisa, Italy
lorenzo.frassinetti42@gmail.com

Abstract: The increased use of Electroencephalography (EEG) for the diagnosis of epileptic disorders in neuropediatrics has led to the development of technologies for the automatic reduction of redundant information. In modern EEG, pre-processing methods of the signal for artefacts removal and artificial intelligence algorithms for the recognition of critical events have been implemented.

This work analyzes the application of sonification techniques to EEG signals for an almost real time assessment of epileptic seizures onset. Sonification concerns the transformation of numerical data into acoustic signals. EEG sonification could speed up its interpretation and reduce the amount of visual information to be analyzed. The dataset consists of 24 EEG recordings coming from 22 children with absence seizures, clinically evaluated at the Meyer Children's Hospital in Firenze, Italy. The obtained low false positive rate (<1%), shows the usefulness of EEG sonification techniques in supporting automatic and early seizure detection.

Keywords: sonification, EEG, epilepsy, patient-specific, entropy.

I. INTRODUCTION

Epilepsy is one of the main brain disorders, affecting about 1% of the world population [1]. It occurs mainly in childhood and increases, in terms of cumulative prevalence, in the elderly population [2]. Epilepsy, an enduring predisposition to manifest epileptic seizures, is also defined as "a transient occurrence of signs and/or symptoms due to abnormal excessive and synchronous neuronal activity in the brain." [3], [4].

Given their high variability both in terms of aetiology, clinical signs, and electrophysiological characteristics, epileptic seizures are divided into several categories [5].

Electroencephalography (EEG) is one of the main technology supporting the diagnosis of epilepsy.

The work presented here focuses on absence seizures, a type of generalized non-motor epileptic seizures [6]. The EEG signal of a patient with a typical absence shows generalized spike-wave discharges at 3Hz.

One of the main technological problems in the analysis of EEG signals in epilepsy is the presence of artefacts that, summed up to the data to be analyzed, can make the analysis in real time complex and burdensome.

In recent years there has been a growing interest in techniques allowing a near real-time clinical assessment of seizures (i.e. low latency times) [7, 8, 9], especially sonification. According to [10], sonification is defined as: "the transformation of data relations into perceived relations in an acoustic signal for the purposes of facilitating communication or interpretation".

Sonification can be used for performing a fast detection of clinical events or monitoring clinical information. Recently, Schwarz et al. [11] presented a psychoacoustic sonification technique for real time monitoring of neonatal oxygen saturation levels.

G.I. Mihalas et al. [12] gave several examples of physiological signal sonification, such as ECG, Heart Rate etc., and discussed their biomedical applications.

F.B. Vialatte et al. [13] used sonification techniques to discriminate EEGs of patients with mild cognitive impairment from those of healthy control subject.

Our recent paper [9] focused on techniques applied to the EEG signal for early seizure detection, along with artificial intelligence methods for the automatic recognition of absence seizures.

In the present work, the sonification techniques presented in [9] are analysed, with the purpose of reducing the false positive rates in the sonified EEG signals.

Patient-specific techniques are proposed to remove background noise from the sonified EEG signal. Moreover, methods are suggested for enhancing amplitude and patterns related to epileptic seizures or possible artefacts. Finally, similarly to [11], a statistical evaluation of the sonified EEG signal is presented, taking into account relevant differences between false

positives not discarded by the automatic recognition process and the true epileptic seizures.

II. METHODS

To compare the performance of the proposed methods, the same EEG recordings used as test set in [9] and the same automatic recognition algorithms are considered here. The test set consists of 24 recordings of 22 children (7 males, 15 females, and age range 4 – 18 years) with absence seizures evaluated at the Paediatric Neurology Unit, Children’s Hospital A. Meyer, Firenze, Italy. The duration of the signals is 47 ± 10 min with sampling frequency $f_s=256\text{Hz}$. Written informed consent was obtained from all patients or their guardians.

The proposed methods are implemented under MATLAB software (version 2017b [14]) installed on a Hp Pavillion 15 notebook (OS Windows 10, 64 bit) Intel Core i7-5500U processor, CPU 2.40 GHz, RAM 16 Gb.

In [9] the automatic recognition gave 1% average percentage of false positives (FPR metric, ONLINE method) in the sonified signal. Here possible techniques that could further reduce false positives in the sonified signal are investigated.

The sonification procedure proposed in [9] is based on the following relationship:

$$[ABS - ABS - ABS] \rightarrow [beep_1 beep_2 \rightarrow sound_{abs}] \quad (1)$$

Eq. (1) is described as follows. If the automatic recognition step detects three (or more) consecutive absences (ABS), the first 2 consecutive seconds will produce 2 beeps (*beep1* and *beep2*) as pre-alarms. This sound represents a first discrimination between a possible absence and a false positive.

As described in detail in [9], the following seconds classified as ABS produce the $sound_{abs}$. This is the sonified signal of the cumulated sum of the 6th levels of decomposition of the EEG signals obtained applying the Stationary Wavelet Transform.

False positives are grouped into: background without intercritical activity, intercritical activity or, more generally, elements of electrophysiological interest of duration $<2s$, and artefacts.

Thus, methods that would allow the reduction or the recognition of these elements in the sonified signal are investigated.

As concerns the intercritical activity, in the dataset such segments have a time duration $<3s$. Being very similar to epileptic activity they are not removed.

To remove the background noise, a patient-specific adaptive method is implemented. The method is based on the assumption that, in this case, the signal shows lower energy contribution as compared to an epileptic

absence seizure, mainly as its characteristic frequencies are concerned.

The proposed procedure is as follows. First, N windows ($N = 100$) with a duration of 2s each are selected from the signal. For each window the local maxima (V_{local}) of the rectified signal used for sonification are identified with a minimum distance between peaks given by $0.25 \cdot f_s$ and their average value is calculated. Once all the windows and their average values are obtained (V_{mean_local}), a baseline is created for each patient given by the global mean value (V_{mean_global}).

From the window $N+1$ on, the windows such that $V_{mean_local} \leq V_{mean_global} + K \cdot std(V_{local} [1:N])$ are excluded.

The evaluation of the artefacts is made directly on the sonified signal. Considering that the energy trend of the artefacts should be discontinuous with respect to an absence seizure (always around 3Hz), an amplification term α is introduced in the sonification equation [9]:

$$sound_{abs} = \alpha \sum_{i=1:N} E_{sound_i} \sin(2\pi f_{osct}) \quad (2)$$

where α is the square of the value of the peaks normalized with respect to their average and standard deviation and repeated for the minimum duration assigned to each peak. Quadratic amplification was chosen as a compromise between a linear amplification (difference between amplitude values difficult to detect) and an exponential one (excessive reduction of the sonified signal even with minimal differences in amplitude).

Therefore a sonified signal is obtained with periods of silence or discontinuous sound corresponding to artefacts, whereas for typical absences the amplitude of the signal is approximately constant.

Finally, likewise in [13], statistical tests are carried out to establish whether sonified signals of absence seizures are statistically different from false positives. In this work, the discriminant is the value of the Sample Entropy (SE) [15] in each time window of the sonified signal of length equal to 1s. The Mann Whitney test with significance level 0.05 is applied. For the Sample Entropy test, the embedding dimension m is set equal to 2 and the tolerance value r is set equal to 1.

III. RESULTS

The performance is evaluated with the same metrics as in [9]: Balanced Accuracy (BACC); F1score; Matthews Correlation Coefficients (MCC); False Positive Rate (FPR). The comparison between the methods is presented in Table 1.

TABLE 1 - Comparison of the proposed methods with the ONLINE method in [9]

Method	BACC	F1 _{score}	MCC	FPR
Frassinetti et al. [9]	89.0% ± 6.0%	69.0% ± 15.0%	0.70 ± 14.0	1.1% ± 1.0%
K=1	88.0% ± 6.0%	73.0% ± 14.0%	0.73 ± 13.0	0.8% ± 0.9%
K=2	86.0% ± 7.0%	72.0% ± 14.0%	0.72 ± 13.0	0.6% ± 0.7%
K=3	85.0% ± 8.0%	71.0% ± 13.0%	0.72 ± 13.0	0.5% ± 0.7%

An example of qualitative comparison is shown in Figure 1 where the sonified signal of an epileptic seizure and its EEG are presented. The upper plot shows the EEG, the middle plot shows the result of sonification without the post-processing procedure, the lower plot shows the result obtained after the application of the proposed additional post-processing and amplification. Only one channel is shown.

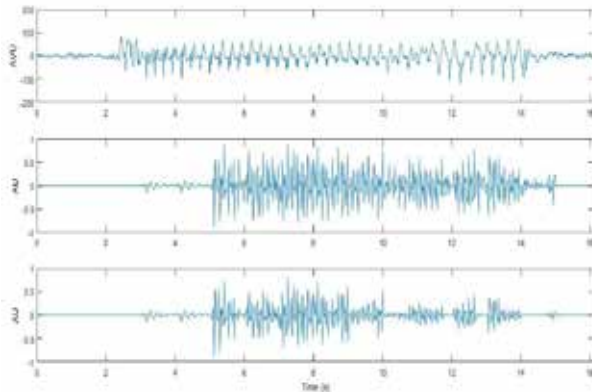


Figure 1 - Example of absence seizure sonification. Above: EEG signal (1 derivation); middle: sonification without post-processing; below: sonification with post-processing

Figure 2 shows an example of artefact and the corresponding sonified signal, before (middle plot) and after (lower plot) the application of the techniques described above.

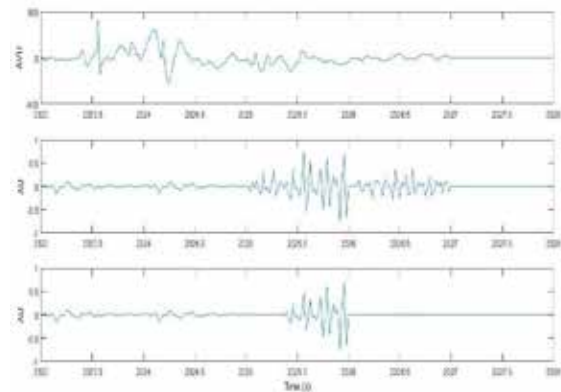


Figure 2 - Example of artefact sonification. Above: EEG signal (1 derivation); middle: sonification without post-processing; below: sonification with post-processing

Finally, the results of the statistical tests described in the previous section are presented in Table 2.

TABLE 2 - Statistical test results. Sample Entropy SE(m=2, r=1) case K=2.

Method	SE(2,1) – True Positive	SE(2,1) – False positive	p-value Mann Whitney
Frassinetti et al. [9]	0.53 ± 0.14	0.51 ± 0.15	0.04
With α	0.35 ± 0.17	0.26 ± 0.16	$3e^{-10}$
With α and post-processing	0.34 ± 0.18	0.14 ± 0.19	$1e^{-25}$

IV. DISCUSSION

Table 1 shows an increase of the performance especially for the FPR parameter that is reduced by almost 50% as compared to the case without post-processing. Also the F1_{score} and MCC parameters increase with respect to the original ONLINE case. The decrease of the BACC parameter is an indication that some seizure windows (True Positive TP) were eliminated by the proposed method. Therefore, the choice of the parameter K must be made with caution to avoid losing useful information: indeed high values, such as $K > 3$, could eliminate more seizures than artefacts. This point is under study.

Regarding the amplification parameter α of the sonified signals, with and without post-processing, the results of Table 2 (case $K = 2$) show significant differences between TP and FP (False Positive) rates.

Therefore, both the proposed techniques could alter the signal complexity, the TP Sample Entropy values decreasing more than the original ones, but less than the FP SE values. Possible reasons for this behaviour will be investigated in future developments of the methods.

Figures 1 and 2 and the statistical results of Table 2 show the possibility of detecting different events in the sonified signal, at least in terms of sound intensity and its repeatability. However, given the large variety of possible artefacts in the EEG signal, the proposed method cannot capture all possible variants. In particular, those that have a rhythmic energetic trend with frequencies close to 3Hz would be difficult to differentiate from the typical absence seizures.

Given the low number of atypical epileptic absence seizures in the dataset (2 recordings), no statistical tests were carried out to assess differences with typical absences in the sonified signal.

V. CONCLUSION

The application of post-processing and signal amplification techniques can lead to improvements both in terms of performance and qualitative recognition of events.

For a seizure detection system that applies sonification techniques the reduction of the FPR parameter seems relevant to reduce the amount of information to be analyzed.

The obtained statistical results allow performing validation experiments on the sonified signal [16]. The statistical evaluation of the sonified signal of atypical absences with respect to the typical ones will be one of the main future developments.

In conclusion, sonification techniques applied to EEG signals could introduce advantages for the real-time interpretation of information related to epileptic absence seizures. However, setting up sonification techniques for an easy removal of noise must be further exploited. If successful, they could help in reducing subjectivity and speed up the analysis.

ACKNOWLEDGEMENT

This work was partially funded by Fondazione Cassa di Risparmio di Firenze, projects: 2016.0952 and 2015.0911.

REFERENCES

- [1] *Epilepsy Atlas 2005 World Health Organization*.
- [2] E. Cesnik et al. "Incidence of epilepsy in Ferrara, Italy" *Springer, Neurol Sci (2013)* 34:2167-2172 doi 10.1007/s10072-013-1442-5.
- [3] <https://www.ilae.org/> last access: 10/09/2019
- [4] R.S. Fisher et al. "Epileptic seizure and epilepsy: definitions proposed by the International League against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE)" *Epilepsia* 2005;46:470-472.
- [5] R.S. Fischer et al. "Operational classification of seizure types by the International League Against Epilepsy: Position Paper of the ILAE Commission for Classification and Terminology." *Epilepsia*. 2017 Apr;58(4):522-530. doi: 10.1111/epi.13670.
- [6] C. E. Stafstrom et al. "Seizures and Epilepsy: An Overview for Neuroscientists" *Cold Spring Harb Perspect Med*. 2015 Jun 1;5(6) doi: 10.1101/cshperspect.a022426.
- [7] M. E. Saab and J. Gotman, "A system to detect the onset of epileptic seizures in scalp EEG," *Clin. Neurophysiol.*, vol. 116, no. 2. pp. 427–442, Feb. 2005.
- [8] Vidyaratne, Lasitha S., Iftekharuddin, Khan M., 2017. "Real-time epileptic seizure detection using EEG". *IEEE Trans. Neural Syst. Rehabil. Eng.* 25 (11), 2146–2156.
- [9] Frassinetti et al. "Automatic detection and sonification of nonmotor generalized onset epileptic seizures: Preliminary results" *Brain Research* 1721 (2019) 146341.
- [10] Kramer, G., et al., 1999. "Sonification report: status of the field and research agenda." Tech. Rep., Int. Commun. Auditory Display
- [11] S. Schwarz et al. "A psychoacoustic sound design for pulse oximetry" 25th International Conference on Auditory Display (ICAD 2019) 23-27 June 2019 <https://doi.org/10.21785/icad2019.024>
- [12] G.I. Mihalas et al. "Can Sonification Become a Useful Tool for Medical Data Representation?" MEDINFO 2017: Precision Healthcare through Informatics A.V. Gundlapalli et al. (Eds.) 2017 International Medical Informatics Association (IMIA) and IOS Press. doi:10.3233/978-1-61499-830-3-526
- [13] F.B. Vialatte et al. "Audio representations of multi-channel EEG: a new tool for diagnosis of brain disorders" *Am J Neurodegener Dis* 2012; 1(3):292-304.
- [14] MATLAB and Statistics and Machine Learning Toolbox Release 2017b. The MathWorks, Inc., Natick, Massachusetts, United States.
- [15] Richaman JS, Moorman JR. "Physiological timeseries analysis using approximate entropy and sample entropy." *Am J Physiol Heart Circ Physiol* 2000; 278: 2039-2049.
- [16] Bonebright, T.L., 2011. "Evaluation of Auditory Display". In: Hermann, Thomas, Hunt, Andy, Neuhoff, John G. (Eds.), *The Sonification Handbook*. Logos Verlag, Berlin, Germany, pp. 111–141.

BIOVOICE: A MULTIPURPOSE TOOL FOR VOICE ANALYSIS

Maria Sole Morelli¹, Silvia Orlandi² Claudia Manfredi¹

¹ Department of Information Engineering (DINFO), Università degli Studi di Firenze, Firenze, Italy

² Bloorview Research Institute, Holland Bloorview Kids Rehabilitation Hospital, Toronto, ON, Canada

msole.morelli@gmail.com, sorlandi@hollandbloorview.ca, claudia.manfredi@unifi.it

Abstract: BioVoice is a user-friendly software for the acoustical analysis of the human voice. Here we introduce the last version of this software tool.

Results of the acoustical analysis of newborn cry signals are described and discussed. Furthermore, results of the acoustical analysis of 2 adults (1 male and 1 female) and 1 child emitting sustained vowels are shown.

Keywords: BioVoice, acoustic analysis, software tool, voice parameters, infant cry, fundamental frequency, formant frequencies.

I. INTRODUCTION

Speech and vocal productions are characterized by acoustic waves generated by physiological processes that involve the central nervous system, the respiratory system and the phonatory apparatus.

The acoustical analysis of the human voice is considered clinically relevant in assessing the health state of the phonatory apparatus and detecting possible neurological disorders [1, 2]. Main clinical applications concern screening, diagnostic support and evaluation of the effectiveness of treatments [3]. To date, few automated methods have been developed for voice analysis, the most used being PRAAT [4]. Another one is the LENA system, that investigates both healthy child recordings and child populations with language disorders [5].

BioVoice, a user-friendly software tool for voice analysis, was developed at the Biomedical Engineering Lab, Firenze University [6–9]. It allows recording the human voice from the newborn to the elder and performing both time and frequency analysis, estimating more than 20 acoustical parameters. Here, a short description of the software is proposed. Furthermore, some results on voice analysis of newborn cries, child sustained vowels and adult sustained vowels are shown.

II. METHODS

A. Biovoice: software description

In Fig. 1, the main interface of BioVoice is shown. The user has to follow few mandatory steps to perform voice analysis.

First, the user selects and uploads at least one audio file. Indeed, it is possible to upload at the same time several files of any time duration and concerning different age range, gender and kind of voice emission. The file selection is allowed from any folder on the computer or HD or USB key. Only wav. files can be analysed.

Before starting the analysis, the user has to specify the settings of the audio file(s). Specifically, age (newborn, child, adult), gender (male or female), and kind of vocal emission (voiced, singing, speech, cry) must be selected (Fig.2).

When the analysis starts, BioVoice first performs the selection of voiced/unvoiced (V/UV) audio segments [10]. Then, all the parameters of interests are extracted from each voiced segment. Specifically, in time domain, information about the number and length of voiced segments, length of silence segments and percentage of voiced segments are extracted and saved in an excel table. A picture shows the V/UV selection. In the frequency domain, fundamental frequency (F0), formant frequencies (F1, F2, F3), noise level (Normalized Noise Energy) and jitter are estimated. For F0 and for each formant, the mean, median, standard deviation, maximum and minimum values are calculated and saved in excel tables. Furthermore, differently from other automatic software tools, BioVoice computes the melodic shape of F0, identifying up to 12 melodic shapes (rising, falling, symmetric, plateau, low-up, up-low, double, frequency step, complex, unstructured, not a cry, other) [11]. It is also possible to perform a perceptual melodic analysis, looking at each melodic shape of F0 and classifying them manually. Some other options are available. Indeed, a selected audio file can be listened to, and new audio files can be saved using a connected external microphone. For newborn and child, it is also possible to perform the perceptual analysis of the melodic shape and compare it with the automatic results.



Fig.1 BioVoice user interface

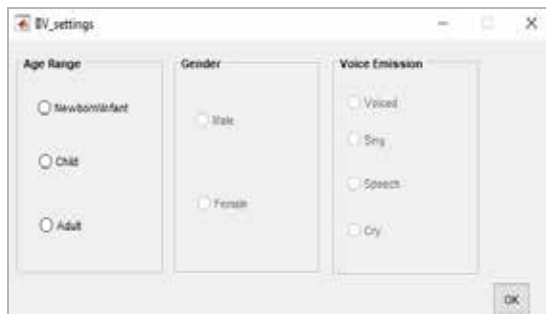


Fig.2. BioVoice settings

At the end of the analysis, BioVoice results and pictures are saved in a specific folder created in the same directory of the audio file. Tables (excel) contain F0 and formants values and statistical information about the parameters. Colour figures (.jpeg) include: V/UV selection, F0 shape and spectrogram with formants values superimposed, for each detected voiced frame.

B. Data analysis

To show some applications of BioVoice, we report the results of the analysis of four different recordings pertaining to different categories of human voice: a newborn cry (1st week of life) and sustained vowels of: a typically developing child (4 years old), an healthy adult male and an adult female (both 24 years old). Specifically, for the child and the adults only \a\, \i\ and \u\ vowels are considered here, as they allow building the vowel triangle, i.e. the plot of F2 vs F1 [12].

III. RESULTS

In Fig.3, an example of the results of newborn cry analysis is shown. Specifically, in this recording six voiced frames (cry units) were detected with BioVoice. For each cry unit, F0, along with its mean and std and

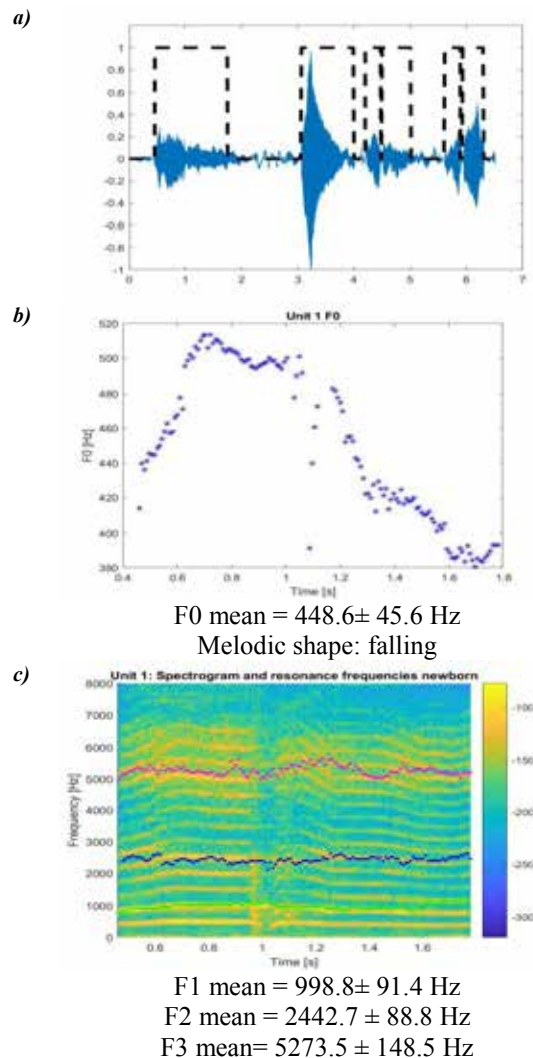


Fig.3. Newborn cry. a) V/UV selection; b) F0 shape; c) Spectrogram and formants. First cry unit.

its melodic shape are computed. Also, the spectrogram

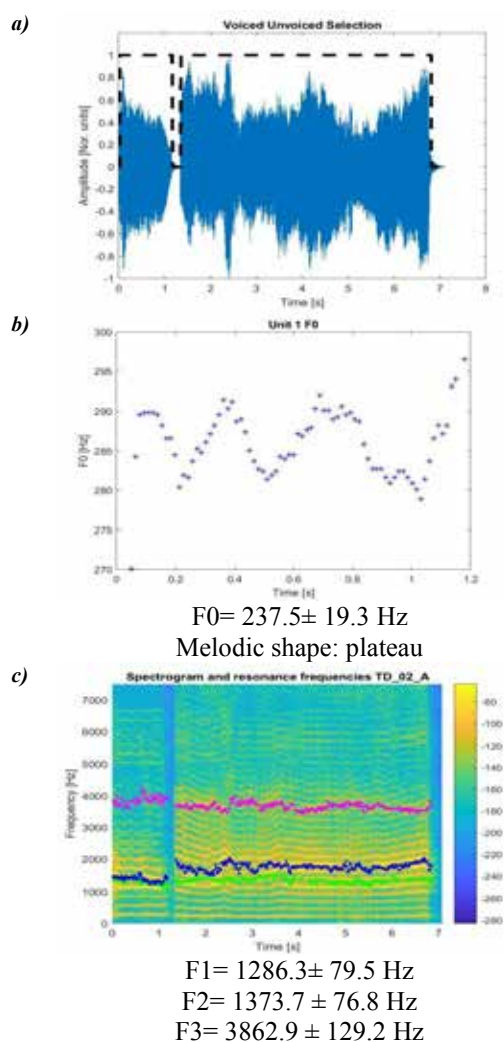


Fig.4. Child voice (sustained 'a')

a) V/UV selection; b) F0 shape; c) Spectrogram and formants for the first voiced unit.

with the first three formants superimposed is reported. In Fig.4, an example of sustained vowel for a child voice is reported (a), as well as its the melodic shape and spectrogram. Finally in Fig.5, an example concerning a male adult voice is shown (a). In Fig. 6, the vowel triangle of a child voice, of a male and a female adult voices are shown and compared to the normative values of male adult voice reported in [12].

IV. DISCUSSION

We presented the last version of BioVoice, a software developed at the Firenze University for the acoustical analysis of human voice. This updated version is improved as of the interface layout is concerned, but most of all concerning the methodologies for parameter estimation. In addition to F0 and formants values, it allows performing melodic

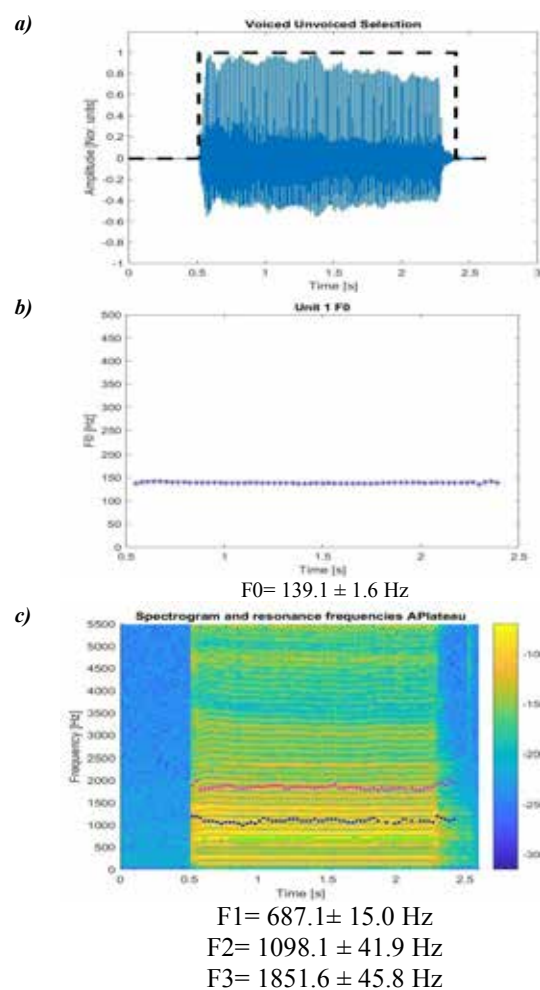


Fig.5. Adult male voice (sustained 'a')

a) V/UV selection; b) F0 shape; c) Spectrogram and formants.

shape analysis in newborn cry and child's voice, both automatic and perceptual.

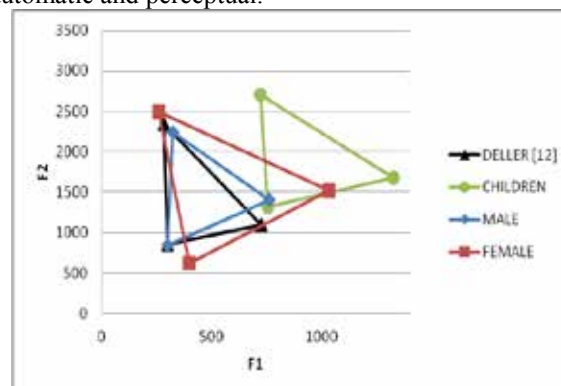


Fig. 6. Vowel triangle: Comparison of children and adults results to normative values in [12].

The reported results show some potential of this software. As for newborn cry analysis, BioVoice provides coherent values of F0 and formants in

different cry units. Concerning the melodic analysis, the software automatically detects the shape of F0, that can be compared to the perceptual one. As for child analysis, the comparison of its vowel triangle with normative values of adult male shows some differences. Indeed, as expected, children voice has higher values of F1 and F2 with respect to adults, the values of formants being related to the shape and size of the vocal tract: the larger the vocal tract cavities, the lower the resonance frequencies [12]. Concerning the adult's voice, the same consideration can be made. The vowel triangle of male voice is similar to the normative triangle reported in [12], while, as expected, for the female voice higher frequency values are shown, due to the smaller vocal tract cavities.

In the current version of BioVoice child and adult voice analysis is limited to voiced frames only. We are working towards including running speech analysis for both groups. Moreover, a revised version of the singing voice analysis is under development. In this case, two more formants, F4 and F5 must be estimated, as well as the so-called singer's formant and other acoustical parameters [13]. Finally, up to now the melodic shape is computed only for newborn cry and children voice, but it will be extended to the adult's voice too, as it might be related to the emotional state of the subject.

V. CONCLUSION

A new version of BioVoice is presented. This software could be very useful for estimating several acoustical parameters, from the newborn to the elder. The software is very intuitive and easy to use also by a less expert user. The executable version is freely available upon request.

REFERENCES

- [1] Niedzielska, G. (2001). Acoustic analysis in the diagnosis of voice disorders in children, *International Journal of Pediatric Otorhinolaryngology*, Vol. 57, No. 3, 189–193. doi:10.1016/S0165-5876(00)00411-0
- [2] Niebudek-Bogusz, E.; Fiszer, M.; Kotylo, P.; Sliwinska-Kowalska, M. (2006). Diagnostic value of voice acoustic analysis in assessment of occupational voice pathologies in teachers, *Logopedics Phoniatrics Vocology*, Vol. 31, No. 3, 100–106. doi:10.1080/14015430500295756
- [3] Laver, J.; Hiller, S.; Beck, J. M. (1992). Acoustic waveform perturbations and voice disorders, *Journal of Voice*, Vol. 6, No. 2, 115–126. doi:10.1016/S0892-1997(05)80125-0
- [4] Boersma, P. (2001). Praat, a system for doing phonetics by computer., *Glott International*, Vol. 5, No. 9/10
- [5] Ganek, H.; Eriks-Brophy, A. (2018). Language ENvironment analysis (LENA) system investigation of day long recordings in children: A literature review, *Journal of Communication Disorders*, Vol. 72, 77–85. doi:10.1016/J.JCOMDIS.2017.12.005
- [6] Orlandi, S.; Bocchi, L.; Donzelli, G.; Manfredi, C. (2012). Central blood oxygen saturation vs crying in preterm newborns, *Biomedical Signal Processing and Control*, Vol. 7, No. 1, 88–92. doi:10.1016/J.BSPC.2011.07.003
- [7] Manfredi, C.; Bocchi, L.; Cantarella, G. (2009). A multipurpose user-friendly tool for voice analysis: Application to pathological adult voices, *Biomedical Signal Processing and Control*, Vol. 4, No. 3, 212–220. doi:10.1016/J.BSPC.2008.11.006
- [8] Rruqja, N.; Dejonckere, P. H.; Cantarella, G.; Schoentgen, J.; Orlandi, S.; Barbagallo, S. D.; Manfredi, C. (2014). Testing software tools with synthesized deviant voices for medicolegal assessment of occupational dysphonia, *Biomedical Signal Processing and Control*, Vol. 13, 71–78. doi:10.1016/J.BSPC.2014.03.011
- [9] Manfredi, C.; Bocchi, L.; Orlandi, S.; Calisti, M.; Spaccaterra, L.; Donzelli, G. P. (2008). Non-invasive distress evaluation in preterm newborn infants., *Conference Proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, Vol. 2008, 2908–2911. doi:10.1109/IEMBS.2008.4649811
- [10] Orlandi, S.; Dejonckere, P. H.; Schoentgen, J.; Lebacqz, J.; Rruqja, N.; Manfredi, C. (2013). Effective pre-processing of long term noisy audio recordings: An aid to clinical monitoring, *Biomedical Signal Processing and Control*, Vol. 8, No. 6, 799–810
- [11] Manfredi, C.; Pieraccini, G.; Orlandi, S.; Viellevoeye, R.; Torres-García, A.; Reyes-García, C. (2019). Automated analysis of newborn cry: relationships between melodic shapes and native language, *Biomedical Signal Processing and Control*, Vol. (In print)
- [12] Deller, J. R.; Hansen, J. H. L.; Proakis, J. G. (2000). *Discrete-Time Processing of Speech Signals*
- [13] Manfredi, C.; Barbagallo, D.; Baracca, G.; Orlandi, S.; Bandini, A.; Dejonckere, P.H. (2015). Automatic Assessment of Acoustic Parameters of the Singing Voice: Application to Professional Western Operatic and Jazz Singers, *Journal of Voice*, Vol. 29, No. 4, 517.e1-517.e9

INDEX OF AUTHORS

- Agostini Tiziano 227
Aichinger Philipp 121, 125, 133, 137, 141
Álvarez A. 25,
Angelakis Evangelos 51
Arora Siddharth 45
- Barba C. 257
Becherini Cosimo 171
Biavati E. 81
Bouvet Anne 149, 153
Bruni Eleonora 81, 181
Bula Viteslav 205
Bulusu Sridhar 141
Burdumy Michael 179
- Calcinoni Orietta 171
Campos-Roca Y. 103
Cantarella Giovanna 167, 185
Chukaeva Tatjana 85
Ciabatta Annaclara 167
Contesse Adrien 219
Coppola Walter 227
- D'Amario Sara 59
Daffern Helena 55
De Arcas-Castro Guillermo 21
Dehak Najim 33
DeJonckere Philippe 129, 163, 167
Devaraj Vinod 133
Di Natale Valentina 167
Dlask Pavel 63
Doroginskaya E.S. 245
Draginskaya E.S.
Drioli Carlo 137
- Echternach Matthias 179
Evain Solene 219
Evdokimova Vera 85, 237
Evgrafova Karina 85
- Franke Ingolf 175
Frassinetti L. 257
Frič Marek 63, 67
Fussi Franco 81, 181
- Gálvez-García Gerardo 21
Ganapathy Sriram 223
Georgaki Anastasia 51
Godino-Llorente Juan Ignacio 33
Gómez A. 21, 25
- Gómez P. 25
Gómez-García Jorge Andrés 33
Gómez-Rodellar Andrés 21
Gómez-Vilda Pedro 21
Gonet Wiktor 233
Greco Alberto 93
Grenez F. 37
Guerrini R. 257
Guhschina Y.S. 245
Gully Amelia 55
- Hampala Vít 209
Hemmerling Daria 29
Henrich Bernardoni Nathalie 219
Hermansky Hynek 223
Hernández-del-Toro C. Tonatiuh 249
Horáček Jaromír 205, 213
Hortis-Dzierzbicka Maria 233
Howard David M. 59
Hruška Viktor 63
- Israelsen Niels Møller 195
- Jønsson Anders Overgård 195
- Kacha A. 37
Kamanina N.A. 245
Kastyro I.V. 241, 245
Kim Y-J. 71, 89
Kovalenko A.N. 241, 245
Krenn Matthias 189
Kumar S. Pravin 141
- Laukkanen Anne-Maria 205, 213
Lebacqz J. 129
Lecouteux Benjamin 219
Lehoux Hugo 209
Leonhard Matthias 189
Lucero Jorge C. 145
- Madruga M. 103
Manfredi Claudia XIII, 167, 253, 257, 261
Marceglia Sara 227
Martínez R. 25
Marzi Claudia 93
Mayr Winfried 189
Melani F. 257
Morales-Vargas E. 253
Morelli Maria Sole 261
Moro-Velázquez Laureano 33

Motlicek Petr 223

Orlandi Silvia 261

Palacios D. 25

Palacios-Alonso Daniel 21

Pavani F. 97

Pedersen Mette 195

Pelorsen Xavier 145, 149, 153

Peregrina-Barreto H. 253

Pérez C.J. 103

Pinchaud Antoine 219

Piras F. 257

Podzimková Iva 67

Prenassi Marco 227

Pützer Manfred 111

Patel Rita 107

Radolf Vojtěch 205, 213

Ramponi Giovanni 227

Reyes-García Carlos A. 249, 253

Richter Bernhard 179

Rosa E. 81

Schlömicher-Thier Josef 175

Schneider-Stickler Berit 189

Schoentgen J. 37

Schwab Didier 219

Scilingo Enzo Pasquale 93

Sidtis John J. 41, 71, 89

Skopich A.

Skrelin Pavel 85

Spahn Claudia 179

Stoffels H. 163

Sundberg Johan 157

Svec Jan G. 141, 209

Sztaho David 29

Ternström Sten 107

Tokuda Isao 149

Torshin V.I. 241, 245

Traser Louisa 179

Tropper Hannes 175

Tsanas Athanasios 25, 45, 201

Van Hirtum Annemie 149, 153

Van Lancker Sidtis Diana 17, 41, 71, 89

Vanello Nicola 93

Vurma Allan 75

Wokurek Wolfgang 111,

Xi Chen 17

Yang Seung-yun 115

Zakharchenko E.A. 237

Zeloni G. 97

Zomkowska Edyta 233

