UNIVERSITÀ
DEGLI STUDI
FIRENZE
DINFO
DIPARTIMENTO DI
INGEGNERIA
DELL'INFORMAZIONE

12th
INTERNATIONAL
WORKSHOP

MODELS AND ANALYSIS OF VOCAL EMISSIONS FOR BIOMEDICAL APPLICATIONS

December 14-16, 2021
Firenze, Italy

PROCEEDINGS

FIRENZE
UNIVERSITY
PRESS

PROCEEDINGS E REPORT

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

131

MODELS AND ANALYSIS OF VOCAL EMISSIONS FOR BIOMEDICAL APPLICATIONS

12TH INTERNATIONAL WORKSHOP

December 14-16, 2021
Firenze, Italy

Edited by
Claudia Manfredi

Firenze University Press
2021

Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021 / edited by Claudia Manfredi. – Firenze : Firenze University Press, 2021.
(Proceedings e report ; 131)

https://www.fupress.com/isbn/9788855184496

ISSN 2704-601X (print)
ISSN 2704-5846 (online)
ISBN 978-88-5518-448-9 (Print)
ISBN 978-88-5518-449-6 (PDF)
ISBN 978-88-5518-450-2 (XML)
DOI 10.36253/978-88-5518-449-6

Cover: designed by CdC, Firenze, Italy.

FUP Best Practice in Scholarly Publishing (DOI https://doi.org/10.36253/fup_best_practice)
All publications are submitted to an external refereeing process under the responsibility of the FUP Editorial Board and the Scientific Boards of the series. The works published are evaluated and approved by the Editorial Board of the publishing house, and must be compliant with the Peer review policy, the Open Access, Copyright and Licensing policy and the Publication Ethics and Complaint policy.

The online digital edition is published in Open Access on www.fupress.com.

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 2021

Firenze, Italy

The MAVEBA 2021 Workshop is sponsored by:

Università degli Studi di Firenze
Department of Information Engineering - DINFO

Department of Information Engineering
www.dinfo.unifi.it

Int. Journal Biomedical Signal Processing and Control Elsevier l.t.d.
www.sciencedirect.com/journal/biomedical-signal-processing-and-control

CoMeT - Collegium Medicorum Theatri
comet-collegium.com

World Voice Day
world-voice-day.org

and is supported by:

Italian Ministry of Foreign Affairs and International Cooperation - MAECI
www.esteri.it

Department of Information Engineering
www.dinfo.unifi.it

CONTENTS

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

MAVEBA
2021
Firenze, Italy

FOREWORD

This book of Proceedings includes the contributions presented at the 12th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications – MAVEBA 2021, held in Firenze from 14 to 16 December, 2021.

The previous edition of MAVEBA in December 2019 was an opportunity to happily celebrate the twentieth anniversary of MAVEBA with "historical" colleagues and many new ones. No one would have imagined that just a couple of months later everyone's life would have changed dramatically due to the COVID-19 pandemic!

This year's edition wants to be a sort of revenge for life and a wish to start again meeting each other for everyone. For this reason it is mostly carried out not in a virtual way but in person.

I thank all the participants who, with their presence, wanted to be next to me once again. I also thank those who could not participate directly for personal or security reasons.

Thus, the question that I asked me two years ago is still the same, and I give the same answer: MAVEBA started because of curiosity and continued thanks to the enthusiasm of the participants: and today? Curiosity and enthusiasm are still there despite pandemic, with the awareness of a fascinating and increasingly interdisciplinary world.

The main subjects of MAVEBA 2021 still concern methods for analyzing the human voice and retrieving its features related to particular physiological or neurological conditions. The aim is that of assessing reliable procedures for objective and quantitative definition of levels of voice disorders, singing voice parameters, newborn cry features, vocal fold and vocal tract modelling and mechanics. The interdisciplinarity, that has always characterized the MAVEBA workshops, is well highlighted by the themes addressed, that are listed below.

The papers presented at MAVEBA 2021 and collected in this volume are divided into eight Sessions.

SESSION I – MODELS AND ANALYSIS
SESSION II – SPEECH
SESSION IIII – NEUROLOGICAL DISORDERS
SESSION IV - BIOMECHANICS
SESSION V - SINGING
SESSION VI – NEWBORNS AND CHILDREN
SESSION VII - PARKINSON
SESSION VIII – COVID-19

I am very grateful to the authors for their contribution and to all participants, near and far, but present anyway, that stimulated the discussion and helped to propose new research themes and methodologies of analysis in the continuously evolving field of the study of the human voice.

Claudia Manfredi
MAVEBA 2021 Chair

ACKNOWLEDGEMENTS

I greatly acknowledge my PhD student Dr.Eng. Lorenzo Frassineti, who manages and constantly updates the website, collaborated in reviewing the Proceedings and in solving the daily difficulties with patience and professionalism, Dr. Eng. Benedetta Olmi, research fellow, the biomedical Engineering students Federico Calà, Angela Parente, Valentina Guarguagli and the ScaramuzziTeam Congress Agency for its professionalism.

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

SESSION I
MODELS AND ANALYSIS

Objective Detection of Amplitude Modulation in Glottal Area Waveforms

Vinod Devaraj[1,2], Philipp Aichinger[1]

[1]Dept. of Otorhinolaryngology, Division of Phoniatrics-Logopedics, Medical University of Vienna, Austria [2]Department of Signal Processing and Speech Communication, Graz University of Technology, Austria
vinod.devaraj@meduniwien.ac.at, philipp.aichinger@meduniwien.ac.at

Abstract: **Traditionally, glottal area waveform (GAW) of modal voice have negligible amplitude modulation. Contrarily, for other voice qualities, like vocal fry and diplophonic voice qualities in particular, the GAWs contain multiple amplitude modulated pulses in a single modulator cycle. In this proposed approach, amplitude modulated vocal fry GAW segments are objectively detected. First, GAWs are modelled using an analysis-by-synthesis approach. This approach fits two modelled GAWs for each of the input GAW. One modelled GAW is modulated to replicate the amplitude and frequency modulations of the input GAW and the other modelled GAW is unmodulated. Modelling errors are obtained by taking root mean squared difference between the input and two modelled GAWs separately. These two modelling errors are used as features for detection of amplitude modulated segments using a support vector machine (SVM) followed by a hidden Markov model (HMM). The sensitivity, specificity and accuracy of detecting the amplitude modulated GAW segments are 0.79, 0.92 and 0.92 respectively.**

Keywords: voice quality, vocal fry, glottal area waveform

I. INTRODUCTION

Typical voice qualities are modal, breathy, hoarse, diplophonic, and vocal fry. In this study, amplitude modulated vocal fry voice quality is investigated. Vocal fry is mainly characterized by a low fundamental frequency which gives an auditory impression of "a stick being run along a railing", "popping of corn" or "cooking of food on a pan" [1, 2, 3]. Other characteristics include shimmer, jitter, and damping of pulses. Also, sub-glottal air pressure and air flow were found to be smaller in vocal fry than in modal registers [2]. However, fundamental frequency is one of the main factors which distinguishes vocal fry from modal and harsh voice [4]. In our work, we identify vocal fry based on the impulsivity of voice samples, i.e., the auditory attribute associated with the separate perception of glottal cycles.

Traditionally, vocal fry was associated with a single pulse and a long closed phase of the glottis in each modulator cycle. However, several studies have investigated the vibration patterns of vocal fry which have amplitude modulated cycles with multiple pulses. Authors of [5] and [6] found single and double pulsed patterns respectively using high-speed videos where the closed phase is longer than the open phase. More evidence for the existence of multiple pulses in a single cycle was reported in [1, 2 and 3]. Also, multiple pulses in a single cycle without a long closed phase were observed using electroglottograms which reflect translaryngeal electrical resistance that is proportional to the contact area of the vocal folds [7]. In this paper, we observe vocal fry GAWs with multiple single-peak pulses in a single cycle as shown in Fig. 1.

For objective detection of vocal fry, several approaches have been used in the past. Presence of vocal fry segments in speech utterances were detected based on the autocorrelation properties of the audio signals [8]. In [9], audio features like inter-frame periodicity, inter-pulse similarity, peak fall and peak rise, H2-H1, i.e., the difference in amplitudes of the first two harmonics, F0 contours in each frame and peak prominence were used for vocal fry detection. A Fourier spectrum analysis approach of the audio signals was also proposed for distinguishing vocal fry segments from diplophonic voice [10]. Distinct characteristics of fry and modal regions were obtained using so called epoch parameters [11]. Epochs are negative to positive zero crossing instances of zero frequency filtered audio signals. Though these methods detect vocal fry or creaky segments, they do not allow a detailed study of voice production. In this study, we use GAWs which provide in depth investigation of voice production.

II. METHODS

Data used:

In this study, 398 GAWs are used which are extracted from the high-speed videos using a segmentation tool. These GAWs are annotated with regard to the presence or absence of amplitude modulation. Unvoiced GAW segments are also marked

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

as unmodulated. Two tasks are performed in this study: 1) detection of amplitude modulated GAW segments and 2) detection of amplitude modulated GAW segments which are audio annotated as vocal fry. For task 1, the positive data consists of amplitude modulated GAW segments. This constitutes seven percent of the total GAW segments. The negative data consists of unmodulated and unvoiced GAW segments. With these annotations as ground truth, we investigate the efficiency of a detector which detects the presence of amplitude modulation present in the GAWs. The detector will be explained later in this section. The detector trained for task 1 is used for task 2 also. For annotating the GAW segments as fry in task 2, audio signals corresponding to the GAWs are annotated with regard to the presence or absence of vocal fry. The amplitude modulated GAWs segments which overlap with the audio annotations labelled as vocal fry, are used as positive data. The amplitude modulated fry GAW segments constitutes of around one percent of the total GAW segments. All the remaining GAW segments (unvoiced, unmodulated and amplitude modulated segments which have the corresponding audio annotation as non-fry) are used as negative data.

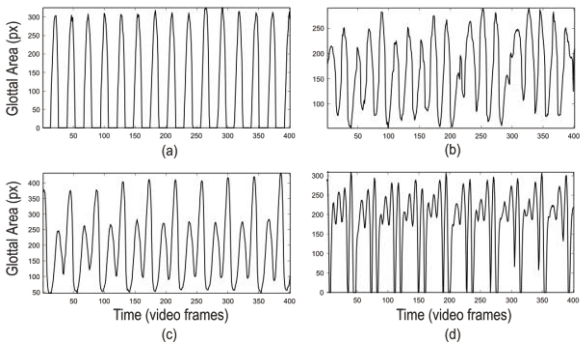


Figure 1: Example GAWs extracted by segmentation of the HSVs:(a) euphonic voice and (b, c, and d) vocal fry. Vocal fry GAWs are amplitude modulated. Euphonic GAW has negligible amplitude modulation.

GAW model:

GAWs are modelled using an analysis-by-synthesis approach [12,13]. Fig. 2 shows the block diagram of the analysis-by-synthesis approach used for modelling each of extracted GAWs. For each input GAW, two modelled GAWs are obtained, where one modelled GAW is modulated and the other modelled GAW is unmodulated. Using this approach, first, the fundamental frequency (f_o) track of each input GAW is estimated using a hidden Markov model (HMM). An unmodulated quasi-unit pulse train u_t is generated by an oscillator driven by the extracted f_o track. The pulse locations of this pulse train approximate the time instants of the maxima of the input GAW. Pulse shapes r_l are obtained by cross-correlating the quasi-unit pulse with each block of the input GAW of length 32 ms obtained using a Hanning window with a 50 percent overlap. These single-peak pulse shapes are then modelled using Chen's model [14]. Fourier coefficients of the pulse shapes are then obtained by transforming the pulse shapes using the discrete Fourier transform (DFT). A Fourier synthesizer (FS) uses the Fourier coefficients R_k and the instantaneous phase Θ(t) extracted from the quasi-unit pulse train to obtain $y_F(\mathrm{t})$. The synthesized GAW $y_F(\mathrm{t})$ is multiplied with an amplitude modulator $m(t)$ to obtain a modelled GAW,

$$\hat{y}(t) = y_F(\mathrm{t}) \cdot m(t). \qquad (1)$$

$m(t)$ is obtained using the pulse heights of the quasi-unit pulse train.

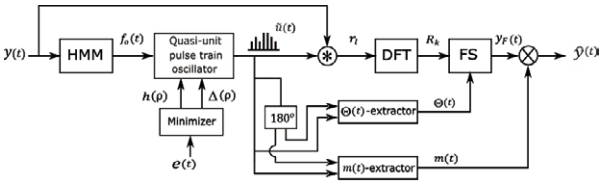


Figure 2: Analyzer used for modelling GAWs. For each input GAW, two modelled GAWs are output by the analyzer, where one modelled GAW is modulated and the other is unmodulated [12].

$\hat{y}(t)$ obtained using the unmodulated quasi-unit pulse train u_t is the output of a non-modulating model where the amplitude and frequency modulations present in the input GAWs are not modelled. To model these random modulations of the input GAW, the quasi-unit pulse train is modulated iteratively on a pulse-to-pulse time scale by minimizing the modelling error between the input and the modelled GAW. $\hat{y}(t)$ obtained using the modulated quasi-unit pulse train $\tilde{u}_t$ is the output of a modulating model. Modelling error E is the root mean squared difference between the input and the modelled GAW.

$$E(t) = 20 \cdot log_{10}\left(\frac{\sqrt{\overline{\left(y(\mathrm{t}) - \hat{y}(t)\right)^2}}}{\sqrt{\overline{y^2(t)}}}\right) \ [dB], \qquad (2)$$

E_{unmod} is the modelling error obtained using an unmodulated quasi-unit pulse $u(\mathrm{t})$ train and E_{mod} is the improved modelling error obtained using a modulated pulse train $\tilde{u}(\mathrm{t})$.

Detector:

An SVM classifier with a Guassian kernel is used to detect amplitude modulated segments using the two modelling errors (E_{unmod} and E_{mod}) as features. The SVM outputs posterior probabilities which indicates the

likelihood of each observation belonging to one of the two classes. To refine the prediction accuracy, a second HMM is used in addition, which eliminates sudden transitions present in SVM prediction. The HMM parameters include a prior probability vector, observation matrix which contains the posterior probabilities output by SVM and a parameter matrix. The parameter matrix is optimized in the training phase to maximize the average of sensitivity and specificity of prediction. Optimum state sequence is estimated using Viterbi algorithm with the given HMM parameters.

III. RESULTS

The scatter plot of the two modelling errors of all the GAW segments is shown in Fig. 3. Here a GAW segment refers to all the adjacent GAW blocks with the same annotation as either amplitude modulated or amplitude unmodulated. Each point on the plot indicates modelling errors of a GAW segment which is obtained by taking the mean of modelling errors of all the blocks corresponding to the GAW segment. For amplitude modulated GAW segments, the modulating model resulted in a better modelling error than the non-modulating model. On the other hand, for most of the unmodulated GAW segments, modulating and non-modulating model resulted in similar modelling errors.

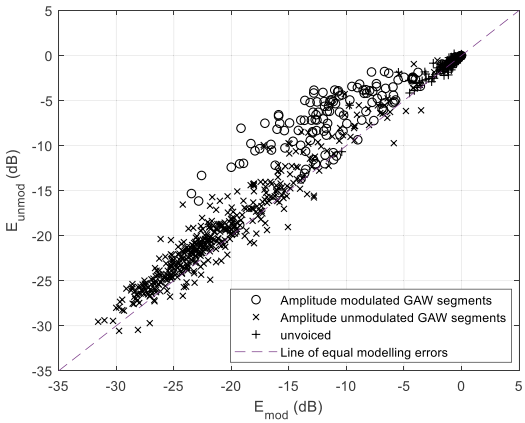


Figure 3: Scatter plot of modelling errors for all the GAW segments obtained using the modulating model (along the x-axis) and the non-modulating (along the y-axis).

Task 1: Detection of amplitude modulated GAW segments:

The two modelling errors are used as features for classification between amplitude modulated and unmodulated GAWs using the SVM. Posterior probabilities are estimated by the SVM. Fig. 3.a shows an example GAW having amplitude modulated segment in between unmodulated segments. The corresponding posterior probabilities increase for amplitude modulated GAW segments and decrease for unmodulated segments. Based on these probabilities, SVM predicts the presence or absence of amplitude modulation in the GAW with the decision threshold at 0.5. Sensitivity, specificity and accuracy of classifying the modulated and the unmodulated segments of all the GAWs are 0.45, 0.98 and 0.94 respectively. Sudden change in posterior probabilities causes instantaneous transition in state prediction, which is undesirable.

Further, to eliminate the instantaneous transitions in state predictions, the HMM is used, which increased the proposition of true positives. The sensitivity and specificity of predicting the amplitude modulated GAW blocks by the HMM are 0.69 and 0.96 respectively.

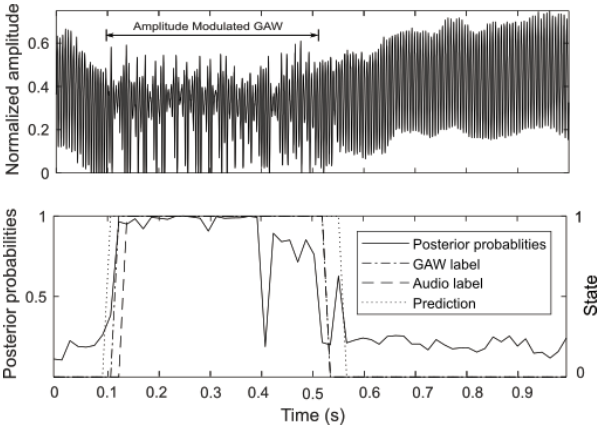


Figure 4: An example GAW containing amplitude modulated and unmodulated segments (top). Corresponding posterior probabilities of the GAW output by the SVM, HMM prediction, GAW and audio labels (bottom). GAW label at level '0' and '1' indicates the absence and presence of amplitude modulation in GAW. Audio label at level '0' and '1' indicates the absence and presence of fry voice quality.

Table 1: Sensitivity, specificity and accuracy of predicting all the amplitude modulated GAW segments and only amplitude modulated fry GAW segments.

Prediction results		Sensitivity	Specificity	Accuracy
Task 1	SVM	0.4575	0.9861	0.9465
	HMM	0.6916	0.9630	0.9426
Task 2	SVM	0.5785	0.9590	0.9547
	HMM	0.7972	0.9221	0.9207

Task 2: Detection of amplitude modulated fry GAW segments:

To analyze the prediction of amplitude modulated fry GAW segments, the detector trained for the

detection of amplitude modulated GAW segments is used. However, the non-fry amplitude modulated GAW segments are also used as negative data along with the negative data used in task one. Only, amplitude modulated fry segments are used as positive data. The sensitivity, specificity and accuracy of predicting the amplitude modulated fry GAW segments by the SVM and the HMM are given in Tab.1.

IV. DISCUSSION AND CONCLUSION

In this paper, amplitude modulation in GAWs is investigated. First, the GAWs are modelled using an analysis-by-synthesis approach to obtain two modelling errors for each input GAW. Amplitude modulated GAW segments are detected based on the modelling errors, E_{unmod} and E_{mod} using a SVM followed by a HMM. In task 1, the sensitivity of detecting amplitude modulated GAW is 0.69. Firstly, this could be improved by annotating the ground truth precisely. Noisy GAWs could have been wrongly annotated as amplitude modulated segments. Secondly, amplitude modulated GAW segments constitutes only seven percent of the total GAW segments used. Therefore, more positive data should be used. In task 2, the specificity of detecting the amplitude modulated fry segments decreased compared to specificity in task 1. This is because, in task 2, only amplitude modulated fry segments are used as positive data. The remaining amplitude modulated segments might belong to any other voice quality. The proposed detector provides a possible detection of amplitude modulation in GAWs. To distinguish amplitude modulated vocal fry from other amplitude modulated voice quality, modulator frequency could be investigated in the future.

V. ACKNOWLEDGEMENT

This work was supported by the Austrian Science Fund (FWF), KLI722-B30 and the University Hospital Erlangen kindly provided the segmentation tool.

REFERENCES

[1] 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.

[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] G. Degottex, "Glottal source and vocal-tract separation." *Dissertation*, Diss. Université Pierre et Marie Curie-Paris VI, **2010.**

[4] J. Laver. "The phonetic description of voice quality." *Cambridge Studies in Linguistics London* 31: 1-186, **1980.**

[5] 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.

[6] 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

[7] 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.

[8] C.T. Ishi, H. Ishiguro and N. Hagita, "Proposal of acoustic measures for automatic detection of vocal fry," *Ninth European Conference on Speech Communication and Technology*, **2005**

[9] T. Drugman , J. Kane and C. Gobl, "Data-driven detection and analysis of the patterns of creaky voice ," *Computer Speech & Language*, **2014**, 28(5), 1233-1253.

[10] P. Martin, "Automatic detection of voice creak," *Speech Prosody*, *Sixth International Conference*, Shanghai, China, **2012**.

[11] N. P. Narendra, and K. S. Rao. "Automatic detection of creaky voice using epoch parameters," *Sixteenth Annual Conference of the International Speech Communication Association*, **2015**

[12] V. Devaraj and P. Aichinger, "Modelling of Amplitude Modulated Vocal Fry Glottal Waveforms using an Analysis-by-synthesis Approach," *Applied Sciences,* 11(5), **2021**: 1990.

[13] P. Aichinger and F. Pernkopf, "Synthesis and Analysis-by-Synthesis of Modulated Diplophonic Glottal Area Waveforms." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, doi: 10.1109/TASLP.**2021**.3053387.

[14] G.Chen *et al*, "Estimating the voice source noise." *Thirteenth Annual Conference of the International Speech Communication Association*, **2012**

FORMANT ADAPTATION IN DIALOGUES

V. V. Evdokimova[1]

[1] Saint Petersburg State University/Department of Phonetics, Saint-Petersburg, Russian Federation
v.evdokimova@spbu.ru

Abstract: **Communicative adaptation is understood as a phenomenon when the speaker adopts the characteristics of the interlocutor. Phonetic adaptation assumes that there is an interaction between the perception and production of speech. The speech corpus SibLing consisting of 90 dialogues in Russian was used. The corpus consists of dialogues between pairs of twins and siblings with the difference of 1-2 years and all of them with close friends, strangers of the same sex, strangers of the opposite sex, strangers of higher rank and greater age. The formant characteristics of the vowels for all speakers were calculated. The average values of the Euclidean distance for the two first formants and the analysis of the formant pictures of vowels showed that in most cases there is a mutual adjustment of the interlocutors in the process of dialogue for vowel formant values.**

Keywords: **Speech adaptation, dialogue, vowel formants, phonetics**

I. INTRODUCTION

Communicative adaptation is a very complex phenomenon. In the process of communication, the speaker adopts the characteristics of their speech to the interlocutor. Phonetic adaptation is also a very complex phenomenon which assumes that there is an interaction between the perception and production of speech. When a listener perceives the speech of his interlocutor, the perceived utterances can influence the subsequent production of speech. In the process of speech production, the difference between the interlocutors can be explained by their physiology. However, if you do not take into account the difference in the physiological characteristics of speakers, the speech characteristics depend the situation, the emotions and physical state. Thus, the study of the phenomenon of adaptation allows us to understand changes in the process of speech production of a speaker [1, 2, 3]

The degree of adaptation is associated with various social factors (gender of participants, social status), individual differences, personality factors (the degree of openness of a person, the ability to concentrate, the number of social connections and the tendency to compromise), race, gender, the status of the interlocutor and the speaker's role in the performed task. The previous research shows that phonetic convergence increases with time and experience. Thus, speakers with very different phonetic characteristics seem to be more similar in later stages of interaction than in earlier stages [4] or after several trials [1, 2, 3, 5, 6, 7]. The previous research also shows that in the case of strangers, there was only slight convergence, while friends were more adapted to each other's speech. Moreover, scientists note that convergence also depends on the type of vowel: rounded vowels are more exposed to this phenomenon. Speakers shift vowels to varying degrees, especially for middle vowels [8].

The degree of adaptation depends on the role of the speaker: the explainers adapted to those to whom they were explaining [7]. Sometimes the experiment was conducted not on the analysis of all acoustic material, on the research of keywords. The results turned out to be quite interesting: in the process of conversation, the realizations of vowels and consonants became closer [9].

In another experiment, the researcher also found out how the formants shift in the process of performing tasks. The photo of the interlocutor on the screen influenced the degree of adaptation. Men showed more adaptation (for women, the situation was the opposite). When there was no photo on the screen, there were no differences in the degree of adaptation (regardless of the gender of the speaker) [2].

II. METHODS

Material: The SibLing Russian speech corpus was used in this study [10]. It was recorded at the Department of Phonetics of Saint Petersburg State University during the project 19-78-10046 “Phonetic manifestations of communication accommodation in dialogues” supported by the Russian Science Foundation. The SibLing corpus consists of 90 dialogues in Russian between pairs of twins and siblings with the difference of 1-2 years and all of them

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

with close friends, strangers of the same sex, strangers of the opposite sex, strangers of higher rank and greater age (a “boss”). Each dialogue consists of two tasks. In the first part the speakers tried to find similar notions on the dixit playing card. In the second task they explained the routes presented in the pictures. Therefore, the interlocutors used the same words in a dialogue.

The previous research on 7 dialogues from the SibLing corpus confirmed the presence of speech entrainment in the values of the formants of stressed vowels [12]. Based on the calculation of the Euclidean distance and the analysis of the vowel formant patterns, the following conclusions can be drawn:

1) In most cases, there is a mutual shift in the formant characteristics of vowels in the process of dialogue.

2) According to preliminary data, the degree of familiarity of the interlocutors quite strongly affects the speed of speech entrainment. The one case of divergence was observed. The degree of entrainment can be affected by the difference in age and social status.

3) The degree of adjustment of the acoustic characteristics depends on the quality of the vowel. To a greater extent, speakers adapt to each other by the rounded vowels of the back row (/o/, /u/), and also actively change the location of the vowels /a/ and /i/, adapting to each other. Moreover, the speakers often shift the focus of pronunciation of all cardinal vowels at once [12].

Method: The formant characteristics of the vowels for all speakers were calculated using the method for determining the characteristics of the transfer function of the vocal tract using the reverse filtering of the speech signal for the whole speech material and all the vowels in the material [11]. The formant characteristics were obtained for all speakers and all the interlocutors for both tasks in the speech corpus and allows analyzing the whole material.

In this paper the calculating the transfer function of vowels was performed on an unlabeled speech corpus. The written software makes it possible to carry out a general assessment of the average position of the formant characteristics for unannotated material and find the average values of the vowel formants. For each speaker, about 30,000 vowel processing 80 ms windows were calculated. Formant characteristics of vowels were calculated for all 90 dialogues in the corpus.

III. RESULTS

To calculate the difference in the relative position of the formant characteristics for two speakers in different types of tasks, the Euclidean distance between a vowel was calculated using two formants in Hertz. The results showed that in part of the dialogues, the formants were adjusted from the first to the second task. However, sometimes the closeness of the values of the formant characteristics was higher in the first task and dropped sharply in the second, which may be due to the fact that during the monologue, the share of which is greater in the second task, the announcers stop adjusting to the interlocutor and speak in their own more familiar mode.

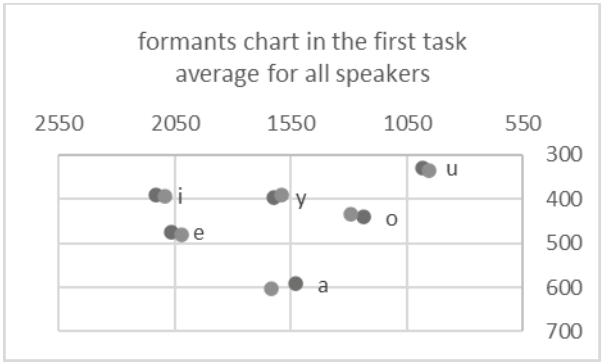


Figure 1. The average location of two first formant of the vowels on the formant chart in the first task.

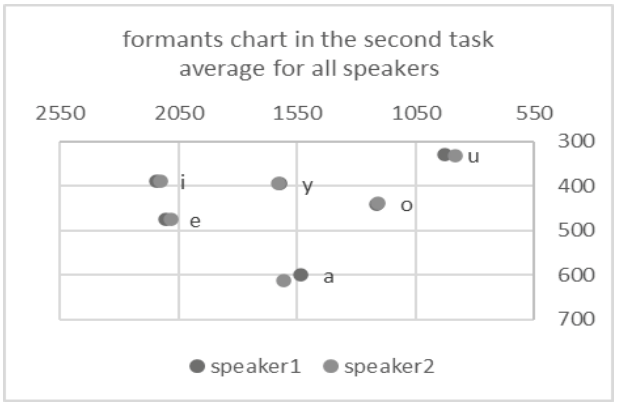


Figure 2. The average location of two first formant of the vowels on the formant chart in the second task.

The average values of the Euclidian distance and the analysis of the formant pictures of vowels showed that there is a mutual adjustment of the interlocutors in the process of dialogue. The formant chart shows that the places of the formants for two tasks become closer (Fig.1 and Fig.2).

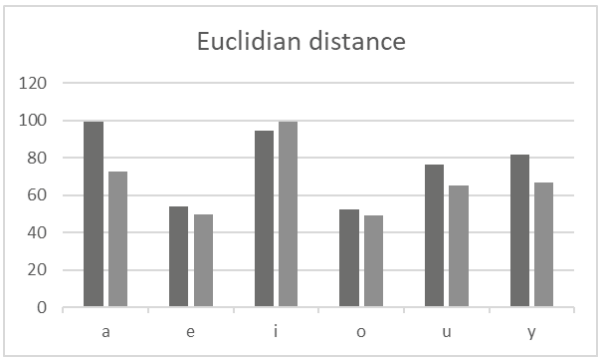


Figure 3. The average Euclidian distance between the formants in all dialogues.

The Euclidean distance decreases for vowels /o/, /u/|, /ɨ/ (sign y in the pictures), /e/. Also, speakers actively change, adjusting to each other, the location of the vowel / a / (Fig,3). Euclidian distance for /i/ increases on average for dialogues.

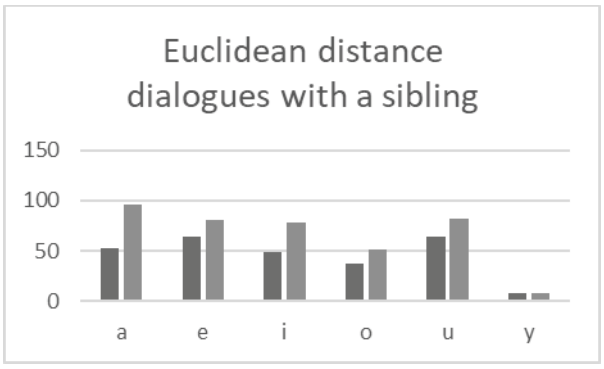


Figure 4. The average Euclidian distance between the formants in dialogues with a sibling.

In dialogues with a twin or sibling there is no adjustment in formants between two tasks. This can be explained by the closeness of their acoustic characteristics and similarity of the vocal tract characteristics.

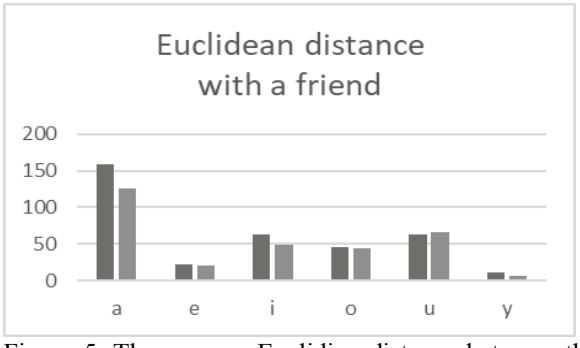


Figure 5. The average Euclidian distance between the formants in dialogues with a friend in the first and second tasks.

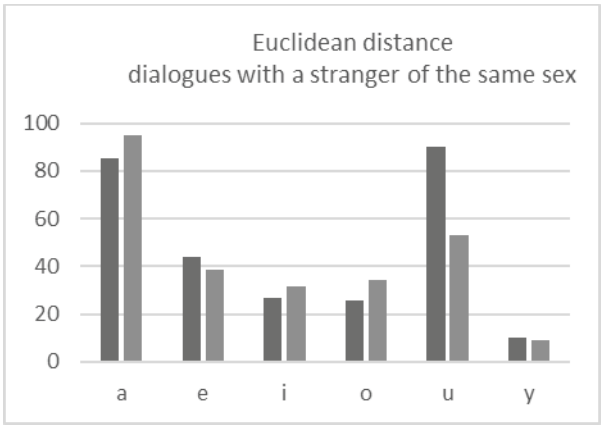


Figure 6. The average Euclidian distance between the formants in dialogues with a stranger of the same sex in the first and second tasks.

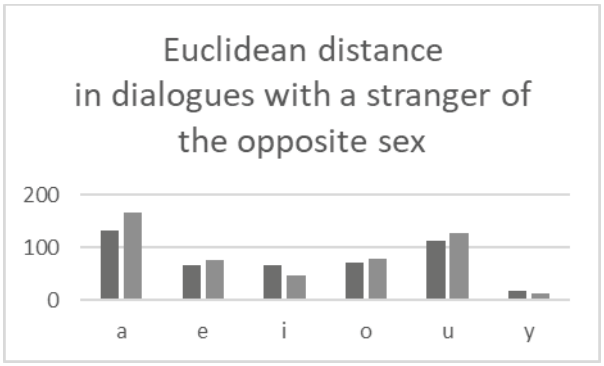


Figure 7. The average Euclidian distance between the formants in dialogues with a stranger of the opposite sex in the first and second tasks.

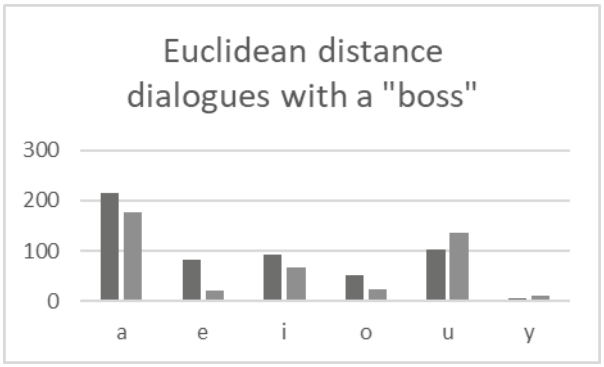


Figure 8. The average Euclidian distance between the formants in dialogues with a stranger of the same sex and higher age and position (a “boss”) in the first and second tasks.

The formant movement showed high variability for vowels within different types of dialogues with different interlocutors (Fig.4-8).

V. Conclusion

The analysis of the places of the first two formants in dialogues with a sibling or twin showed that there is no formant adaptation for speakers. In dialogues with a friend speakers try to adapt more. In dialogues with a strangers the Euclidean distance for formants for different vowels show different tendencies. However the average values of the Euclidean distance for the two first formants and the analysis of the formant pictures of vowels for all 90 dialogues in the corpus showed that in most cases there is a mutual adjustment in formants of the interlocutors in the process of a dialogue.

VI. Acknowledgements

The research is supported by the Russian Science Foundation (grant 19-78-10046 "Phonetic manifestations of communication accommodation in dialogues").

REFERENCES

[1] M. Babel, "Selective Vowel Imitation in Spontaneous Phonetic Accommodation", *UC Berkeley PhonLab Annual Report*, 5, 2009.

[2] M. Babel, "Phonetic and Social Selectivity in Speech Accommodation", University of California, Berkeley. ProQuest Dissertations Publishing, 2009. 3382831.

[3] T.V. Kachkovskaia, A.D. Mamushina, " PHONETIC MANIFESTATIONS OF COMMUNICATION ACCOMMODATION IN DIALOGUE", Voprosy Jazykoznanija (Topics in the study of language), 2021, 2: pp. 123–141.

[4] J.S. Pardo "On phonetic convergence during conversational interaction", *The Journal of the Acoustical Association of America*, 2006, 119(4): pp. 2382–2393.

[5] S.D. Goldinger "Echoes of echoes? An episodic theory of lexical access", *Psychological Review*, 1998, 105(2): pp. 251–279.

[6] U.C. Priva, Ch. Sanker. "Convergence is predicted by particular interlocutors, not speakers." (2019).

[7] Cohen Priva, U., Edelist, L., Gleason, E. 2017. "Converging to the baseline: Corpus evidence for convergence in speech rate to interlocutor's baseline", The Journal of the Acoustical Society of America 141(5), pp. 2989–2996.

[8] G. Bailly, A. Lelong, "Speech dominoes and phonetic convergence", *Proc. of Interspeech 2010*. T. Kobayashi, K. Hirose, S. Nakamura (eds.), "International Speech Communication Association, 2010, pp. 1153–1156. https://www.isca-speech.org/archive/interspeech_2010/i10_1153.html

[9] V. Aubanel, N. Nguyen, "Automatic recognition of regional phonological variation in conversational interaction", 2010, Speech Communication. 52. Pp. 577-586. 10.1016/j.specom.2010.02.008.

[10] A. Menshikova, T. Kachkovskaia, T. Chukaeva, V. Evdokimova, P. Kholiavin, N. Kriakina, D. Kocharov, A. Mamushina, A. Menshikova, S. Zimina, "SibLing Corpus of Russian Dialogue Speech Designed for Research on Speech Entrainment". In: Proc. Of LREC2020

[11] V. Evdokimova, D. Kocharov, P. Skrelin, (2020). "Method for Constructing Formants for Studying Phonetic Characteristics of Vowels", *SPIIRAS Proceedings*, *19*(2), pp. 302-329. https://doi.org/10.15622/sp.2020.19.2.3

[12] S.Zimina, V.Evdokimova "Acoustic Characteristics of Speech Entrainment in Dialogues in Similar Phonetic Sequences", SPECOM 2021, LNAI 12997, 2021 (in print).

TESTBED DESIGN FOR IN-VITRO CHARACTERISATION OF BIOMIMETIC VOCAL FOLDS

R. Girault[1], H. Yousefi-Mashouf [1,2], P.Luizard [1],
L. Bailly[2], L. Orgéas [2], X. Laval [1], N. Henrich Bernardoni [1]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
[2] Univ. Grenoble Alpes, CNRS, Grenoble INP, 3SR, 38000 Grenoble, France
lucie.bailly@3sr-grenoble.fr , nathalie.henrich@gipsa-lab.fr

Abstract: ***An in-vitro testbed was designed to study the vibromechanical behaviour of biomimetic and extensible vocal folds. The paper describes the several steps of its conception. It consists of a deformable laryngeal envelope in which stretchable vocal-fold replicas of adjustable material and structural properties can be inserted for testing. The folds are able to oscillate for a wide range of aerodynamic conditions and material elongations.***
Keywords: **biomimetic vocal fold, in vitro testbed, self-oscillation**

I. INTRODUCTION

The understanding of human vocal-fold vibratory properties is based on the development of testbeds that allow to reproduce self-sustained oscillations [1-4]. Such *in vitro* set-ups are used to study the physics of phonation. Most models already developed in past decades correspond to deformable folds, albeit fixed in a given geometry and pre-tension prior oscillations. Their microstructure has been progressively refined, ranging from isotropic and mono-layered oscillators to anisotropic and multi-layered ones [1,4]. To our knowledge however, only two studies present extensible folds, even though vocal-fold stretching is a major aspect of phonation biomechanical control [5,6]. The aim of this work was to design a testbed dedicated to the study of vibromechanical behaviour of biomimetic vocal folds. Emphasis was placed on the possibility of reproducing the actions of crico-thyroid tilt (fold stretching) and inter-arytenoid compression (fold abduction and adduction).

II. METHODS

A. A deformable laryngeal envelope

The chosen approach was to enable folds actuation in stretching and compression. We designed a flexible laryngeal envelope made up of silicone (Ecoflex™ 00-50), into which vocal folds to be tested could be inserted while maintaining a seal. This three-part envelope is shown in Figure 1. It consists of i) a subglottal tract attached to the air inlet tube, and representing the trachea upper part (subglottal stage); ii) a divergent tract that joins the subglottal and glottal stages; iii) a case in which deformable folds can be positioned (glottal stage). An air pressure sensor is inserted in the subglottal stage.

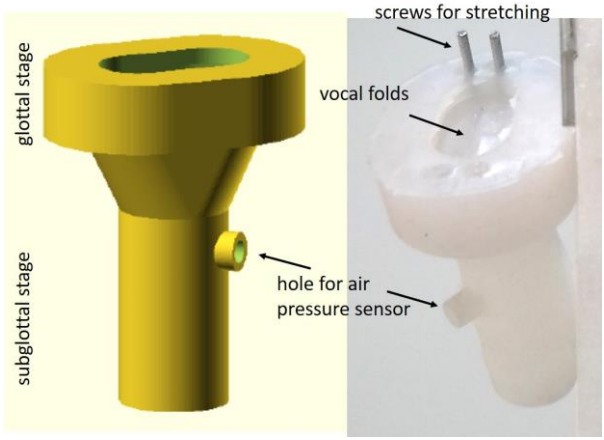


Figure 1 : Design of a deformable laryngeal envelope

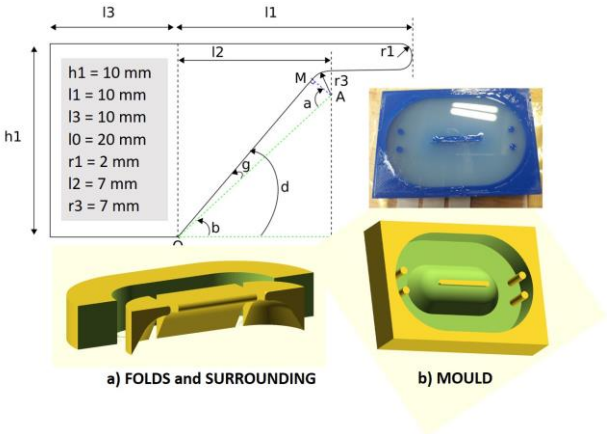


Figure 2 : (a) : 3D initial geometry chosen for one deformable fold and its surrounding (subglottal convergent tract and glottal plane). (b) : illustration of the 3D-printed mould of the folds, and its CAO design.

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

B. Choice of material and vocal-fold design

Our ultimate goal is to design architectured vocal folds. In a first step, homogeneous mono-layered oscillators were conceived, made of isotropic polymers and processed using 3D-printed moulds (see Figure 2). Two complementary approaches were tested, using either : (i) silicone rubbers of reference (Ecoflex™ 00-10, 00-30, 00-50) owning distinct mechanical properties (*e.g.* shore hardness, tensile strength, elongation at break, viscosity); (ii) cross-linked gelatin-based hydrogels chosen for their ability to retain high tissue-like water content. In a second step, a two-layered version was made. The surface layer can be homogeneous or composed of a fiber mat.

C. Mechanical behaviour in tension and compression

The mechanical behaviour of the selected materials were explored in tension and compression, using a uniaxial machine (INSTRON® 5944) equipped with a ±10 N load cell, and combined to an hygro-regulated chamber. These mechanical tests made it possible to better understand the impact of each material's formulation on its stress-strain behaviour, and to characterize the impact of various loading conditions (e.g. loading mode, cyclic paths, deformation rate) on its mechanical properties. This step allowed to compare the mechanical properties of each vocal-fold model to reference database previously acquired on native vocal folds [7], and to better understand their vibromechanical behaviour observed under fluid-structure interaction in step E.

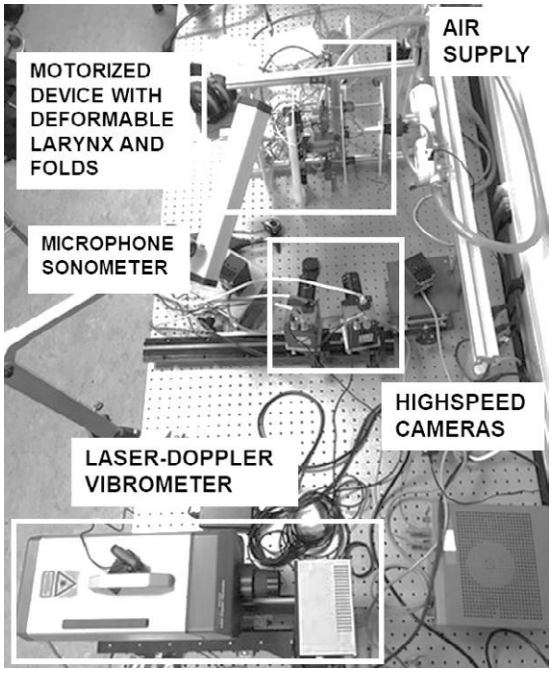


Figure 3 : Photo of the testbed with all equipments

D. Vibromechanical characterisation

In addition to the mechanical characterisation of the selected materials, the vibromechanical behaviour of the 3D vocal folds were measured by laser-Doppler vibrometry. The tesbed is shown in Figure 3. The structure was installed on the testbed in its phonatory position and without air supply. The vibromechanical eigenmodes were excited with a sine sweep signal delivered by a vibration shaker.

E. Behaviour in fluid-structure interaction

Pressurised air injected into the model through the subglottal tract enabled the synthetic folds to vibrate. Subglottal air flow was controlled by the degree of valve opening and measured with a flow-meter. For different levels of fold elongation, the aero-acoustic behaviour was characterised by measuring aerodynamic parameters (subglottal phonation pressure, vocal efficiency) and acoustic parameters (intensity, frequency, spectral centroid). Vibratory behaviour was characterised by highspeed cinematography through computing kymograms and glottal area waveforms.

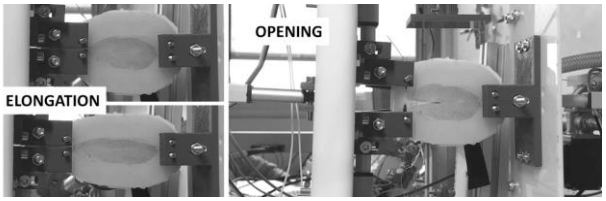


Figure 4 : Illustration of stretching action and abduction/adduction on deformable folds.

III. RESULTS

A. Ability to be stretched and abducted/adducted

Figure 4 illustrates the stretching action on vocal-fold replica. In a first measurement campaign, the stretch ratio λ=L/L0 was varied from 1.05 to 1.21, L (resp. L0) being the length of the fold in the deformed (resp. undeformed) configuration. A greater elongation could be obtained, yet not tested for allowing the whole measurement campaign without damaging the biomimetic folds.

B. Ability to oscillate in fluid-structure interaction

As illustrated in Figures 5 and 6, vocal folds were able to oscillate in response to an increase of subglottal air flow in most cases. The greater the material stiffness, the higher the subglottal pressure measured during stabilized vibration. For a given material, the greater the applied stretching, the higher the subglottal

pressure needed to self-oscillate. Homogenous vocal folds in Ecoflex™ 00-10 demonstrated self-oscillations on a limited mid-to-high range of airflow (from 1.5 to 3.5 L/s), with lesser subglottal pressure variation. On the other side, oscillations of hydrogel-based folds were limited to low-to-mid range of airflow (from 0.5 to 2.5 L/s). The linear relationship between subglottal pressure and airflow was similar for all materials and stretching conditions, except for folds in Ecoflex™ 00-10 for which subglottal pressure did not vary much.

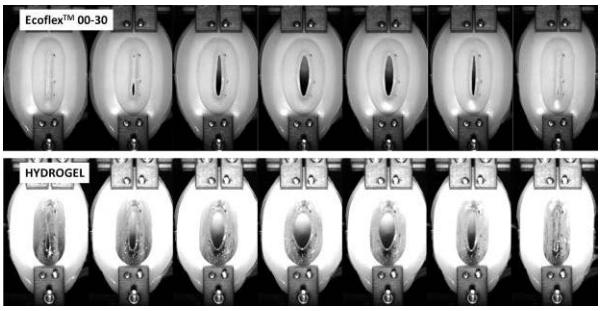


Figure 5 : Visualisation of a glottal vibratory cycle for two types of material. Top : silicone. Bottom : cross-linked gelatin. The laryngeal envelope is identical in both cases.

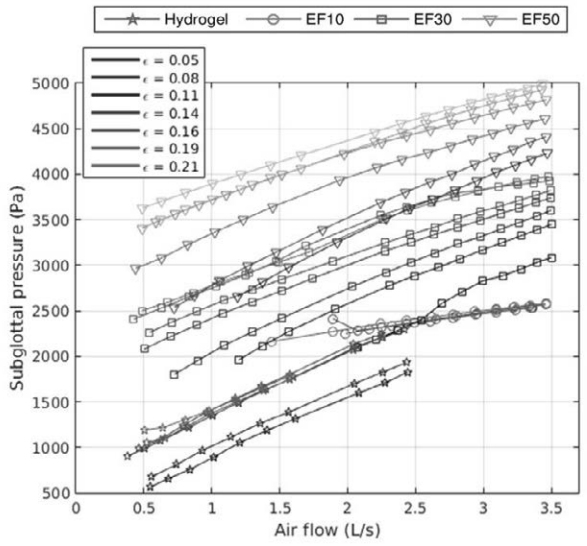


Figure 6 : Subglottal pressure as a function of air flow (control parameter) for three different types of silicone (variable stiffness poperties) and for the cross-linked hydrogel. Each dot represents mean value computed on a sequence of stabilized vocal-folds oscillation.

IV. Discussion

With a simple and homogenous model of vocal folds and laryngeal envelope, self-sustained oscillations were obtained on a wide range of subglottal airflow and pressures. Stretching did not impede the oscillatory capabilities for Ecoflex™ 00-30 and Ecoflex™ 00-50 folds. Yet Ecoflex™ 00-10 folds could not oscillate at low airflow rate, and hydrogel-based one at high airflow rate.

Several improvements can be made. Concerning the folds, a two-layer version has been designed, yet not tested. Other geometries of vocal folds can also be designed, closer to synthetic models used in the literature [7]. Replicas of arytenoid cartilages can be easily 3D printed and inserted in the posterior part.

V. Conclusion

The implemented approach allows the development of biomimetic vocal folds step by step, so as to better understand the relationships between material and structural properties of the replica, and their vibratory outcomes under fluid/structure interaction. Vocal folds designed separately (left-right) or as a whole (monobloc setting) were able to sustain self-oscillation at a wide range of subglottal flows and pressures, and at various degrees of stretching.

References

[1] T.E. Greenwood, and S.L. Thomson, “Embedded 3D printing of multi-layer, self-oscillating vocal fold models.” *Journal of Biomechanics*, 121, 110388, 2021.

[2] L. Bailly, X. Pelorson, N. Henrich, and N. Ruty, “Influence of a constriction in the near field of the vocal folds: Physical modeling and experimental validation,” *J. Acoust. Soc. Am.*, 2008, 124, pp. 3296–3308, 2008.

[3] S. Kniesburges, S.L. Thomson, A. Barney, M. Triep, P. Sidlof, J. Horacek, C. Brucker and S. Becker, “ In Vitro Experimental Investigation of Voice Production”, Current Bioinformatics 2011; 6(3)

[4] Murray, P. R., & Thomson, S. L. (2012). Vibratory responses of synthetic, self-oscillating vocal fold models. The Journal of the Acoustical Society of America, 132(5), 3428-3438.

[5] S. M. Shaw, S. L. Thomson, C. Dromey, and S. Smith, “Frequency response of synthetic vocal fold models with linear and nonlinear material properties.”, JSLH, vol. 55 (5), pp. 1395-1406, 2012

[6] S. Kniesburges, R. Veltrup, S. Fattoum, and A. Schützenberger,“Modeling the pre-phonatory vocal fold posture in the larynx model SynthVOICE”, Proc. ICA, 2019

[7] T. Cochereau, L. Bailly, L. Orgéas, N. Henrich Bernardoni, S. Rolland du Roscoat, Y. Robert, M. Terrien (2020). “Mechanics of human vocal folds layers during finite strains in tension, compression and shear”. *Journal of Biomechanics*, 110:0021-9290.

LARGE-EDDY SIMULATION OF HUMAN PHONATION USING THE ANISOTROPIC MINIMUM-DISSIPATION MODEL

M. Lasota[1,3], P. Šidlof[1,2]

[1] Faculty of Mechatronics, Technical University of Liberec, Studentská 2, 461 17 Liberec 1, Czech Republic
[2] Institute of Thermomechanics, Czech Academy of Sciences, Dolejškova 5, 182 00 Prague 8, Czech Republic
[3] Department of Technical Mathematics, CTU in Prague, Karlovo náměstí 13, 121 35 Prague 2, Czech Republic

Abstract: ***This contribution is focused on numerical modeling of 3D incompressible laryngeal flow through healthy vocal folds oscillating at a fundamental frequency of 100 Hz. The investigation is based on a realistic CFD simulation of turbulent flow by Large-Eddy Simulation with various subgrid-scale models and monitoring their influence on the aeroacoustic spectrum during human phonation of five vowels /u, i, a, o, æ/.***

Keywords: **human phonation, turbulent flow, aeroacoustic simulations, vocal tract shapes, vowels**

I. INTRODUCTION

Voice production has been investigated by experimental measurements and numerical simulations. However, the experiments often bring numerous limitations, especially when the in-vivo measurements have to be carried out. High-performance computing can be used as an alternative method An extensive list of numerical models of human phonation is available in [1]. The current contribution presents a computational aeroacoustic model of human phonation based on high-resolution Large-Eddy Simulation of glottal flow with advanced turbulence modeling.

II. METHODS

Considering the enormous disparity of scales in the flow and acoustics, the aeroacoustic simulation has been divided by computing the flow by the finite-volume method, and subsequently the sound sources and wave propagation by the finite-element method, see [2] for more details. The Large-Eddy Simulation of flow resolves the large-scale vortices, while the influence of the subgrid-scale vortices is modeled by a subgrid-scale closure model. Various subgrid-scale models have been studied to cope with near-wall modeling in the glottis, where inaccurate prediction of the shear stress at the surface of vocal folds delays the transition to turbulence. The dominant sound source caused by flow-induced vibration of vocal folds lie within the glottis. Thereby the geometry and kinematics were specified with care, but some necessary simplification had to be included.

Fig. 1 presents the simplified 3D model of the larynx in a coronal view with a square cross-section in the straight subglottal and supraglottal segments 12x12 mm (y-z plane). The kinematics of vocal folds is prescribed by sinusoidal displacement of inferior-superior margins in medial-lateral (y) direction with two degrees of freedom with the amplitude A=0.3 mm for both vocal folds, allowing closing/opening the glottal gap g in the range 0.42-1.46 mm, having the medial surface convergence angle $\psi/2$ (in clockwise) for convergent and divergent position -10° and +10°, respectively, and with the same phase difference $\pi/2$ between the inferior and superior vocal fold margin on both vocal folds. The distance (y) between both ventricles and false vocal folds is equal to 16 and 6.15 mm, respectively. The boundary conditions for the fluid flow are listed in Tab. 1.

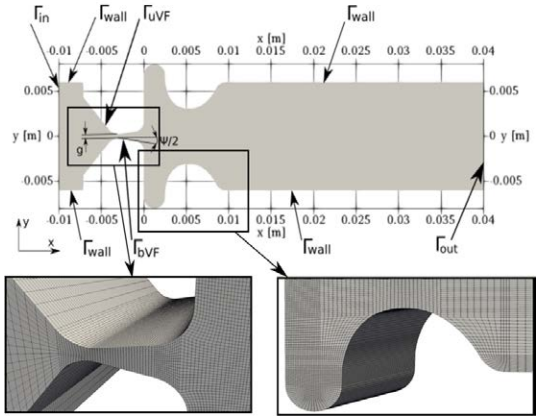


Figure 1 - Geometry, boundaries and mesh of the larynx

Table 1 – Boundary conditions of filtered flow variables velocity and static pressure.

Boundary	$\overline{U}$ [ms^{-1}]	$\overline{p}$ [Pa]
Inlet Γ_{in}	from flux,$\overline{U}_i n_i < 0$ 0,$\overline{U}_i n_i > 0$	307.4
Outlet Γ_{out}	$\nabla(\overline{\boldsymbol{U}}) \cdot \boldsymbol{n} = 0$,$\overline{U}_i n_i > 0$ $\overline{\boldsymbol{U}} = 0$,$\overline{U}_i n_i < 0$	0
Vocal folds Γ_{bVF}, Γ_{uVF}	$\overline{U}_2 = \frac{\partial}{\partial t} h(\boldsymbol{x}, t)$ $\overline{U}_1 = \overline{U}_3 = 0$	$\nabla(\overline{p}) \cdot \boldsymbol{n} = 0$
Fixed walls Γ_{wall}	$\overline{\boldsymbol{U}} = 0$	$\nabla(\overline{p}) \cdot \boldsymbol{n} = 0$

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

Healthy phonation reaches the Reynolds number in the range (100-10,000), thereby the flow field is highly turbulent. The large eddies carrying the most energy in the flow are resolved directly by Navier-Stokes equations (NSE), whereas the small scales are modeled applying a spatial filter ($\bar{\cdot}$) to the NSE, that gives

$$\partial_t + \partial_j(\overline{u_i}\overline{u_j}) - \partial_j(\nu\partial_j\overline{u_i}) = -\partial_i(\bar{p}) - \partial_j\tau_{ij},\ \partial_i\overline{u_i} = 0, \quad (1)$$

where the subgrid-scale tensor $\tau_{ij}(u)$ represents the effect of small scales on directly resolved large eddies. To include local increasing of turbulent (eddy) viscosity ν_t from the unresolved eddies into molecular viscosity ν is applied the eddy-viscosity equation

$$\tau_{ij} - \frac{1}{3}\tau_{kk}I_{ij} = -2\nu_t S_{ij}, \quad (2)$$

where I_{ij} is the identity matrix and S_{ij} is resolved rate-of-strain tensor. This study is focused on approximation of ν_t by various turbulence subgrid-scale models, namely the standard One-Equation (OE) [3], Wall-Adapting Local-Eddy (WALE) [4] and a newly implemented Anisotropic Minimum Dissipation (AMD) model [5]. For completeness, the fourth case (LAM) is included with no turbulence modeling. The used acoustic grids are shown in Figs. 2-6, varying in shapes for each vowel.

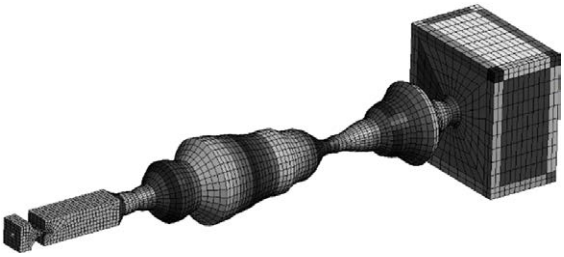


Figure 2 - Acoustic mesh for /u/

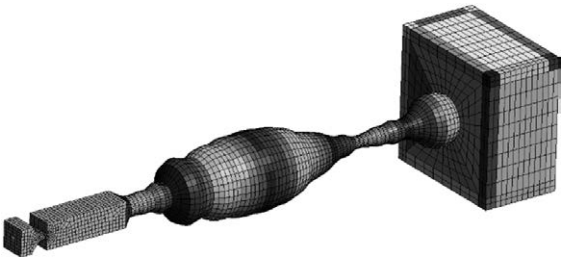


Figure 3 - Acoustic mesh for /i/

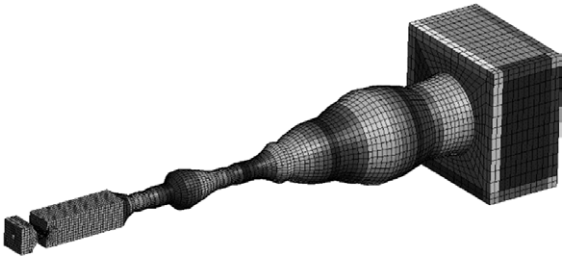


Figure 4 - Acoustic mesh for /a/

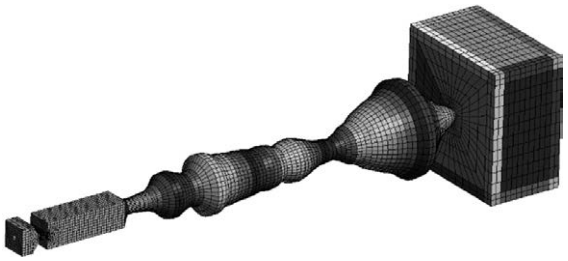


Figure 5 - Acoustic mesh for /o/

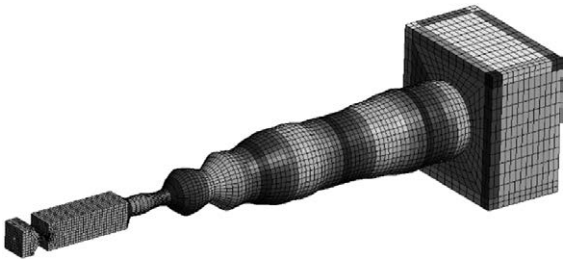


Figure 6 - Acoustic mesh for /æ/

At the laryngeal and vocal tract walls, a sound-hard boundary condition perfectly reflecting the sound waves are specified. At the boundary of the radiation region and subglottal inlet, a Perfectly Matched Layers suppressing the acoustic reflections are used. The Perturbed Convective Wave Equation [6] were used to solve the acoustic potential ψ^a from the partial differential equation

$$\frac{1}{c_0^2}\frac{D^2\psi^a}{Dt^2} - \nabla\cdot\nabla(\psi^a) = -\frac{1}{\rho c_0^2}\frac{D\bar{p}^{ic}}{Dt}, \quad (3)$$

where the acoustic potential is equal to the acoustic pressure in this case of $\rho \approx 1$. The probe location MIC1 (Fig. 7), 1 cm from mouth, is used for the Fast Fourier Transform ($\Delta f \approx 5$ Hz).

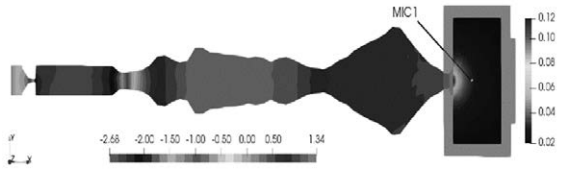


Figure 7 – Coronal view of the computational domain for /o/. Distribution of acoustic potential $\psi^a(t)$

III. RESULTS

Four CFD simulations over 20 periods of vocal fold oscillations were realized, with subsequent aeroacoustic simulations yielding the acoustic spectra (Figs. 8-12). The usage of subgrid-scale turbulence models does not modify positions of formant frequencies, but it modifies considerably the SPLs. The simulations using the new AMD model enforced higher harmonics compared the rest of models, except the higher harmonics at low frequencies in the spectrum of vowel /i/.

Vowel /u/. SPLs at F_2=1000 Hz and $F_3 \approx$ 2500 Hz are nonuniform, at the second formant AMD is higher by 22 % than WALE, and subsequently at the third formant WALE is higher by 28 % than AMD. This trend occurs only for vowels /u, *a*/.

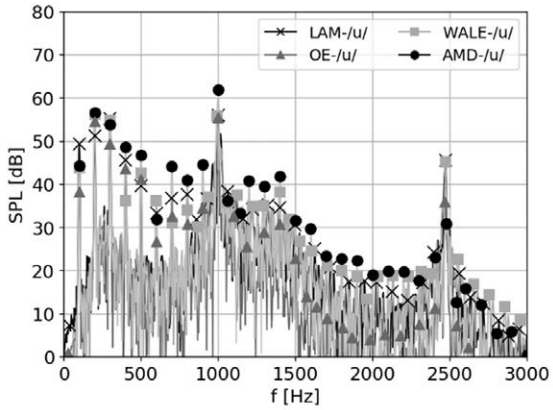


Figure 8 - Acoustic sound spectrum for /u/

Vowel /i/. At f_o=100 Hz the SPL for the AMD model refers to a very low value around 35 dB, whereas the other SPLs are minimally 17 % higher. In general the SPLs for the AMD model are low in low-frequency bandwidth, but for the $F_2 \approx$1400 Hz the AMD model is enforced by 21 % compared to WALE. In the situation at F_3=2500 Hz the WALE and AMD models are close to each other and 1-2 dB higher than the laminar case. This could mean that the AMD and WALE models have a similar behavior at high-frequency bandwidth, but this assumption holds only for /i, o, æ/. The WALE model also enforced the sound pressure levels at F_2 and F_3 stronger than the OE model, by 6 dB and 15 dB, respectively, which is related to higher flow rate through the glottis.

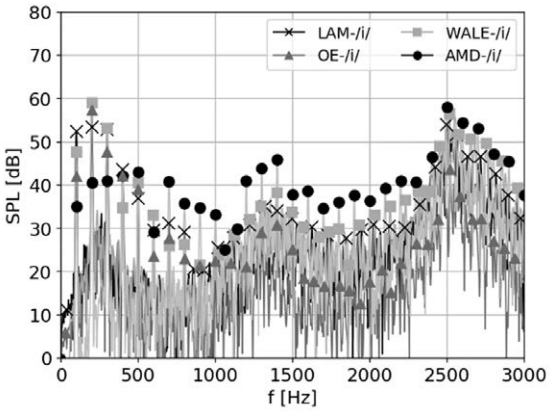


Figure 9 - Acoustic sound spectrum for /i/

Vowel /a/. SPLs at f_o stayed at similar levels for the WALE and AMD models, which happened only twice, in cases /*a*, æ/. The space between formants F_1-F_2 is typical for vowels /u, *a*, o/, but in simulation of /*a*/ the second formant (around 1300 Hz) was not distinct. On the other hand, the third formant is clearly visible and presents the same behavior as it was seen in /u/, i.e. a big drop predicted by AMD up to 9 and 13 dB compared to WALE and LAM, respectively.

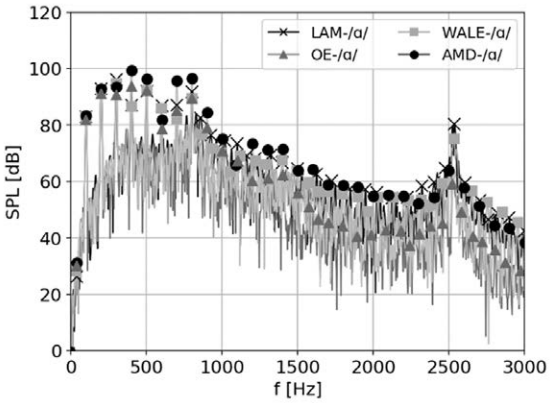


Figure 10 - Acoustic sound spectrum for /a/

Vowel /o/. The SPLs are held at high levels up to 3 kHz, with some little skips, however the vocal tract shape (see again Fig. 5 and 7) has contained the widest throat 7.25 cm^2 of presented acoustic grids.

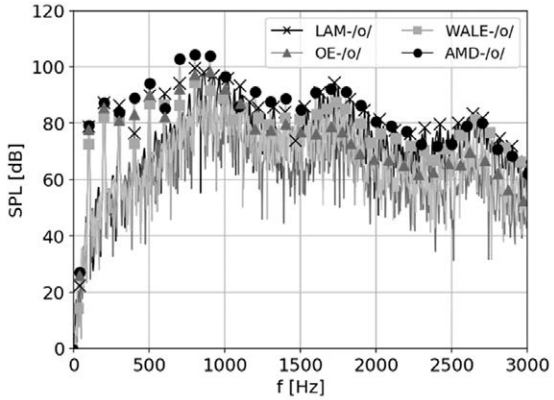


Figure 11 - Acoustic sound spectrum for /o/

Vowel /æ/. In this case, the AMD model highlighted the first formant very well, 14 dB higher than the WALE. The predictions of the SPL of the second and third formants by various SGS models were similar, with differences less than 3 dB.

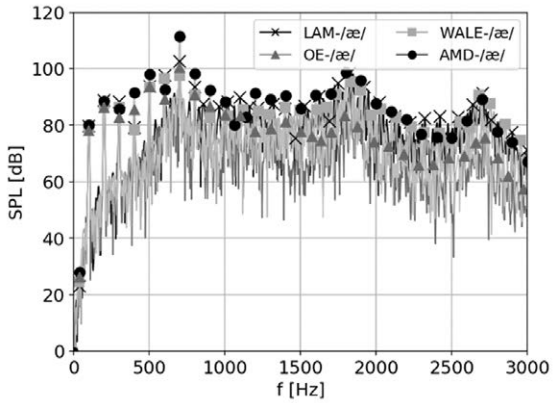


Figure 12 - Acoustic sound spectrum for /æ/

IV. Discussion

So far, the AMD subgrid-scale model has not been studied in the application associated with human phonation. In addition, the computational efforts can be reduced compared to conventional subgrid-scale models, due to the more accessible algorithm computing no invariants from the rate-of-strain tensor.

V. Conclusion

In the result section was demonstrated that the subgrid-scale models have a considerable impact on sound pressure levels. In the simulations using the OE model, formants were hardly visible and significantly weaker compared to other models. Usage of the WALE model, which is well-known to handle the turbulent viscosity at near-wall and high-shear regions more precisely than the OE model, predicted 7 % higher volumetric flow rates of air through the glottis compared to the OE model, and only slightly lower than the LAM and AMD models (by 9 and 4 %, respectively). The (widely-used) WALE model was able to uncover all characteristics for identification of formant frequencies, even so it gave the best recognition of third formants in high-frequency bandwidth in cases /u, *a*/. In contrary, the newly implemented AMD model has proven a good agreement with the WALE model, and even more it identified the SPLs at lower formants F_1 and F_2 the most evident, with the exception of the vowel /i/.

The research was supported by the Czech Science Foundation, project 19-04477S Modelling and measurements of fluid-structure-acoustic interactions in biomechanics of human voice production, and by the Student Grant Scheme at the Technical University of Liberec through project no. SGS-2020-3068.

REFERENCES

[1] Alipour, F., Brücker, C., Cook, D., Gömmel, A., Kaltenbacher, M., Willy, M., et al. Mathematical models and numerical schemes for the simulation of human phonation. *Current Bioinformatics.* **2011**, 6(3).
[2] Lasota, M.; Šidlof, P.; Kaltenbacher, M.; Schoder, S. Impact of the Sub-Grid Scale Turbulence Model in Aeroacoustic Simulation of Human Voice. *Applied Sciences*. **2021**, 11, 1970
[3] Davidson L. Hybrid LES--RANS: back scatter from a scale-similarity model used as forcing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. **2009**. Vol. 367. No. 1899. pp. 2905-2915.
[4] Nicoud, F. and Ducros, F. Subgrid-scale stress modelling based on the square of the velocity gradient tensor. *Flow, Turbulence and Combustion*. **1999**. Vol. 62. No. 3. pp. 183-200.
[5] Rozema, W., Bae, H., Moin, P. and Verstappen, R. Minimum-dissipation models for large-eddy simulation. *Physics of Fluids*. **2015**. Vol 27. No. 8.
[6] Hüppe, A., Grabinger, J., Kaltenbacher, M., Reppenhagen, A., Dutzler, G., Kühnel, W. A non-conforming finite element method for computational aeroacoustics in rotating systems. *20th AIAA/CEAS Aeroacoustics Conference.* 2014.

PRACTICAL GUIDELINES FOR IMPLEMENTING VOCAL TRACT RESONANCES CHARACTERIZATION WITH EXCITATION AT THE LIPS

T. Maison[1], B. Allain[1], P. Hoyer[2], F. Silva[1], P. Guillemain[1], N. Henrich Bernardoni[3]
[1] Aix-Marseille Univ., CNRS, Centrale Marseille, LMA, Marseille, France
[2] Fraunhofer Headquarters, Hansastr, 27c, 80686 Munich, Germany
[3] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, Grenoble, France

timothee.maison@ens-lyon.org, baptiste.allain@centrale-marseille.fr, patrick.hoyer@zv.fraunhofer.de, silva@lma.cnrs-mrs.fr, guillemain@lma.cnrs-mrs.fr, nathalie.henrich@gipsa-lab.fr

Abstract: **The measurement of vocal tract resonances is crucial to understand voice acoustics. Their characterization using a broadband excitation signal and pressure measurements both at the lips is a good compromise between accuracy, speed and intrusiveness. In this paper we address some practical guidelines for performing such measurements in order to provide reliable estimates for resonance frequencies and quality factors. We also experiment the possibility to move away microphone from excitation at the lips, with a cylindrical waveguide as ideal vocal tract and its line-transmission model. Using a far-field model in which pressure decreases as the inverse of the distance from excitation at the lips, the microphone could be placed anywhere as soon as the Signal-to-Noise Ratio is good enough: measurements become more sensitive to interferences with other acoustics sources. Modal analysis is performed with a robust method using both amplitude and phase of measured functions. Resonance frequencies and quality factors estimations at distances up to 30 cm deviates by less than 0.2% and 10% respectively from reference measurement at the lips validating accurate characterization with a distant microphone from the lips.**

Keywords: **vocal tract, resonance, measurement, radiation, acoustics**

I. INTRODUCTION

Vocal tract resonances are essentials for spoken communication and they enhance singing voice efficiency [1]. Their characterization has led to the development of several devices with a compromise between intrusiveness and accuracy. Indirect methods based on formants analyses in the voice signal (such as linear predictive coding) evidenced severe limitations for high pitches or closed vowels. We focus here on a non-invasive device based on a broadband excitation signal at the lips, first proposed by Epps [2]. It has undergone several developments and variations such as the use of sine-sweep signals [3], a buzzer excitation device [4], or the direct estimation of pressure and velocity [5]. Recently the measurement model of this approach has been clarified, also showing its limitations [6]. While the measurement setup seems to be quite simple, accurate estimate of resonance frequencies and quality factors requires some precautions. From the hardware system implementation to the signal processing of a measurement and the estimation of modal parameters, each step needs precise adjustments. We first address a full methodology and some practical guidelines for performing such measurements in the best possible conditions (Sec. II), leaving aside for the moment the question of real-time estimation. Motivated by the high voice level at the lips that can saturate and damage the microphone (physical clipping), we explore the possibility to separate pressure measurement from excitation (traditionally both done at the lips), by means of experiments and radiation theory (Sec. III). We study here the unvoiced case. All results are expected to remain valid in the voiced case assuming a perfect separation between the excitation signal and the voice.

II. METHODS

A. Measurement device and theoretical context

The device consists of a flexible capillary (diameter *4 mm*) connected to an impedance adapter on a loudspeaker (Beyma CP850Nd) with amplifier (Flying Mole DAD-M100 proII BB) and sound card (Focusrite 6i6), positioned at the inlet of a vocal tract model (an ideal open-closed cylindrical waveguide with length *15* cm and diameter *21* mm for validation purpose). The flexible capillary is coated by a thicker tube in order to remove its sides radiation (see Figure 1): the only acoustic source should be the output of the capillary at the lips – otherwise interferences may

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

occur. The loudspeaker and impedance adapter with the tube must also be acoustically isolated for the same reason.

Figure 1 – Photo of excitation at the waveguide inlet.

As usually done [1-3, 5, 6], the microphone is positioned close to the excitation system at the lips. It records the waveguide (vocal tract) responses to a broadband signal in a calibration condition (closed-mouth) and in a measurement condition (open-mouth). The pressure measurement model at the lips, valid for all kinds of excitation signal, has been described in [6]. The spectrum in open-mouth condition $P_{meas}(\omega)$ calibrated by closed-mouth spectrum $P_{cal}(\omega)$ gives access to a frequency response $H(\omega)$ characterizing the vocal-tract acoustics including its radiation (Equation 1). The excitation capillary is small enough (high impedance) to be assumed independent from the load (i.e. it provides the same source flow for open or closed vocal tract).

$$H(\omega) = \frac{P_{meas}}{P_{cal}} = \frac{Z_{VT}}{Z_{VT} + Z_R} \qquad (1)$$

with Z_{VT} the vocal tract input impedance seen from the lips and Z_R the vocal tract radiation impedance. The resonances of $H(\omega)$ are identical to a usual vocal tract transfer function pressure at the lips over flow at the glottis (P_{lips} / $U_{glottis}$). The lips radiation is included, so these resonances corresponds to the acoustics of the radiating vocal tract.

B. Excitation signal and analysis method

A synchronized exponential swept-sine is used (frequency range *0,1-5,5* kHz, duration *1* s, with sinusoidal fade in and fade out). Deconvolution in Fourier domain method detailed in [7] is applied, which enables to properly separate non-linear contributions of the excitation (and measurement) chain from the waveguide linear impulse response. Briefly, each pressure measurement (closed and open mouth conditions) is convoluted in Fourier domain with inverse sweep, then the linear impulse is windowed in time domain. For the cylindrical waveguide in use here, the linear impulse is around *120* ms. A long-enough flat-top window is used to minimize distortion of the impulse response. Finally the corrected frequency response $H(\omega)$ is calculated (Eq. (2)) with a regularized (parameter $\epsilon(\omega)$) spectrum inversion as discussed in [8].

$$H(\omega) = \frac{P_{meas}.P_{cal}^*}{|P_{cal}|^2 + \epsilon(\omega)} \qquad (2)$$

Time-domain signals are sampled at $f_s = 44100$ Hz, Fourier transforms are performed with the power of 2 greater than the length of signal to optimize computation with zero-padding. The final frequency resolution of $H(\omega)$ is less than *1 Hz*. The excitation signal sound pressure level measured with a dBmeter (Nor131) reaches around *86* dB (LAF_{max}) at *9* cm from the source (capillary output and waveguide inlet). All measurements were made in the anechoic room of LMA laboratory.

C. Microphone position testbed

In order to explore the possibility of placing the microphone away from the lips, a microphone (GRAS 40PR/PL with conditioner MMF M108) is horizontally moved away between *0.4* cm and *38* cm from the waveguide inlet. For each microphone position, two records are made in 1) open-mouth condition and 2) closed-mouth condition. Root-Mean Square value of pressure level is computed over short time windows (ten periods long) along the sweep, with respect to the distance. It is compared with a monopolar far-field radiation model (RMS decreasing as the inverse of the distance).

In a second step, a 4-microphones (*M1*, *M2*, *M3*, *M4*) antenna (at distance *d = 0.5, 10, 20* and *30* cm) allows simultaneous measurements (see Figure 2) of the frequency response at the waveguide inlet. Resonance frequencies and quality factors estimated at each microphone distance from the vocal tract inlet (*0.5* cm reference) were compared with each other, and with a transmission line numerical model.

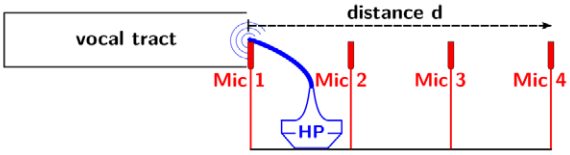


Figure 2 – Four microphones (*M1*, *M2*, *M3*, *M4*) antenna scheme at distances *0.5*, *10*, *20* and *30* cm.

D. Modal estimation

We focus on the resonances (amplitude maxima) of the calculated frequency response $H(\omega)$. The chosen method to determine a resonance frequency and quality factor should be robust to noise and to the proximity of nearby zeros (see Figure 4). Even if frequency is not

biased, quality factor estimation may suffer from the skewness of the amplitude peak and from the phase shift: bandwidth at -3dB of $H(\omega)$ resonances deviates from real bandwidth of vocal tract resonances. First the frequency areas of maxima are delimited (see Figure 3 for the first resonance case). The values of $H(\omega)$ in the considered frequency range are displayed on the complex plane and fitted on a geometric circle (so-called Kennelly circle). Then fitting the curve of angle relative to the circle's center enables the evaluation of the resonance frequency and of its quality factor [9].

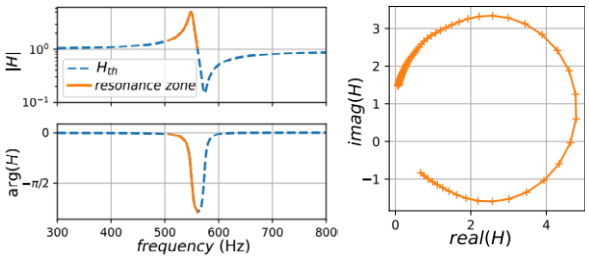


Figure 3 – Definition of first resonance zone and associated Kennelly circle for numerical model.

III. RESULTS

A. Far-field pressure decrease

Figure 4 displays the RMS pressure value (over ten periods) of recorded signals for the mid-frequency *2400* Hz with respect to the distance from the waveguide inlet for open and closed mouth conditions. Plots are normalized by the reference value at the inlet (*0.4* cm). They are compared to the far-field model.

The sound pressure level decreases globally as the inverse of the distance, equivalent to a far-field monopolar radiation model on the considered frequency ranges, as soon as distance is superior to *2 cm* from the inlet. For other studied frequencies, little variation is observed: the decrease in pressure remains proportional to the inverse of distance.

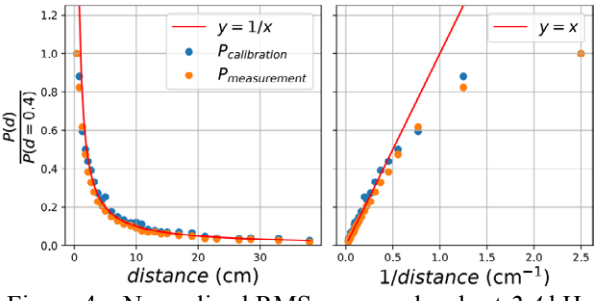


Figure 4 – Normalized RMS pressure levels at *2.4* kHz with distance (left) and with inverse of distance (right).

The decrease behavior is similar between closed and open conditions which indicates that the method can compensate the position of the microphone. Therefore the measurement model (Eq. (1)) remains the same for remote microphones.

B. Far-field modal estimation

Figure 5 displays measurements made with the 4-microphones antenna with distances *0.5 (M1), 10 (M2), 20 (M3)* and *30* cm *(M4)*. Modal estimations (frequencies in Table 1, quality factors in Table 2) obtained from measurements are compared with the one at *0.5* cm and with the frequency response at the waveguide inlet from transmission line model with visco-thermal losses. Model data are computed with the same modal analysis as Sec. II.D.

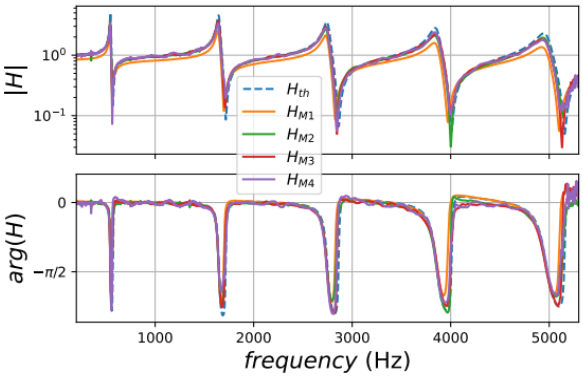


Figure 5 – Example of *H* measurements with distance for 4 microphones and numerical model (dashed-line)

Measurements at different distances are very similar and close to the numerical model. The distance of the microphone and the decrease in pressure is well compensated between calibration and measurement. Noise appear for far-field positions of microphone due to a lower Signal-to-Noise Ratio. The measurement at the lips (*M1*) has a lower amplitude probably due to the bias introduced by the thick cover used for closed mouth condition (*1* mm thick rigid latex cut to the size of the cylindrical waveguide diameter).

***f* (Hz)**	***R1***	***R2***	***R3***	***R4***	***R5***
Model	*546.8*	*1645.4*	*2746.5*	*3851.2*	*4960.7*
M1 (std)	*543.8 ± 0.1*	*1641.3 ± 0.4*	*2736.9 ± 0.6*	*3846.5 ± 1.0*	*4947.1 ± 1.6*
M2 (std)	*543.9 ± 0.1*	*1641.7 ± 0.4*	*2737.4 ± 0.8*	*3845.6 ± 1.6*	*4951.1 ± 2.0*
M3 (std)	*544.0 ± 0.1*	*1642.1 ± 0.5*	*2737.2 ± 1.1*	*3851.7 ± 3.0*	*4947.5 ± 3.3*
M4 (std)	*543.5 ± 0.2*	*1641.2 ± 0.7*	*2735.1 ± 2.0*	*3847.8 ± 3.0*	*4950.4 ± 3.4*

Table 1 – Mean estimations of resonance frequencies (and standard deviation) for *20* measurements (with new calibration for each one) and numerical model.

For all studied distances, resonance frequencies estimates deviate by less than *0.2 %* relatively to the reference measurement at *0.5* cm (*M1*). They are underestimated by less than *0.6 %* relatively to the computational model at the inlet. Standard deviation of frequency estimates slightly increases with the distance (microphone number) and with the frequency (resonance number).

Q	***R1***	***R2***	***R3***	***R4***	***R5***
Model	*59.1*	*55.6*	*45.4*	*39.0*	*35.0*
M1 (std)	*46.6 ± 0.2*	*45.5 ± 1.2*	*39.9 ± 0.4*	*32.5 ± 1.2*	*30.8 ± 0.4*
M2 (std)	*46.9 ± 0.6*	*46.2 ± 1.2*	*39.9 ± 1.2*	*31.8 ± 1.8*	*32.0 ± 1.3*
M3 (std)	*47.0 ± 0.5*	*46.8 ± 1.1*	*39.2 ± 2.1*	*32.4 ± 3.0*	*31.9 ± 3.4*
M4 (std)	*47.1 ± 1.2*	*48.7 ± 0.8*	*38.7 ± 2.4*	*36.1 ± 3.5*	*31.9 ± 4.5*

Table 2 – Mean estimations of resonance quality factors (and standard deviation) for *20* measurements (with new calibration for each one) and model

For all microphones, quality factors estimates deviate by less than *10 %* relatively to the reference measurement (*M1*). They are underestimated compared with results from the model, particularly for the first resonance: measurement of a low-loss waveguide would require a longer excitation time to obtain a better estimate of amplitude and quality factor. This should not be an issue for the estimation of quality factors of real vocal tract whose low-frequency resonances have larger bandwidth [10]. Standard deviation increases with the frequency and the distance.

IV. Conclusion and guidelines

Results evidence that the microphone can be positioned away from the lips. If the study focuses on a horizontal motion of the microphone in order to limit the different radiation patterns of frequencies, the calibration and measurement are always performed at the same microphone position so that radiation pattern is compensated. Therefore the microphone could be placed anywhere. However, the transfer function is more sensitive to ambient noise as the distance of the microphone increases. For accurate measurement, any other acoustic sources than the excitation at the lips should be avoided or limited. Interferences between different sources would appear as noise or peaks on measurements.

A microphone placed at *30* cm distance as recommended by the Union of the European Phoniatricians [11] for stand mounted microphone voice measurements can be used without loss of precision. The device could also be based on head-mounted microphone on which an excitation system could be added. Those possibilities are advantageous for comfort of subjects and to encourage an ecological vocal gesture during studies.

REFERENCES

[1] N. Henrich Bernardoni, J. Smith, and J. Wolfe, “Vocal tract resonances in singing: Strategies used by sopranos, altos, tenors, and baritones”, *J. Acous. Soc. Am.*, vol. 129, no 2, pp. 1024-1035, 2011.

[2] J. Epps, J.R. Smith, and J. Wolfe, “A novel instrument to measure acoustic resonances of the vocal tract during phonation”, *Meas. Sci. and Tech.*, vol. 8, no 10, pp. 1112, 1997.

[3] B. Delvaux, and D. Howard, “Sinesweep-Based Method to Measure the Vocal Tract Resonances”, *Proc. 9th MAVEBA*, September 2-4, 2015.

[4] P. Hoyer, and S. Graf, “Adjustment of the vocal tract shape via biofeedback: a case study”, *J. Voice*, vol. 33, no 4, pp. 482-489, 2019.

[5] M. Kob, and C. Neuschaefer-Rube, “A method for measurement of the vocal tract impedance at the mouth”, *Med. Eng. & Physics*, vol. 24, n 7-8, pp. 467-471, 2002.

[6] T. Maison, F. Silva, Ch. Vergez, and N. Henrich Bernardoni, “Measuring vocal-tract impedance at the lips: model, hypotheses and limits”, *Proc. 12th ICVPB*, Grenoble, France, December 2020.

[7] A. Novak, P. Lotton, and S. Laurent, "Synchronized swept-sine: Theory, application and implementation”, *J. Audio Engineering Society*, vol.63, no 10, pp. 786-798, 2015.

[8] M. Rébillat, R. Hennequin, E. Corteel, and B. Katz, “Identification of cascade of Hammerstein models for the description of non-linearities in vibrating devices”, *J. Sound and Vibration*, vol.330, no 5, pp. 1018-1038, 2010.

[9] A.E. Kennelly, and K. Kurokawa, “Acoustic impedance and its measurement”, *Proc. American Academy of Arts & Sciences*, vol.56, no 1, pp. 3-42, 1921.

[10] N. Hanna, J.R. Smith, and J. Wolfe, “Low frequency response of the vocal tract: acoustic and mechanical resonances and their losses”, *Proc. Australian Acoustical Society*, Fremantle, Australia, pp. 317-323, 2012.

[11] J.G. Svec, and S. Granqvist, “Guidelines for selecting microphones for human voice production research”, *Am. J. Speech-Language Pathology*, vol.19, no 8, pp. 356-369, 2010.

DEVELOPMENT OF AN ACOUSTIC COUGH ANALYSIS METHOD

S. Mootassim-Billah[1], J. Schoentgen[2], D. Van Gestel[3]

1. Department of Radiation Oncology, Speech Therapy Unit, Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium; sofiana.mootassim-billah@bordet.be
2. Department of Biomechatronics, Université Libre de Bruxelles, Brussels, Belgium; jean.schoentgen@ulb.be
3. Department of Radiation Oncology, Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium; dirk.vangestel@bordet.be

Abstract: **Background: The goal of our project is to develop an objective assessment method for dysphagia and aspiration in HNC-patients using acoustic features related to voluntary and/or reflex cough as biomarkers for dysphagia and/or aspiration. This presentation describes the development of an acoustic cough analysis method. The data collected with a free-standing and a skin-contact microphone are compared for a population of healthy subjects.**

Methods: Twenty-one healthy subjects produced five single coughs. A software developed for the purposes of this study enables to analyze cough signals in terms of spectral and temporal features. Cough samples were simultaneously recorded using a free-standing microphone and a skin-contact microphone.

Results: Our study presents the descriptive statistics of spectral and temporal features as well as the correlations observed. Results suggest that the skin-contact microphone under-reports acoustic energy in high-frequency bands, ascribed to turbulence noise.

Keywords: **cough, acoustic analysis, free-air microphone, skin-contact microphone.**

I. INTRODUCTION

Late radiation associated dysphagia (RAD) is defined as impaired swallowing efficiency and/or safety following (chemo)radiotherapy in head and neck cancer (HNC) patients [1]. The two hallmarks of RAD are residue (food sticking in the mouth/throat) and aspiration (food entering the airways). The latter ideally results in a cough reflex protecting the airways and lungs. In HNC-patients, the efficacy of this elicited cough is often diminished due to changes in the local physiology (sensory deterioration). As such, up to 83% of HNC-patients are at risk of lung aspirations and consequent lung infections, fatal for one third of them [2]. Although cough efficacy is considered as a reliable predictor of aspiration in the framework of dysphagia, cough investigation has been minimal in patients with RAD.

An acoustic cough emission is usually defined as a transient signal comprising three sequential phases: a burst/release, followed by a "fricated" fragment (turbulence noise) and a "voiced" fragment (oscillations) [3]. This academic view is inspired by the analogy between a cough sound and a glottal stop – a consonantal sound used in many spoken languages, produced by obstructing airflow in the vocal tract or, more precisely, the glottis.

Because of the transient attributes of the cough signal, conventional software for voice and speech analysis are not appropriate. Indeed, the assessment of voice quality is based on sustained voiced speech sounds, selected for reasons of technical feasibility and ease of reproducibility of the analysis.

Also, voice and speech samples are usually recorded with validated professional acoustic microphones (free-standing microphones or head-mounted microphones). In practice, however, one observes that head-mounted microphones are not suited for recording cough signals. The intensity of cough signals may be so high that the transducer and/or pre-amplifier saturate. For this reason, freely placeable microphones are more appropriate than head-mounted microphones. In addition, a skin-contact microphone is the most suitable sensor for recording elicited cough sounds because a facemask is necessary for tussigen nebulization.

A free-standing microphone as well as a skin-contact microphone have therefore been used to record voluntary or elicited cough sounds in the framework of our study. The recordings obtained with these transducers are idiosyncratic and non-interchangeable [4].Consequently, it is necessary to take into account the specificities of each of these transducers with regard to the application.

The acoustic microphone is placed at a fixed distance from the mouth of the subject and is protected by a metallic anti-pop screen. The skin-contact microphone is attached to the throat skin and records laryngeal vibrations directly.

The sound recorded by a skin-contact microphone is muffled because it transits through the tissue of the neck and the resonances of the vocal tract do not fully contribute to the timbre [5]. The skin-contact

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

microphone has other disadvantages. Its position and orientation may influence the recording of the signal [6]. Moreover, the recorded signal may be impacted by the skin properties or by extra-sounds owing to blood-flow and heartbeat as well as by tissue or muscle movement when the speaker is in motion [5].

The overall goal of our project is to develop an objective assessment method for dysphagia and aspiration in HNC-patients using acoustic features related to voluntary and/or reflex cough as biomarkers for dysphagia and/or aspiration. This presentation describes the acoustic cough analysis methods developed for this research. The data collected with a free-standing and a skin-contact microphone are compared for a corpus of healthy subjects.

II. METHODS

A. Corpus

Twenty-one healthy individuals, including 13 women and 8 men, participated in this study. The average age of the participants was 33.1 ± 5.09 years (range, 24 to 53 years). Exclusion criteria were 1) history of head and neck cancer; 2) dysphagia; 3) dysphonia (G > 0 on GRBAS-I scale); 3) history of smoking for within less than one year; 4) significant or chronic respiratory disease or illness.

B. Recordings

Participants were seated in an audiometric booth. The recordings were simultaneously collected using a skin-contact microphone and a professional quality acoustic free-standing microphone. The skin-contact microphone was the Albrecht AE 38 S2, valid and reliable for recordings in a natural setting [7]. The free-standing microphone was the AKG Perception 420 Omnidirectionnal, fixed to a flex arm fastened to a table facing the participants. A metallic anti-pop filter was placed in front of the microphone to prevent the exhaled air hitting the microphone, but also for easy disinfection with wipes. Intensity (in dB) was measured with an external sound level meter Bruel & Kjaer 2236 placed at 40 cm on the right of the mouth of the participant, also for reasons of hygiene.

All participants produced 5 voluntary coughs. Each participant was verbally instructed as follows “Take a maximal breath and cough as if you have something stuck in your throat”.

As recommended by Union of European Phoniatricians’ guidelines, investigators wore a protective visor for face and eye protection, a surgical facemask, a single use protective gown and a single use cap. A time interval of ten minutes between participants was scheduled for purifying and sterilizing the room with a Hextio Radic8 device and for cleaning all surfaces and equipment.

Cough samples were recorded with an HP ProBook computer (Hew1ett-Packard Company, USA) using the computer program PRAAT and the preamplifier 2 channel interface Presonus Audiobox USB 96 Audio, with a sampling frequency of 44.1 kHz.

Cough samples were analyzed with a software developed for the purposes of this study.

C. Segmentation

Cough bouts are segmented by hand into single coughs leaving silent intervals before and after. The segregation of a single cough from its preceding and succeeding silent intervals by hand would be difficult because the offset of a single cough is drawn out without a well-defined boundary. Segregation from silence is therefore carried out automatically via the signal contour by assigning to the onset the first contour sample and to the offset the last contour sample the value of which is larger than -20 dB with regard to the signal contour maximum.

For spectral analyses, the signal contour is estimated by smoothing the absolute signal samples via a rectangular window the length of which equals the sampling frequency in Hz divided by a cutoff frequency equal to 50 Hz.

Before analysis, the segmented cough signals are normalized so that the maximum of the absolute value is equal to one.

D. Spectral analysis

Cough signals are transient signals, which are therefore unsatisfactorily represented by spectrograms. We have therefore focused on a smaller number of frequency intervals, the energies and frequencies of which are reported band by band.

The signals are broken up into constituent signals via a filter bank that is based on the discrete cosine transform (DCT). The frequency boundaries are equal to 400 Hz, 800 Hz, 1600 Hz and 3200 Hz. The difference between the discrete cosine and discrete Fourier transforms is that the former periodically extends the analyzed signal by pivoting the signal with regard to its onset and offset so that the periodically extended signal is even. The juxtaposition of a slow and low-amplitude offset with a rapid and high-amplitude onset is so avoided, as well as the ensuing spectral artefacts. The decomposition of the cough signal by means of a DCT is exact, that is, the sum of the band-filtered signals as well as their signal energies is equal to the original cough signal and its energy [8].

The spectral features are the relative signal energies in the bands (0Hz, 400Hz), (400Hz, 800Hz), (800Hz, 1600Hz), (1600 Hz, 3200Hz) as well as in the interval between 3200 Hz and half the sampling

frequency. The average frequency in each band is estimated via the number of zero-crossings. The per-band frequencies are weighted by the relative band energies and summed. The weighted sum is an approximation of the spectral centroid that subdivides the signal spectrum into two halves that have equal energies.

E. Temporal analysis

The temporal analysis involves the evolution with time of the cough signal amplitude, the sample entropy as well as the kurtosis. Each of these quantities is obtained once per analysis frame. The frame length is equal to 30 ms and the hop ratio is equal to 0.5. The frame-wise calculated quantities are then interpolated to obtain the contour of each quantity sample-by-sample.

The amplitude reports the strength of the cough transient. The amplitude is estimated by taking the square root of the sum of the squared samples divided by the number of samples in the analysis frame.

The sample entropy reports the degree of randomness in an analysis frame. The samples are z-normalized before the entropy is calculated by comparing the distance between all sample pairs and all sample triplets that is smaller than a threshold. The threshold is equal to 0.2 times the standard deviation. The sample entropy segregates analysis frames according to whether they report turbulence noise or locally-periodic oscillations because turbulence noise is expected to be less predictable than locally-periodic oscillations or a mix thereof [9].

The kurtosis reports the impulsive quality of the signal samples within an analysis frame. It involves the fourth moment of the samples divided by the square of the second moment. The kurtosis may be interpreted in terms of the peakedness of the histogram of the sample values. Sample histograms that are between normal and uniform have kurtosis values between three and zero. Histograms the peakedness of which is stronger than normal have kurtosis values larger than three. Burst-like onsets are therefore expected to have larger kurtosis values than turbulence noise or oscillations [10].

The shape of the contours of the cough amplitude, sample entropy and kurtosis is described by means of the first three DCT coefficients. Inspecting the pattern of the first three co-sinusoidal basis functions shows that the first coefficient is the contour average. The second coefficient describes the contour trend. A positive coefficient value indicates a trend that is decreasing with time. The third coefficient reports the contour curvature. A positive coefficient value indicates a downward-upward (convex) curvature and negative values an upward-downward (concave) curvature of the contour with regard to the horizontal.

III. RESULTS

A. Descriptive statistics

The medians of the acoustic features collected with the free-standing (Fmic) and skin-contact microphones (SCmic) and a comparison of the medians by means of a non-parametric Wilcoxon test are reported in Tables 1 and 2. The average intensity in dB of the 5 times 21 cough signals was 97.05 ± 4.69 (median = 97.4).

Table 1: Median cough length, median relative energy in each frequency band, median spectral centroid and the statistical significance of the difference between free-standing and skin-contact microphone.

	Fmic	SCmic	Wilcoxon tests
Length (sec)	0.387	0.335	$p<0.05$
< 400 Hz	0.495	0.488	$p=0.61$
400-800Hz	0.103	0.414	$p<0.05$
800-1600Hz	0.095	0.071	$p<0.05$
1600-3200Hz	0.136	0.007	$p<0.05$
>3200Hz	0.084	0.000	$p<0.05$
Weighted freq. (Hz)	1245	477	$p<0.05$

Table 2: Median length, average, trend and curvature of the amplitude, sample entropy and kurtosis contours and the statistical significance of the difference of the medians between free-standing and skin-contact microphone.

	Fmic	SCmic	Wilcoxon tests
Length (sec)	0.907	0.792	$p<0.05$
Amplitude			
Average	0.109	0.117	$p<0.05$
Trend	0.034	0.026	$p<0.05$
Curvature	0.000	0.019	$p<0.05$
Sample entropy			
Average	0.584	0.290	$p<0.05$
Trend	0.117	0.005	$p<0.05$
Curvature	-0.107	-0.060	$p<0.05$
Kurtosis			
Average	3.451	3.840	$p<0.05$
Trend	0.562	0.474	$p=0.121$
Curvature	0.487	0.661	$p<0.05$

B. Correlations

Spearman correlation coefficients were calculated between all temporal and spectral features, considering both transducers separately.

Fmic correlations: Significant correlations were found between the average amplitude contour and the relative

energy < 400 Hz (0.3), the average sample entropy (-0.6), and the spectral centroid (-0.4).

SCmic correlations: the same significant correlations were found (but to a different degree) between the average amplitude contour and the relative energy < 400 Hz (0.2), the average sample entropy (-0.3) and the spectral centroid (-0.2).

IV. Discussion

The spectral analysis shows that the energy distribution between frequency bands differs for the free-air and skin-contact microphones. The latter reports less energy in frequency bands > 800 Hz than the former. This is confirmed by the spectral centroid, which is significantly lower for cough samples collected with the skin-contact microphone. A possible explanation involves the acoustic radiation characteristics through the tissues at the neck as well as the attenuation of the acoustic propagation through that tissue, which weaken the contribution of high-frequency turbulence noise to cough signals recorded via a skin-contact microphone.

Similarly, the temporal analysis shows significant differences between the average sample entropies and their trends reported by free-standing and skin-contact microphones. The latter report lower average entropy values and flatter entropy trends than the former. Knowing that turbulence noise boosts entropy values compared to locally-periodic vibrations and that skin-contact microphones de-emphasize the spectral energy in high-frequency bands, one may therefore conclude that a skin-contact microphone under-reports the acoustic energy involved in turbulence noise compared to a free-air microphone.

One other issue that is likely to influence temporal features that report the shapes of the amplitude, entropy and kurtosis contours of a single cough is segmentation. Consistent segmentation is a delicate task because single coughs lack a well-defined offset. Minor changes in the segmentation criteria, in the estimation of the amplitude contour or the use of distinct transducers are therefore likely to cause the segmented cough lengths to differ, which may result in differences in the values of the temporal features.

Finally, the statistically significant positive correlation between the average amplitude contour and the relative spectral energy < 400 Hz, as well as the negative correlation between the average amplitude contour and the average entropy contour as well as spectral centroid suggest that the overall size of the amplitude contour of the normalized cough signal mainly co-evolves with the low-frequency locally-periodic cough signal oscillations, which are larger than the fricative signal fragments.

V. Conclusion

The study presents the development of an acoustic cough analysis method. Here, it is used to compare the features of single coughs recorded by a free-standing and a skin-contact microphone. Results suggest that the observed differences are attributable to the under-emphasis by the skin-contact microphone of high-frequency bands, which therefore under-reports the acoustic energy ascribed to turbulence noise.

References

[1] M.J. Awad, A.S.R. Mohamed, J.S. Lewin, et al, “Late Radiation-Associated Dysphagia (Late-RAD) with Lower Cranial Neuropathy after Oropharyngeal Radiotherapy: A Preliminary Dosimetric Comparison.” *Oral Oncol*, vol. 50(8), pp746-754, 2014.

[2] N. Rogus-Pulia, M.C. Pierce, B.B. Mittal et al, “Changes in swallowing physiology and patient perception of swallowing function following Chemoradiation for head and neck cancer”, *Dysphagia*, vol. 29, pp. 223-233, 2014.

[3] G.A. Fontana & J. Widdicombe, “What is cough and what should be measured?”, *Pulmonary Pharmacology & Therapeutics,* vol. 20, pp. 307-312, 2007.

[4] F. Movahedi, A. Kurosu, J.L. Coyle et al, “A comparison between swallowing sounds and vibrations in patients with dysphagia”, *Computer Methods and Programs in Biomedicine*, vol. 144, pp. 179-187, 2017.

[5] F. Scherbaum, S. Rosenzweig, M. Müller et al, « Throat microphones for vocal music analysis », *in Demos and Late Breaking News of the International Society for Music Information Retrieval Society Conference*, 2018.

[6] E. Sejdic, J. Dudik, A Kurosu et al, “Understanding differences between healthy swallows and penetration-aspiration swallows via compressive sensing of tri-axial swallowing accelerometry signals”, *in Proc SPIE Compressive Sens III*, 2014.

[7] F. Scherbaum, N. Mzhavanadze, S. Rosenzweig et al, “Multi-media recordings of traditional Georgian vocal music for computational analysis”, *in 9th International Workshop on Folk Music Analysis, Birmingham,* 2019.

[8] S.J. Orfanidis, *Introduction to signal processing*, Prentice Hall, N. J., 1996, p. 472

[9] A. Delgado-Bonal & A. Marshak, “Approximate Entropy and Sample Entropy: A Comprehensive Tutorial”, *Entropy*, vol 21(6), 2019, 541; doi:10.3390/e21060541

[10] W. Qiu, W.J. Murphy & A. Suter, “Kurtosis: a new tool for noise analysis”, *Acoustics Today*, 16(4), pp. 39-47, 2020.

LUNG VOLUME AND VOICE EFFICIENCY

P. H. DeJonckere[1], J. Lebacq [2]

[1] Federal Agency for Occupational Risks, Brussels, Belgium
ph.dejonckere@outlook.com
[2] University of Louvain, Neurosciences, Brussels, Belgium
Jean.lebacq@uclouvain.be

Abstract: **At the end of a vocal emission, when the voicing is not interrupted by a laryngeal closure, a damped oscillatory motion of each vocal fold can be observed after the last contact phase of the two fold edges on the midline. It can be precisely analyzed using a photometric method. Actually, during modal phonation, the vocal oscillator mainly comprises two components: the vocal folds themselves and the vibrating air mass. In order to investigate the effect of the vibrating air mass, a voicing protocol was elaborated for validly measuring and comparing damping characteristics in two conditions: at high and at low lung volume, ceteris paribus. Glottal area, intraoral pressure, EGG and sound were recorded simultaneously. The results show that the decay of vocal fold oscillation is influenced by the amount of lung air that is set into oscillation. A reduction of the air volume leads to a significant increase in the rate of decay, thus voicing at low lung volume requires more energy, which is of importance for voice hygiene.**

Keywords: **Lung volume, damping, vocal folds, photoglottography. fundamental frequency.**

I. INTRODUCTION

At the end of a vocal emission, a damped oscillatory movement on each vocal fold (VF) can be observed after the last contact phase of the two fold edges on the midline [1]. The amplitude decrement from cycle to cycle reflects the energy input requested to maintain a steady state oscillation. A fast repetition (3 to 4 s^{-1}) of a vowel followed by an abrupt bilabial occlusion (e.g. /ερερερερερερ/) at comfortable pitch and loudness is a convenient protocol for analyzing this. The oscillating system itself consists in two components: the two VFs and the air mass of lower and upper airways. The size of the vibrating mass of the VFs tissue can be roughly estimated on the basis of MR-imaging. Thickness and width of each vibrating fold can be estimated to 4 and 5 mm respectively. The vibrating length, as seen on videolaryngoscopic images, is around 16 mm (male subject, modal register, comfortable pitch and loudness). So 0,5 g is a reasonable upper limit estimate of the total mass of vibrating tissue *in vivo* (2 VF). In a female subject, one may expect 0.35 g. A rough assumption is that modal speech occurs with an average lung volume slightly above the upper limit of the tidal volume. Hence the internal air volume set into vibration consists in about 50% of the vital capacity (i.e. a half of 3000 – 4500 ml), to which has to be added a probably large part of the residual volume (on average 1,1 - 1,2 l) and the supraglottal vocal tract (around 75 ml). Globally, the weight of the vibrating air can be estimated to around 2,7 to 3,7 g (1,14 g / l), clearly larger than even the high estimate of the VF mass. Varying the air volume set into vibration would allow checking its importance for the damping characteristics. This is possible by comparing two conditions: voicing with respectively high and low lung volume while the above-mentioned protocol is applied. Our hypothesis is that an increase of the air volume (of about 2,5 l) put into vibration by the VFs should improve the mechanical quality of the global oscillating system, which should be reflected in a lower damping when the driving force is abruptly suppressed.

II. METHODS

The glottal area was derived from a photometric record obtained by transilluminating the trachea. The light source for this transillumination was a tungsten filament light bulb driven by a constant ripple-free current source. The light flux was detected by a photovoltaic transducer positioned as dorsally as possible in the pharynx (photoglottography) [2]. The light signal is the most important one, as it serves to compute the damping. The calibration procedure has been described previously [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 amplitude of the light signal was normalized and expressed - in the damping phase - as a fraction of the amplitude of the first 'free oscillation' after the last closed plateau. The intra-oral pressure was measured by means of a Millar Mikro-Tip catheter (Model SPC-751, Millar

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

Instruments, Inc. Houston, USA). The pressure signal allows to precisely identify the moment of lip opening (pressure drop). When the lips are closing, the intraoral pressure increases up to nearly the level of the lung pressure, which remains approximately constant. The electroglottographic (EGG) signal, used as a reference for monitoring the changes in contact surface of the VF, was detected using a portable electroglottograph (Laryngograph Ltd, London, UK) Model EG90. The EGG-signal however fails to show the final phase of the damping, since there is no contact between the VF during this phase. The last sinusoidal EGG-cycles probably correspond to small (reduced amplitude) impedance fluctuations at the level of the ventral commissure. The start of free oscillations of the VF is indicated by a strong reduction in amplitude of the EGG-signal. Sounds were detected by a Sennheiser MD 421 U microphone at 10 cm of the mouth.

All signals were recorded by means of a 4-channels Pico Scope 3403D module (Pico Technology Ltd, St Neots, England, UK) driven by the PicoScope 6 programme, and stored in a computer.

The subject was a healthy trained male vocalist, experienced in controlling voicing parameters [1,3]. During three sessions, a total of 227 recordings of series of short repetitive vocal /pɛp/ emissions were achieved with the photoglottograph and pressure sensor in situ. The vocalist made series of fast repetitions (3 to 4 s^{-1}) of the vowel /ɛ/ (determined by mechanical constraints of the experimental procedure) , each vocalization being followed by an abrupt bilabial occlusion (/ɛpɛpɛpɛpɛpɛp/) at comfortable pitch and loudness (105 – 130 Hz, corresponding to the average speaking frequency of the subject, and 63 – 68 dB_A at 10 cm of the lips).

These series of fast repetitions of the vowel /ɛ/ were carried out in two lung volume conditions: high and low lung volume. Fig. 1 shows a spirographic diagram with personalized values for the subject, showing the traditional lung volume compartments, and the situation of the two zones (of 500 ml each) in which the sequences of interrupted vocalizations were produced and recorded. The two zones correspond to the 'high' and 'low' lung volume conditions respectively. The difference in lung volume between the two zones is approximately 2410 ml.

A total corpus of 105 selected polygraphic recordings corresponding to the condition 'high lung volume' (54) and to the condition 'low lung volume' (51) was created. An example is given in Fig. 2.

Counting the number of free oscillations on the glottal area trace started just after the last closed plateau. However, identifying this last closed plateau requires expanding the time scale.

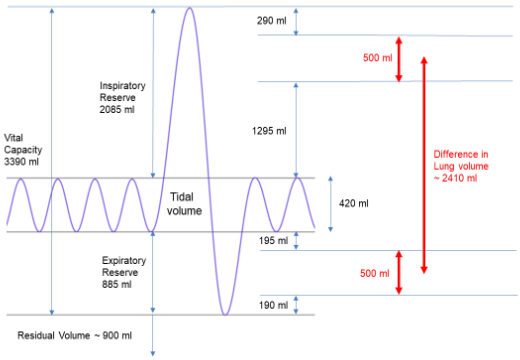


Fig. 1: Spirographic diagram, with personalized values, showing the traditional lung volume compartments, and the situation of the two zones (of 500 ml each) wherein the sequences of interrupted vocalizations were produced and recorded.

Counting was made blindly, i.e. the rater being unaware of the condition (high or low lung volume).

Measurement of amplitude decay was done by first identifying - after strong enlargement of the Pico picture (vertical expansion) - the successive maximum and minimum of each cycle.

III. RESULTS

Fig. 2 shows a global view of a polygraphic recording of a single vocalization /pɛp/ in the 'high lung volume' condition. The /pɛp/ is extracted from a /ɛpɛpɛpɛpɛp…/ sequence at a rhythm of three to four vocalizations per s. The vowel / ɛ / is determined by the constraints of the oral and pharyngeal sensors. F_0 is around 130 Hz and intensity around 64 dB (at 10 cm). Subglottal pressure (estimate) is 4.9 hPa.

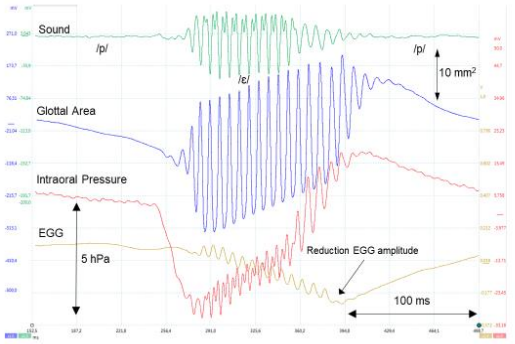


Fig. 2: Global view of a polygraphic recording of a single vocalization /pɛp/ in the 'high lung volume'

condition. The /pɛp/ is extracted from a /ɛpɛpɛpɛpɛp…/ sequence at a rhythm of three to four vocalizations per s. Fo is around 130 Hz and intensity around 64 dB (10 cm). Subglottal pressure (estimated) is 4.9 hPa.

Fig. 3 is focusing on the voicing offset in an example of the 'high lung volume' condition: on the glottal area trace, seven free oscillations can be identified after the last closed plateau.

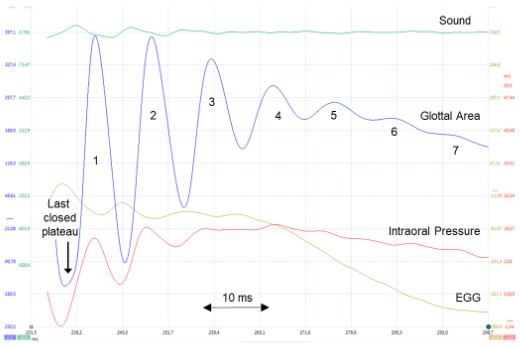


Fig. 3: Example of a voicing offset in the 'high lung volume' condition. Seven free oscillations can be identified on the glottal area trace after the last closed plateau.

Average counts (blinded for condition) of the numbers of 'free oscillation' cycles after the last VF contact, were 4,89 +/- 0,79 in the 'high lung volume' condition and 3,65 ± 0,72 in the 'low lung volume' condition. The difference is highly significant ($p < 0.0001$). This is confirmed by the superimposed histograms with Gaussian fits (Fig. 4).

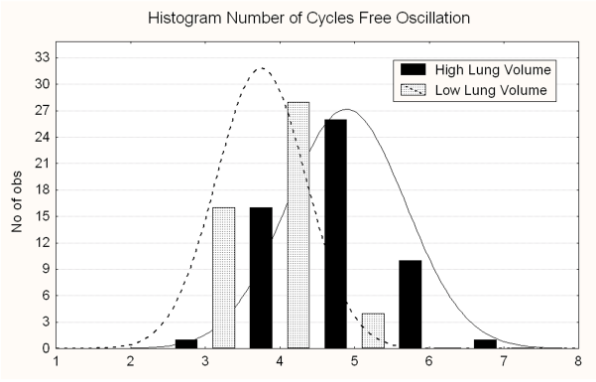


Fig. 4: Histogram of the number of cycles (free oscillations) that can be identified after the last closed plateau (= last contact between vocal fold edges on the midline). Gaussian fits. N = 54 and 51. The average number of cycles is highly significantly lower in the case of low lung volume ($p < .0001$).

The cycle by cycle decay of the normalized amplitude after the last closed plateau for the two conditions is shown in Fig. 5. Cycle # 1 is the first free oscillation, defining 100% amplitude. The decay is stronger and faster in the 'low lung volume' condition. The difference in amplitude mainly appears in cycles #2 and #3. In cycle #4, the difference is smaller although still just significant, but there are only a few cases for the 'low lung volume' condition. For cycles #6 and #7, there are only data for the 'high lung volume' condition.

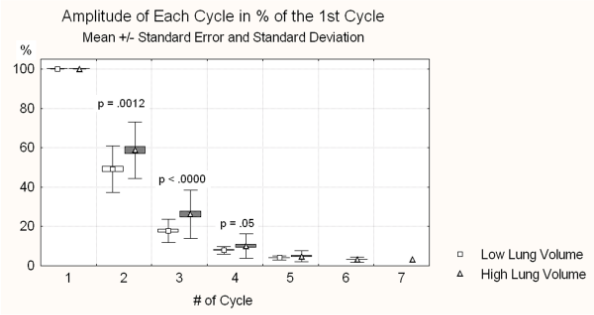


Fig. 5: Comparison of amplitudes between the 'high lung volume' and the 'low lung volume' conditions for each successive free oscillation. The amplitude of the first identifiable free oscillation is set at 100 % in both the 'high lung volume' and in the 'low lung volume' condition (normalization). The difference in amplitude mainly appears in cycles #2 and #3. In cycle #4 the difference is still just significant, but there are only a few cases for the 'low lung volume' condition. For cycles 6 and 7, there are only data for the 'high lung volume' condition.

The logarithmic decrement is defined as the natural log of the ratio of the amplitudes of any two successive positive peaks: ($\ln [x_n / x_{n+1}]$). The global average logarithmic decrement is 0,72 +/- 0,31 in the 'high lung volume' condition (n = 212 logarithmic decrements) and 0,88 +/- 0,26 in the 'low lung volume' condition (n = 133 logarithmic decrements). This difference is highly significant ($p < .001$).

IV. Discussion

In phonation physiology, the concept of 'vocal oscillator' may obviously not be limited to the VF, but it includes internal air volume set into motion by the lung pressure and into vibration by the VFs. The mass of the air appears to be around sevenfold that of the VF tissue. During speech and singing, after the subject has taken a small or a larger deep breath, this volume

progressively declines, and this in turn influences the physical properties and the energy required for of the voice production. At high lung volume, the elastic effect of a larger vibrating air mass reduces the rate of decay of the glottal oscillations.

In our experiments, the calculated global average logarithmic decrement is 0,72 +/- 0,31 in the 'high lung volume' condition and 0,88 +/- 0,26 in the 'low lung volume' condition. For comparison, the logarithmic decrement computed on a graph made by Tanabe and Isshiki [4] and based on high speed cinematography of an autopsy larynx is clearly higher: 1,65 (oscillation stops after 2 cycles).

As to clinical significance, Lowell & al. [5] compared - during teaching-related speaking tasks - teachers with voice problems (in the absence of laryngeal lesions) with 'healthy' teachers, and observed decreased levels of lung volume initiation and termination in the former with respect to the latter. Actually, teachers frequently have to speak at increased loudness levels while teaching. At higher lung volume initiation levels, greater respiratory recoil forces are available for expiratory speech. By starting their breath groups at higher levels, teachers with healthy voices capitalize on these passive recoil forces. Initiating breath groups at a higher volume facilitates an increased lung pressure and consequently a louder voice. Also, by ending their breath groups at higher levels, they avoid the muscle effort required for producing speech below the resting respiratory level.

Similarly, Schaeffer & al. [6] compared patients with abuse-related dysphonia with a normal control group in a reading task of a 60-syllable paragraph: significant results indicated that the end-expiratory lung volume levels of the dysphonic group were further below the resting expiratory level than those of the control group. In a later study, Schaeffer [7] showed that a significant improvement in speech breathing data (higher end-expiratory levels) could be obtained by voice therapy, with a reduction of perceived dysphonia. The average termination of speech relative to the resting respiratory level was – 0.224 l before therapy and + 0.063 l after therapy.

Along the same line, Iwarsson & Sundberg [8], using respiratory inductive plethysmography, investigated female voice patients with vocal fold nodules, They concluded that "females with vocal nodules were shown to inhale more often, and, when shouting, initiated phrases at lower lung volume levels than females without nodules, thus refraining from taking advantage of the increased recoil contributions to subglottal pressure associated with high lung volumes."

Our experiments point to an additional mechanism to this physiological rationale: speech at low lung volume requires significantly more energy for voicing due to the enhanced damping of the oscillating system.

V. Conclusion

With an adequate methodology, it is possible to control, to standardize and to quantify the damping characteristics of the oscillating system (vocal fold tissue and air mass) during a physiological voicing offset with abrupt interruption of the airflow. This allows investigating specifically the role of lung volume. The mechanical quality of the oscillating system appears to be, to a non-negligible extent, determined by the lung volume that is set into oscillation; a reduction of the air volume leads to a significant increase in the rate of decay of oscillations, resulting in a higher energy demand for voicing.

REFERENCES

[1] DeJonckere PH, Lebacq J. Damping of vocal fold oscillation at voice offset. *Biomed Signal Process Control.* 2017; 37: 92–99.
[2] DeJonckere PH, Lebacq J, Titze IR. Dynamics of the driving force during the normal vocal fold vibration cycle. *J Voice*. 2017; 31: 649–661.
[3] DeJonckere PH, Lebacq J. In vivo quantification of the intraglottal pressure: Modal phonation and voice onset. *J.Voice* 2019.
DOI: https://doi.org/10.1016/j.jvoice.2019.01.001
[4] Tanabe M, Isshiki N. Rheological characteristics of the vocal cord. *Studia Phonologica Kyoto* 1979; 13: 18 – 22.
[5] Lowell SY, Brakmeyer-Kraemer JM, Hoit, JD, Story BH. Respiratory and laryngeal function during spontaneous speaking in teachers with voice disorders. *J Speech Lang Hear Res.* 2008; 51: 333 – 349.
[6] Schaeffer N, Cavallo SA, Wall M, Diakow C. Speech breathing behavior in normal and moderately to severely dysphonic subjects during connected speech. *J Med Speech Lang Pathol*. 2002; 10: 1–19.
[7] Schaeffer N. Speech breathing behavior and vocal fold function in dysphonic participants before and after therapy during connected speech: Preliminary observations. *Contemporary Issues in Communication Science and Disorders* 2007; 34: 61-72.
[8] Iwarsson J, Sundberg J. Breathing behaviors during speech in healthy females and patients with vocal fold nodules. *Logopedics Phoniatrics Vocology* 1999; 24: 154 – 169.

F0 ESTIMATION IN IRREGULAR VOCAL EMISSIONS USING RIDGE DETECTION METHODS.

A. Gomez[1], A. Tsanas [1]
[1] Usher Institute, Edinburgh Medical School, University of Edinburgh, Edinburgh, UK;
a.gomezrodellar@ed.ac.uk,
Athanasios.Tsanas @ed.ac.uk

Abstract: **The estimation of the fundamental frequency (F0) is one of the most important problems on characterizing (quasi)-periodic signals, and is particularly relevant in speech signal analysis. Although the concept theoretically is straightforward, it becomes complicated in practical scenarios. Most F0 estimation algorithms make implicit assumptions about the underlying stationary properties which may not apply in naturally produced signals, and exploring these irregularities is particularly important when processing pathological voices. This study explores the F0 estimation using artificially generated |a| vowels by employing exploratory functions (kernels) to analyze the repetitive structure found in the Auto Correlation Function (ACF). The F0 contour is then extracted by applying ridge detection techniques. The results are defined using; Mean Absolute Error (MAE) 2.79 ± 5.72 and Root Mean Squared Error 4.003 ± 7.91.**

Keywords: **F0 estimation, ridge detection.**

I. INTRODUCTION

Fundamental frequency F0 estimation of a speech voiced signal is an important task in Speech Sciences, as many specific features of voiced speech rely on its accurate estimation [1]. These algorithms are typically based on three major components of an F0 estimator: a signal conditioning stage, a generator of candidate estimates of the true period sought, and a post-processing stage to select the best candidate of the estimation, given a specific criterion (often, there is some sort of smoothing embedded within this stage). The estimation methods are classified as time-domain (classically based on autocorrelation) and frequency-domain (spectral or cepstral approaches). Along these years, multiple algorithms that have been proposed [2]

The simplest way of defining F0 in a strictly periodic signal is to identify a time structure that repeats in time (i.e. the period of a signal), select a characteristic of the repeating pattern, and measure the minimum time delay between the two points where it repeats. The problem with defining the presence of a periodic signal is that the periodic pattern must be sustained for a required duration of time. In periodic signals this pattern would stretch infinitely in time, whereas in real signals periodicity has to be observed for a minimum duration. This causes a compromise between time resolution and frequency precision [3]. If the window is too small, the F0 estimation might not extend long enough and if its too large the variability among events is lost and the frequency estimation does not become responsive enough. When applying this definition to speech it has to be taken into consideration accordingly to the underlying biological phenomena that amount to its generation. As the energy stored in the lungs is released through the vocal folds they vibrate generating a pattern, when they are open, air escapes and when they are closed the airflow is cut and pressure builds up inside the lungs and drops down in the larynx, producing a glottal pulse. This pattern is released through the oronasal pharyngeal cavities, that act as a filter shaping the signal into what is recognizable as voice [1]. With this in mind the definition of the F0 becomes quite clear. The F0 pattern is subject to physical laws (inertia and elasticity), meaning it pattern cannot abruptly change under normal oscillatory conditions, (i.e. if there is no damage to the vocal folds or any other pathology [4]). Defining the fundamental frequency in a low quality or pathologic signal is a tricky task as the definition has to encompass the variability that manifests in the irregular production. Such an environment where the quality of the recordings may not be met is telephonic audio, where the devices are not standardized, there are external interferences, and the recording instructions are not fully understood or followed by participants in crowdsourcing database generation.

The aim of this study is to present a new F0 estimation algorithm extending the use of standard time-series approaches based on the Auto Correlation Function (ACF) towards determining more accurately the underlying short time variability of the vocal fold vibrations and hence robustly estimating F0.

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

II. METHODS

The proposed approach is composed of two steps; Kernel decomposition and Ridge detection. The estimation is based on the exploration of the Auto Correlation Function (ACF), which is a well-established methodology [5]. If a repetitive pattern is present, the ACF shows a repetitive oscillation that allows to identify the F0 from the delay between successive peaks. The characteristics of this pattern are related to a single F0, so a logical step would be to compare the ACF with a set of predefined functions (kernels) that are constructed to capture specific delays in the signal under test. The ridge estimation is based on the quasi-stationarity principle of the F0 to make a first proposal. This is based on the knowledge that abrupt changes of the F0 estimation are unlikely, considering small time observation segments; this means that the true frequency can deviate only within a finite range. Therefore the actual F0 value corresponds to the ridge traced by the kernel decomposition matrix.

The estimation begins with the well-established method using the ACF;

$$R_{ff} = \sum_{k=1}^{N} x[n] * \bar{x}[n-k] \qquad (1)$$

with x being the function under exploration and $\bar{x}$ the conjugate function delayed k samples. The speech signal is split into overlapping windows of 60 ms with a stride of 10 ms, resulting in M windows of N_w samples e.g. for a 1 second signal sampled at 8 KHz we would have M=100 and N_w =2646 samples. For each of these elements the autocorrelation is extracted and then stored as the rows in the autocorrelation matrix ($\mathbf{ACFM} \epsilon \mathbb{R}^{M x N_w}$). The traditional way of estimating F0 using the ACF is to calculate the time delay of the second maximum of the auto-correlation function with respect to the maximally aligned auto-correlation, this being the Normalized Correlation Function (NCF)[4].

This method works sufficiently well for most signals with no imperfections, but as signals increase in complexity and randomness, the definition of second maxima is not such a clear cut, yielding imprecisions on the estimation. An example of this situation is the doubling frequency effect due to the insertion of higher order vibration modes or closure defects in the vocal folds. As the speaker is phonating, the vocal folds might close irregularly letting airflow escape, creating a pattern in the middle of the glottal signal that causes the ACF to present secondary spurious maxima, which can lead the estimator to wrongly asses the signal to be of double the frequency. The proposed method expands on this methodology making two assumptions; (1), it assumes that there exists a single F0 to begin with (the estimation always picks one of the exploratory kernels) and secondly, it assumes that the fundamental frequency cannot change abruptly, i.e. there is a frequential range where the possible candidates exist.

Kernels are a set of predefined functions that work as benchmarks upon which to compare a test signal. The properties of the original signal are then inferred from the kernel it resembles the most. The structure of the kernel allows for a degree of freedom to explore the properties of the signal under test, as the functions can be designed with the idea that they explore a set of characteristics of interest. The proposed setting uses vectorial kernels to test the ACF to make an estimation of the F0. The simplest kernel function for this task is a train of pulses (Kronecker delta functions) with a fixed delay between successive peaks; this signal is composed of a step value (for simplicity taken as the unity), each one separated by a fixed quantity of zeros as shown in the diagram in Fig. 1.

$$k_i = \begin{cases} \delta[n-(i+1)\tau+1];\ \tau=(k-1);\ k \epsilon N \\ 0\ otherwise \end{cases} \qquad (2)$$

where δ is the Kronecker delta and i denotes the i-th kernel.

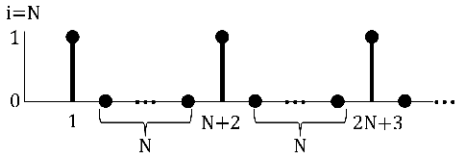


Fig. 1) Representation of the N-th kernel vector.

With this kernel, the maximum resolution in frequency would be one step followed by a zero (i=1); this pattern repeats for the duration of the window N_w, matching the length of the ACF. This is the maximum frequency detected, and corresponds to F_{max}=1/2T_s (Hz), where T_s is the sampling interval, F_{max} corresponds to the Nyquist limit. On the other end, the lowest detectable frequency corresponds to the kernel that has a peak in the offset zero position and another at the position N_w-1, corresponding to the frequency F_{min}=1/(T_s(N_w-1)) (Hz). Therefore the number of kernels that can be defined ranges from 1 to N_w-1, with $K \epsilon$ [1, N_w-1]. This has an interesting property, as it allows to set a frequential search range that doesn't have to cover all of the available frequencies, but one where its ends can be set within [0, F_s/2].

Once the kernels are defined, they are then arranged in columns forming the kernel matrix $\mathbf{Ker} \epsilon \mathbb{R}^{N_w x K}$. The kernel decomposition matrix $\mathbf{D} \epsilon \mathbb{R}^{MxK}$ is the result of the product of the **ACFM** and **Ker** matrices. Each

element of matrix **D** corresponds to the dot product of a row of **ACFM** and a column of **Ker**.

$$\mathbf{D}=\mathbf{ACFM}\cdot\mathbf{Ker} \quad (3)$$

Therefore the better aligned the non-zero elements of the kernel with the ACF pattern maxima the higher the product. This acts as an emphasis filter, making the ridges of the ACFM sharper creating a greater difference between contiguous points on the ACF as exemplified in Fig. 2.

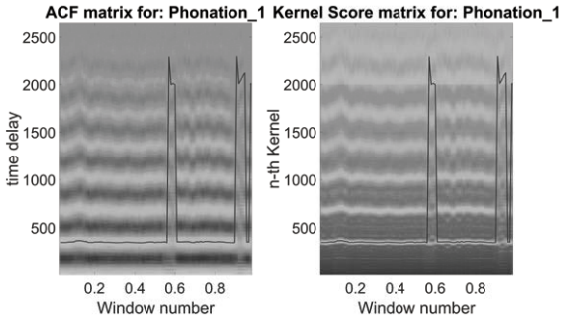


Fig. 2) ACFM (**left**), Kernel decomposition matrix **D** (**right**) and initial F0 estimation profile (**red**). The columns of each matrix correspond to the estimation for individual windows and by selecting the maxima the F0 profile is estimated.

Under a well-conditioned voice signal, this would suffice to estimate the fundamental frequency by searching for the element with the highest value per column as temporal evolution of the F0 (F0 contour) can be extracted. The problem is that as can be seen in Fig. 2 the ACF may present irregular patterns or other phenomena that may cause the estimation to fail. The next step in the estimation (post-processing of the candidate F0 estimates) is the use of a ridge recognition algorithm that tracks the vertical variations of the maximum and selects the path along the resulting ridge. The algorithm decomposition operates on the assumption that abrupt changes are not possible (as expected e.g. when estimating F0 in sustained vowels), and if present they are most likely the result of an erroneous estimation rather than because of actual oscillations. The function that estimates the ridge introduces a penalty to abrupt shifts; this penalty ranging from zero to one penalizes changes in frequency, limiting the response of the ridge function, the closer it is to zero the looser the estimation; otherwise, it approaches the straight line that best covers the ridge. The estimator finds multiple solutions and picks the one with the largest average value. In Fig. 3 the estimation can be observed.

In order to contrast the results of the proposed algorithm against a well-known F0 estimation algorithm was applied, this algorithm is based on the ACF traditional estimation [6] (normalized autocorrelation). To test the proposed method and establish a comparison with the mentioned reference, the database analyzed was composed of a set of 130 artificially generated signals using a sophisticated mathematical model [7]. The artificial signals contained in the database were constructed in order to emulate a sustained vowel /a/, with differing degrees of pathological effects built into them. Alongside with the audio recording, a ground truth for the F0 was provided; this allowed to compare all the estimations against a benchmark of quality. In order to be directly comparable to the results reported in [2] the same performance measures to assess errors have been used. Specifically, the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|e_i - g_i| \quad (4)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(e_i - g_i)^2}{N}} \quad (5)$$

where e_i and g_i are the estimation and the ground truth.

Using Eq. (4) and (5) the multiple estimations for each signal are compared to the provided ground truth. Then the average errors with their standard deviations are extracted and this generates an estimation of how well each approach performed.

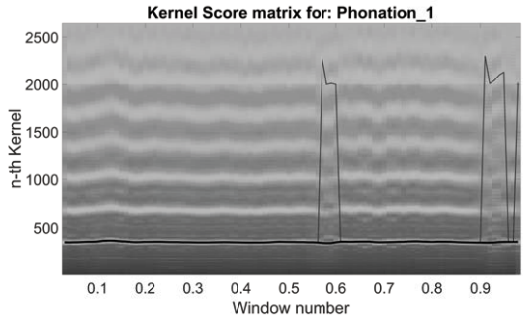


Fig. 3) Kernel decomposition matrix D (background), initial kernel estimation (**red**) and F0 ridge estimation profile (**black**).

III. RESULTS

For each recording the Kernel, Kernel-Ridge and NCF estimations are compared with the provided ground truth. The three RMSE scores for each recording in the database are presented in Fig. 4. It can be observed that on the average the best performing approach is the Kernel-Ridge estimation and the worst is the NCF algorithm. The results show that the NCF was the worst performing estimation for 112 of the recordings. The simple kernel estimation was the worst on 13 recordings and the ridge

estimation on 2 of the recordings. Table 1 shows the average value of the MAE and RMSE errors and the variance of each approach.

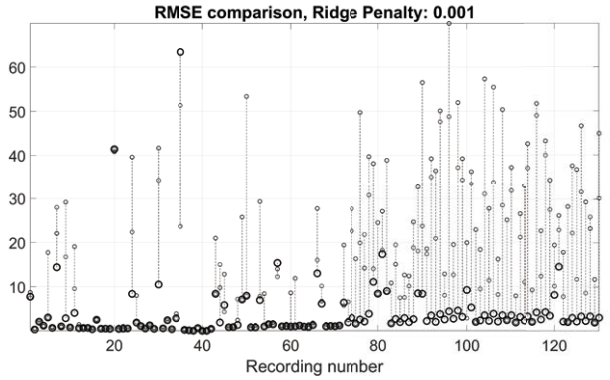


Fig. 4) RMSE scores for each estimation; Kernel (**red**), Kernel-Ridge (**black**) and NCF estimation (**blue**).

Table 1: summary of the results given as mean ± standard deviation.

Algorithm	Mean MAE(HZ)	Mean RMSE(Hz)
NCF	9.11 ± 10.21	17.67 ± 17.46
Kernel	5.67 ± 7.40	11.27 ± 13.25
Kernel-Ridge	**2.79 ± 5.72**	**4.003 ± 7.91**

Once the methodology has been shown to have a successful performance it is interesting to observe its performance on a real signal. The signal that was tested was recorded trough a telephonic line, it was selected for its low quality and the manifestation of a phenomenon of doubling frequency. This process is exemplified in Fig. 5.

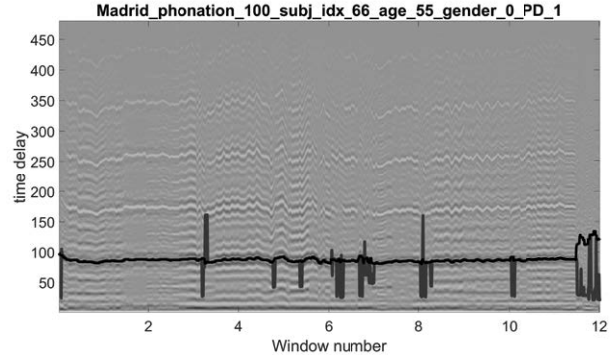


Fig. 5) Estimation of F0 on a real signal. Kernel-Ridge estimation (black), NCF estimation (red), ACFM (background). It can be appreciated that the Kernel-Ridge estimation provides a more stable and robust estimation than the NCF approach.

IV. DISCUSSION

A new F0 estimation algorithm has been introduced relying on ACF and the introduction of kernel functions. Results show that the best performing approach towards F0 estimation is the Kernel-Ridge: the use of kernel exploration and the ridge estimation produces the overall more accurate results. It must be stated that even though the simple Kernel estimation outperforms the NCF estimation it still has a substantial amount of variability. This variability is caused by the simple composition of kernel function. As the kernel is built of deltas the number of non-zero points is small. Due to this fact, temporal precision is lost and leads to variability in the estimation. For future work the kernels should be designed keeping in mind this additional degree of freedom.

V. CONCLUSION

The results show that the Kernel-Ridge is a competitive approach which outperforms a widely established method of estimating the F0 (NCF) using a temporal approach. This study introduced a new methodology that is robust against instability in the F0, finding the most likely F0 estimate.

VI. ACKNOWLEDGEMENTS

MRC DTP Precision Medicine PhD Studentship. Usher institute Edinburgh Medical School: Molecular, Genetic and Population Health Sciences.

REFERENCES

[1] I. R. Titze, *Principles of Voice Production*, Prentice-Hall, Englewood Cliffs, N. J., 1994.

[2] A. Tsanas, M. Zañartu, M. A. Little, C. Fox, L. O. Ramig, and G. D. Clifford, "Robust fundamental frequency estimation in sustained vowels: Detailed algorithmic comparisons and information fusion with adaptive Kalman filtering", *The Journal of the Acoustical Society of America*, Vol. 135, No. 5, 2014, pp. 2885-2901.

[3] C. Manfredi, M. D'Aniello, P. Bruscaglioni, A. Ismaeli, "A comparative analysis of fundamental frequency estimation methods with application to pathological voices", *Medical Eng. & Physics*, Vol. 22, 2000, pp. 135-147.

[4] S. R. Kadiri and P. Alku, "Analysis and Detection of Pathological Voice Using Glottal Source Features", IEEE Journal of Selected Topics in Signal Processing, Vol 14, No. 2, 2020, pp. 367-379.

[5] B. S. Atal, "Automatic Speaker Recognition Based on Pitch Contours." *The Journal of the Acoustical Society of America*, Vol. 52, No. 6B, 1972, pp. 1687–1697.

[6] https://mathworks.com/help/audio/ref/pitch.html

[7] M. Zañartu, "Acoustic coupling in phonation and its effect on inverse filtering of oral airflow and neck surface acceleration," Ph.D. dissertation, School of Electrical and Computer Engineering, Purdue University (2010).

SESSION II
SPEECH

SUPRAGASTRIC BELCHING: SPEECH THERAPY INTERVENTION REDUCES EXCESSIVE BELCHING SYMPTOMS

L. ten Cate[1],
[1]Practice for Voice and Speech, Logopedie aan de Amstel, Amsterdam, The Netherlands,
ltencate@xs4all.nl

Abstract: **A specific speech therapy treatment program was performed in 113 patients with supragastric belching. We retrospectively compared the visual analogue scale scores (VAS) pre- and post-treatment of 73 included patients and found that speech therapy significantly reduces physical and psycho-social symptoms of belching.**
Keywords: **Supragastric belching, impedance measurement, inhalation, injection, speech therapy.**

I. INTRODUCTION

Supragastric belching (SGB) is considered a behavioral disorder and thought to be caused by stress factors in which a person unconsciously inhales or injects air into the esophagus, after which the air is immediately expelled without reaching the stomach. Intra-esophageal impedance measurement can establish the direction of the air passage through the esophagus and therefor differentiate between gastric and supragastric belches [1]. A supragastric belch is characterized by a rapid increase of impedance (≥ 1000 Ω) in aboral direction, immediately followed by a decrease of impedance to normal values in opposite direction within a second (Fig. 1). From impedance measurement combined with manometry it is known that one can perform the esophageal air influx in two different ways. The most frequent mechanism is inhalation, caused by an aboral displacement of the diaphragm, followed by a pressure decrease in the esophagus and relaxation of the upper esophageal sphincter (UES). The air is sucked into the esophagus. The other mechanism is injection, caused by an increase in pharyngeal pressure that is build up with tongue movements and simultaneous UES relaxation. In this mechanism the air is pushed into the esophagus [2,3]. In both mechanisms the sudden closure of the larynx(glottis) plays an important role [4].

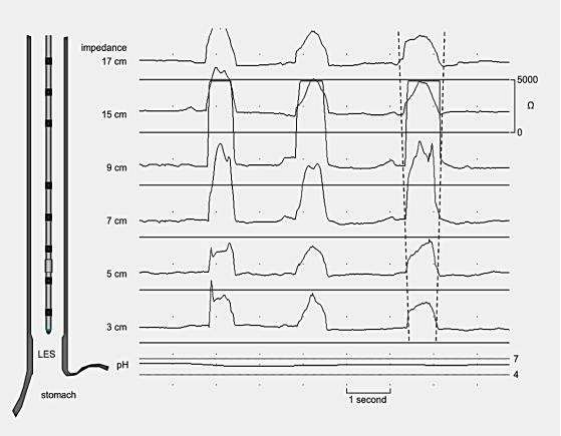


Figure 1. Impedance signal pattern of supragastric belching (three times). The 'V-shaped' dashed lines indicate the direction of air. LES: Lower Esophageal Sphincter.

Patients suffering from SGB can have large numbers of belches, like hundreds a day, sometimes continuously, up to 20 per minute [5].
Excessive belching often turns out to be supragastric and leads to great physical and social discomfort. In the past, a specific speech therapy program [2] was developed to unlearn the supragastric belching mechanism.We described the effect of this therapy in a previous publication [4].

II. METHODS

To enlarge the previously studied group we retrospectively collected 40 additional files of patients who were treated from 2017 until June 2021 for SGB as the main symptom.
One hundred and thirteen patients were treated by the author from 2007 to 2021. Fifty-one patients were objectively assessed by impedance measurements in the referring hospital clinics, and 62 were clinically assessed following criteria described [4].

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

Belching symptoms were scored pre- and post-treatment on a six-item visual analogue scale (VAS) of 100 mm (Table 1) on the severity of symptoms filled in by the patient.

Table 1. VAS-Questionnaire

1	How bothering do you experience your symptom of excessive belching?
2	In your opinion, how disturbing is your excessive belching for others?
3	Can you suppress your belching?
4	Does excessive belching hamper your work or activities?
5	Are your social activities hampered by excessive belching?
6	Do you experience any level of control over your excessive belching?

Speech Therapy consisted of
1. Explanation of the supragastric belching mechanism, besides giving attention to the patient's own ideas about the belching.
2. Creating awareness of the preceding acts and sensations before belching, and recognition of the glottal (laryngeal) closure as part of the inhalation or injection mechanism of esophageal air influx.
3. Exercising a fluent abdominal breathing. Open mouth breathing with a finger or a cork between the teeth is performed if belching occurs continuously and severely.
4. If necessary, exercises to normalize functions of the lingual-laryngo-cricopharyngeal complex were done (maxilla relaxation, voicing, swallowing) depending on the patient's presentation.
5. Implementation into daily life, practicing situations in which supragastric belching occurs based on belching diaries.

Statistics: The treatment effect of the enlarged group was evaluated by means of the Wilcoxon Signed Rank test. We compared the results of the objectively and clinically assessed patients, and the results of the previous studied group and the additional group by use of the Mann-Whitney U-test. The data are presented as median (interquartile range (IQR).

III. RESULTS

Forty patients of the total group of patients were excluded because of missing one or two VAS questionnaires. Reasons for missing data were problems in recollecting the questionnaires (forgot to ask, not filled in or not returned, language problems misinterpretation of the questionnaire). Of them, 17 patients achieved good results according to the patient reports, and 4 had insufficient or no results. In 16 cases there was a premature termination of the treatment (health issues, movement, job, disappointing results). Table 2 shows the reason for exclusion of the objectively assessed group and clinically assessed group separately.

Table 2. Diagram of the excluded cases

Treated patients with SGB Impedance-confirmed (N= 51)	Treated SGB patients Clinically assessed (N=62)
formal termination therapy, but no VAS-post questionnaire good results (patient reports): 7 insufficient results: 2	formal termination therapy but no VAS-post questionnaire good results (patient reports) :10 insufficient results: 2
premature termination: 5	premature termination: 11
	misinterpretation VAS: 3
Included N=37	Included N=36

The pre- and post-treatment scores of 73 patients (34 male, median age 49 (27-60; range 8-90) (7 children of 8, 11, 13, 15(3) and 17 years old) were used for analysis. Median symptom duration was 24 months (12 - 39), median therapy duration was 13 weeks (8 – 18,5) and the median number of sessions was 9 (6 – 11).
The Speech Therapy program resulted in a significant reduction of the supragastric belching and related symptoms on all items of the VAS questionnaire (Table 3).

Table 3. Therapy outcome (median, IQR) of belching symptoms

items	VAS score (mm) pre-treatment	VAS score (mm) post-treatment
'how bothering'	88 (78 – 98)	22 (6.5 – 36)
'disturbing others'	53 (23 – 84.5)	8 (2 – 26.5)
'suppress'	73 (37 – 92.5)	15 (2.5 – 34)
'work/activities'	60 (14.5 – 83)	6 (1 – 19.5)
'social'	51 (20 – 78)	10 (0 – 28.5)
'level of control'	86 (54.5 –94.5)	18 (3 – 36)

The VAS scores decreased after Speech Therapy from a total median score (six items) of 395 (296 – 461) pre-treatment to 101 (30-195) post-treatment ($p<0.001$) (Fig. 2.).

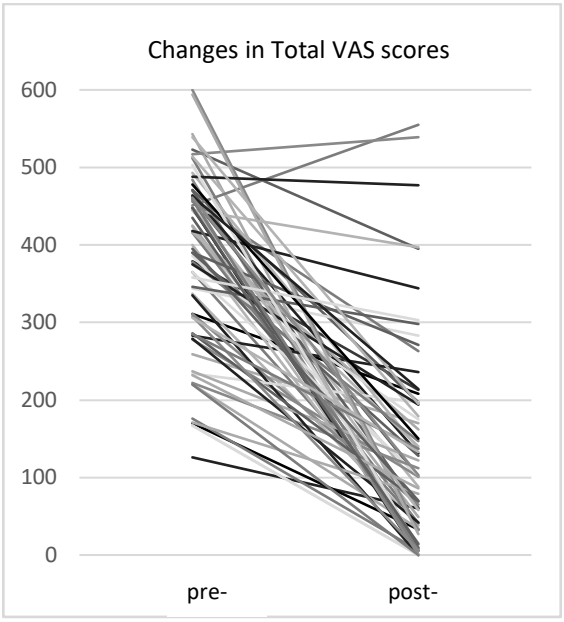


Figure 2. Changes in Total VAS-scores (mm) pre- and post-treatment

The results showed a significant decrease of the symptoms concerning the (physical) belching (items 1,3 and 6) from a subtotal median of 233 (183-269) to 58 (15-110) ($p<0.001$). Furthermore, for the items concerning the social impact (items 2, 4 and 5) from a subtotal median of 164 (107-217) to 32 (5-69) ($p<0.001$). In 84% Speech Therapy had a positive effect of which 55 patients had a sufficient (> 180 mm total VAS change) or major improvement (> 360 mm total VAS change). Twenty patients ended the therapy completely symptom-free. The total VAS score changes did not differ significantly between the patients with objectively assessed SGB and with clinically assessed SGB ($p=0.573$). Age ($p=0.778$), gender, therapy duration ($p=0.687$) and number of sessions ($p=0.833$) did not significantly differ between objectively and clinically assessed patients. A significant difference ($p = 0.033$) in symptom duration was found between the impedance confirmed (Mdn = 36 months and the clinically assessed groups (Mdn = 18 months). Also, a significant difference ($p=0.026$) was found in number of sessions between the previously studied group (Mdn = 10 sessions) and the additional group (Mdn = 7 sessions).

IV. DISCUSSION

Supragastric belching is a disorder that can have serious consequences for the physical well-being of the patient and his/her quality of life [6]. The therapy of choice is a special program of Speech Therapy [4] or a specific Cognitive Behavioral Therapy [7] that have proven to be effective. In this study we enlarged our previously described group of patients who were treated with Speech Therapy with attention to cognitive aspects and fluent abdominal breathing. The aim of therapy is to prevent air inhalation and air injection into the esophagus. Continuous quiet (pulmonary) breathing makes inhalation and injection difficult. Patients with SGB almost always have a high thoracic breathing pattern with many breath stops. Restoring abdominal breathing through an open glottis is therefore the most important physical intervention of therapy. Comparable to our previous study the treatment resulted overall in improvement of the excessive belching symptoms. In several patients the improvement was impressive and was achieved quickly, sometimes in two or three sessions. In others it took (much) more time and effort to change the belching behavior. Probably the capability to recognize physical sensations, stress, and the influence of stress on patient's breathing pattern play important roles. We found a remarkably longer symptom duration of 36 months in the impedance-confirmed group in comparison to 18 months of the clinically assessed group. A possible explanation could be that the patients who finally get an impedance measurement, already underwent a long period of medical investigations such as gastroscopy to rule out abnormalities. We notice in our practice that the diagnosis SGB is often made late. Earlier recognition will be beneficial to the patient. The additional group had lower number of sessions than the earlier described group. This could point to less severe cases, practical circumstances (large geographic distances) or maybe to the advancing experience of the therapist. The considerable number of excluded patients might have influenced the therapy outcome. Closer analysis of these cases in which therapy was not successful (both excluded and included patients) points to difficulties in accepting and adherence to the explanation of SGB and therapy means. Also, these cases show more complex representations of symptoms (co-morbidity) and more psychological problems. A limitation of this study is that this is a retrospective and open label study in which the treatments were done by the same therapist. Another limitation is that the belching and related symptoms were not assessed objectively after treatment, but only evaluated by VAS. Because the most important thing in therapy practice is to resolve the problem, it can be difficult to motivate the patient for a second measurement, especially if the complaints have disappeared. A less invasive way to objectively diagnose and quantify supragastric belches would be very useful.

V. CONCLUSION

A specific program of Speech Therapy intervention reduces symptoms of supragastric belching.

REFERENCES

[1] A.J. Bredenoord, B.L. Weusten, D. Sifrim, R. Timmer, and A.J. Smout, "Aerophagia, gastric, and supragastric belching: a study using intraluminal electrical impedance monitoring", *Gut,* vol. 53, pp 1561-1565, 2004.
[2] G.J. Hemmink, L. ten Cate, A.J. Bredenoord, R. Timmer, B.L. Weusten, and A.J. Smout, "Speech Therapy in patients with excessive supragastric belching – a pilot study", *Neurogastroenterol. Motil*, vol. 22, pp 24-e3, 2010.
[3] B.F. Kessing, A.J. Bredenoord, and A.J. Smout, "Mechanisms of gastric and supragastric belching: a study using concurrent high-resolution manometry and impedance monitoring", *Neurogastroenterol. Motil*, vol. 24, pp e573-e579, 2012.
[4] L. ten Cate, V.K. Herregods, P.H. Dejonckere, G.J. Hemmink, A.J. Smout, and A.J. Bredenoord, "Speech Therapy as Treatment for Supragastric Belching", *Dysphagia*, vol. 33, pp 707-715, 2018.
[5] A.J. Bredenoord, "Management of Belching, Hiccups and Aerophagia", *Clin Gastroenterol Hepatol,* vol.11, pp 6 -12, 2013.
[6] A.J. Bredenoord, and A.J. Smout, "Impaired health-related quality of life in patients with excessive supragastric belching", *Eur J Gastroenterol Hepatol*, vol. 22(12), pp 1420-1423, 2010.
[7] E.Glasinovic, E.Wynter, J. Arguero, J. Ooi, K. Nakagawa, E. Yazaki, P. Hajek, P. Woodland, and D. Sifrim, "Treatment of supragastric belching with cognitive behavioral therapy improves quality of life and reduces acid gastroesophageal reflux", Am J Gastroenterology, vol. 113, pp 539–547, 2018.

A COMPARATIVE STUDY OF EUROPEAN PORTUGUESE STOP CONSONANTS AND FRICATIVES IN WHISPERED AND NORMAL SPEECH FOR REAL-TIME OPERATION OF VOICE CONVERSION

João P. Silva[1], Clara F. Cardoso[1], Marco A. Oliveira[1], Luís M. T. Jesus[2], Aníbal J. S. Ferreira[1]

[1] Department of Electrical and Computer Engineering, University of Porto, Portugal
[2] ESSUA and IEETA, University of Aveiro, Portugal
joaomiguelppsilva@gmail.com, clara.f.cardoso@gmail.com, marcoantmoliveira@gmail.com, lmtj@ua.pt, ajf@fe.up.pt

Abstract: **Contrary to normal speech, whispered speech is produced without the contribution of the vocal folds. Therefore, its acoustic projection is weak, its intelligibility is easily hampered by concurrent sounds and noise, and the speaker-specific sound signature is essentially lost. This has motivated the development of an assistive technology aiming at reconstructing normal speech from whispered speech, in real-time, by carefully implanting synthetic voicing on the latter. The success of this approach depends on the phonemic durations in both normal and whispered speech realizations by the same speaker. In this paper, we focus on European Portuguese stop consonants and fricatives. A European Portuguese database has been built that contains isolated words and sentences, uttered both in normal and whispered speech, by female and male speakers. A study of the duration of stop and fricative consonants was carried out to assess if there exist statistically significant differences between normal and whispered speech both in isolated word and sentence contexts. Results show that despite a few non-representative exceptions, in most cases of interest, differences are not statistically significant. This confirms that when reconstructing Portuguese voiced sounds from whispered speech the algorithm operation is not required to enforce any special duration compensation strategy.**

Keywords: **Stop consonants, fricative consonants, closure duration, whispered speech**

I. INTRODUCTION

Speech communication is the most important modality of human social and professional interaction [1, 2]. In normal speech, most sounds involve vocal fold vibration, but when a health condition affects the vocal folds, as in certain cases of laryngectomy, then the associated speech is known as whispered speech. Whispered speech is problematic because its acoustic projection is weak, its intelligibility is strongly affected by concurrent sounds and noise and, although short-distance voice communication is still possible, most of the sound signature of a specific speaker is lost. This causes communication difficulties, which has a negative impact in professional and social life. This motivated the development of an assistive technology (DyNaVoiceR, www.dynavoicer.com) whose objective is to reconstruct natural speech sounds from whispered speech, in real-time, to allow effective and comfortable communication by patients while using their speech production system seamlessly. The assistive technology that we are developing [3,4,5,6] takes the input whispered speech as a baseline signal, identifies those regions in the signal that would be voiced in natural speech, and implants, in these regions, synthetic voicing creating a replacement for the missing vocal folds contribution. This replacement is carefully shaped in frequency and time such as to enhance the linguistic content of the resulting synthetic speech, to improve voice projection, and to convey elements of the sound signature of a given speaker. The success of this approach depends on the phonemic durations in both natural speech and whispered speech realizations by the same speaker, so that a realistic reconstruction of the former can be done by implanting synthetic voicing on the latter. In this paper, we focus on European Portuguese (EP) stop consonants and fricatives.

Duration studies for fricatives in American English, as reported in [7], and based on listening tests with 12 subjects, concluded that the minimum frication duration required for correct identification depends on the particular fricative, ranging from approximately 30 to 50 ms. The author also notes that, unsurprisingly, identification improves as the duration of the frication noise increases. Previous studies [8] have also looked at the relative importance of the transitions and the frication duration on the perception of the voiceless fricatives /f/ (as in <face>), /s/ (as in <soap>), and /S/ (as in <shame>). The authors note that transition phase spectral characteristics dominate over frication noise duration in terms of fricative identification in several of the tested scenarios. Jesus and Jackson [9] examined the phonetic detail of voiced and voiceless fricatives. In that study, duration statistics were derived from the voicing and frication labels to distinguish between

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

voiceless and voiced fricatives in British English and EP. They concluded that, in normal speech, clusters for voiceless and voiced fricatives are centered at 115 ms and 50 ms, respectively. In a cross-linguistic (Portuguese, Italian and German) devoicing study, Pape and Jesus [10] included stops and fricatives in four vowel contexts and two-word positions and computed the devoicing of the time-varying patterns throughout the stop and fricative duration. They showed that consonant durations are very similar across languages and that considerably longer durations are prevalent for the voiceless consonants when compared to their voiced counterparts. For EP, durations of approximately 100 ms were identified for voiced stop consonants, while the voiceless group presented durations of approximately 150 ms. The above studies, as well as other research results in the literature, consider voiced speech only (*i.e.*, normal speech). In this paper, we focus on a comparison of durational patterns for stops and fricatives between normal and whispered speech. Therefore, an EP database has been created for the DyNaVoiceR project that contains isolated words and sentences uttered in both modes: normal and whispered speech. A study of the stop and fricative consonants was carried out to analyze their duration and to assess whether or not there exist statistically significant differences between normal and whispered speech, both in isolated word and sentence contexts, for female and male speakers.

II. METHODS

Thirty volunteer speakers (15 females and 15 males) were recruited using convenience sampling in the districts of Aveiro and Coimbra, in Portugal, and a database containing whispered and normal speech material was recorded for the DyNaVoiceR project. Recording and manual phonetic annotation tasks for the entire database were performed at the University of Aveiro. The recordings took place in a sound booth with 45 dB sound reduction and using a Sennheiser Ear Set 1 microphone. The sampling frequency was 48 kHz and the sample resolution 16 bits. The database includes 28 isolated words and 6 sentences, among other tasks. Each task was repeated 3 times both in normal speech, and whispered speech modes, by each speaker.

In this paper, we use an underscore W to identify the whispered version of each task (*e.g.*, <nuca$_W$> represents the whispered version of the Portuguese word <nuca>). The analyses conducted in this study were performed using MATLAB R2016b 64-bit.

In our study, we include both voiceless /f, s, S/ and voiced /v, z, Z/ fricatives, and voiceless /p, t, k/ and voiced /b, d, g/ stops.

All stop consonants and fricatives produced in isolated words and sentences contexts have been analyzed regardless of their syllable and sentence position. However, for closure duration analysis, only voiceless stop consonants in intervocalic contexts were considered. It should be noted that the number of samples (*i.e.*, the number of instances in the database) per consonant is not always the same because in the manual annotation process it was detected that the participants did not always produce the correct stop and fricative consonants. Only correct and clearly identifiable stop and fricative consonants were used in the study. For this same reason, the sample number may also differ between female and male speakers.

Durational patterns of fricative and stop consonants of normal and whispered speech were carefully analyzed. In particular, a detailed statistical analysis of the results was performed focusing on 95% confidence intervals around the means, and on the statistically significant differences between those means.

III. RESULTS

This section presents an analysis of the closure and total duration of stop consonants, and the total duration of fricatives via box-plots, as well as statistical inference results regarding normal and whispered speech. The intervocalic stop consonants' duration labelling was performed manually using the waveform and corresponding spectrogram concerning the second repetition of each word in our database. We carried out a statistical hypothesis Wilcoxon signed-rank test with 5% significance level in order to draw statistical inferences from normal/whispered speech recordings.

Figure 1 shows the particular closure duration distribution for the words containing stop consonants, both in normal and whispered speech modes, and based on the recordings of 15 male participants. Each box-plot reflects 15 data points, one for each of the participants. The symbol '+' represents outliers, the symbol 'x' represents the average value, and the horizontal line corresponds to the median value.

An analysis of the overall closure duration results, for both male and female speakers, shows that, except for the words <nuca> and <ripa> produced by female speakers, the average whispered speech closure duration is slightly longer than the normal speech average closure duration. However, none of the *p*-values are below the level of significance of 5% ($p>0.2185$ and $p>0.2747$ in the case of female and male speakers, respectively), which indicates that there are no statistically significant closure duration differences between normal and whispered speech for the 6 words analyzed in this study. This is expected, as

the mean closure duration differences are rather small between normal and whispered speech realizations.

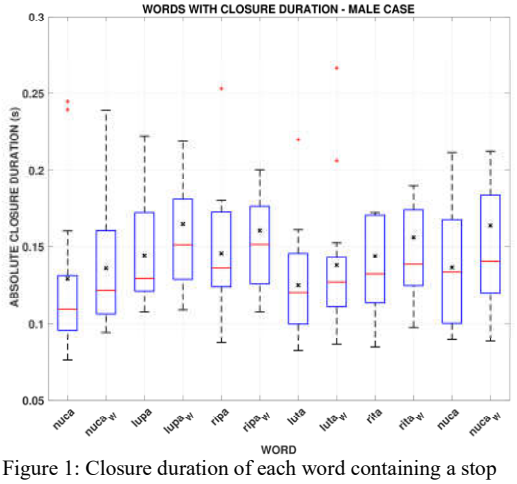


Figure 1: Closure duration of each word containing a stop consonant, produced by male speakers, for both normal and whispered speech modes.

A similar analysis of the stop consonants total duration was also carried out, as illustrated in Figure 2.

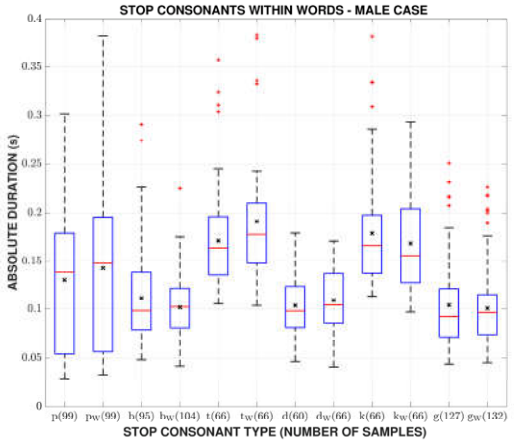


Figure 2: Duration of stop consonants in isolated words, produced by male participants, for both normal and whispered speech scenarios.

It can be observed that, in most cases, the average total duration of whispered stop consonants tends to be slightly longer than the corresponding duration in normal speech. However, in general, the differences are small and not statistically significant, with all p>0.07 in the case of isolated words, similarly to the sentences results with only one significant difference p=0.0187 for the stop consonant pair [g]-[g$_W$]. In the case of female speakers, the average whispered stop consonants total duration is also slightly longer than that of the voiced counterparts. Albeit the pair [d]-[d$_W$] in isolated words showing a statistically significant difference (p=0.0053), this is not relevant because none of the differences in the case of sentences has been found to be statistically significant (all p>0.1137).

A similar duration analysis was carried out for fricative consonants with the similar goal to ascertain if there exist statistically significant differences between normal speech and whispered speech, both in words and sentences contexts, for both female and male speakers.

As an illustrative example, Figure 3 shows the distribution of the duration of each fricative in isolated words regarding male participants. Differences in the mean duration results in the case of words and sentences contexts are minor, and mixed, without a clear trend of a tendency for whispered fricatives to be longer or shorter than normal speech fricatives. With respect to sentences, none of the differences were found to be statistically significant (all p>0.66), however, with respect to words, 5 in 6 cases show p-values less than the level of significance (p<0.044), which means that the average duration of fricatives in isolated words tends to differ significantly. However, isolated words are not as representative of normal speech as sentences are.

In the case of female speakers, similar conclusions were reached concerning the isolated words tests. However, in the case of the sentence tests, 2 out of 6 cases were found to exhibit statistically significant differences (p<0.026) between the average duration of fricatives, specifically in the case of the fricative pairs [S]-[S$_W$], and [z]-[z$_W$].

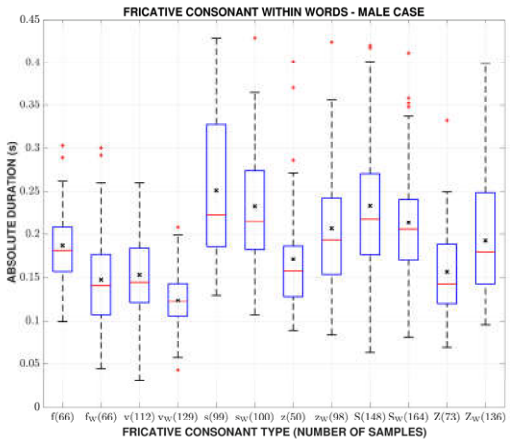


Figure 3: Duration of fricative consonants in isolated words, produced by male participants for both normal and whispered speech scenarios.

IV. Discussion

The paper discusses two sets of results: those regarding intervocalic stop consonants (including closure duration and total duration) and those regarding voiced and voiceless fricatives. Results show that in

both stops and fricatives, while in some non-representative cases statistically significant differences can be found, in most cases of interest, especially regarding sentence contexts, differences were not found to be statistically significant. This important outcome confirms that when reconstructing "on the fly" Portuguese voiced sounds from whispered speech, in real-time, in addition to a careful phoneme-oriented segmentation, the algorithm operation does not need to adopt any special compensation strategy in the whispered speech to normal speech conversion process and regarding the duration of fricatives or stop consonants.

V. Conclusion

While most stop and fricative duration studies available in the literature consider voiced speech only (*i.e.*, normal speech), in this paper we focused on a comparison of durational patterns between normal and whispered speech, using male and female recordings.

The first conclusion that can be drawn from our work is that the voiceless stop consonants average closure duration tends to be slightly longer in whispered speech than in normal speech. Despite a few non-representative exceptions, a statistical analysis comparing whispered speech and normal speech shows that there are no consistent statistically significant differences between the two speech modes.

Regarding stop consonants total duration, in the case of male speakers, no statistically significant differences were found between whispered and normal speech realization in word contexts, and only one statistically significant difference was found in sentence contexts. Regarding female speakers, the opposite was verified.

Therefore, in general, it can be concluded that regarding the average closure duration and the average total duration of the stop consonants analyzed in this paper, there is a tendency for the whispered speech realizations to be slightly longer than in normal speech, however, differences are rather small and negligible.

Regarding fricatives, considering the results of both male and female speakers, it was observed that the average duration of fricatives in isolated words tend to differ significantly although not in a consistent manner. In sentence contexts, which are more representative of normal speech, fricative duration differences are not statistically significant in the case of male speakers, and, in the case of female speakers, in only 2 (out of 6) cases differences were found to be significant.

As a summary, representative and systematic statistically significant stop and fricative duration differences between whisper and normal speech realizations have not been found. This important outcome confirms the real-time operation feasibility of the DyNaVoiceR assistive technology converting Portuguese whispered speech into naturally sounding synthetic speech. This is because in its "on the fly" operation, the algorithm does not need to implement any stop/fricative consonants duration compensation. The linguistic implications of this decision will be fully assessed as the DyNaVoiceR algorithm approaches the final stages of development, in the near future.

ACKNOWLEDGMENTS

This work was financially supported by Project PTDC/EMD-EMD/29308/2017 - POCI-01-0145-FEDER-029308 - funded by FEDER funds through COMPETE2020 - POCI and by national funds (PIDDAC) through FCT/MCTES. Support was also received from National Funds through the FCT - Foundation for Science and Technology, in the context of the project UIDB/00127/2020.

REFERENCES

[1] Gunnar Fant, "Acoustic theory of speech production", The Hague, Netherlands. Mouton, 1970

[2] Douglas O'Shaughnessy, "Speech Communication: Human and Machine", *Wiley-IEEE Press*, 1999

[3] Aníbal Ferreira, "Implantation of voicing on whispered speech using frequency-domain parametric modelling of source and filter information", *ISIVC2016* (invited paper), Tunisia, 2016

[4] J. P. Silva, M. A. Oliveira, C. F. Cardoso, A. J. Ferreira, "Manipulation of the Fundamental Frequency Micro-Variations Using a Fully Parametric and Computationally Efficient Speech Model", *IEEE IWSPS*, Portugal, 2020

[5] A. J. Ferreira, J. Silva, F. Brito, D. Sinha, "Impact of a Shift-Invariant Harmonic Phase Model in Fully Parametric Harmonic Voice Representation and Time/Frequency Synthesis", *ICASSP2020*, Spain, 2020

[6] J. Silva, M. Oliveira, A. Ferreira, "Flexible parametric implantation of voicing in whispered speech under scarce training data", *EUSIPCO 2020*

[7] A. Jongman, "Duration of frication noise required for identification of English fricatives", *JASA* 85 (4), 1989, pp. 1718–1725.

[8] K. Nataraj, P. Pandey, H. Dasgupta, "Effect of frication duration and formant transitions on the perception of fricatives in VCV utterances", ICASSP 2020, pp. 6259- 6263

[9] L. Jesus, P. Jackson, "Frication and voicing classification" In A. Teixeira, V. Lima, L. Oliveira, and P. Quaresma (Eds.), Computational Processing of the Portuguese Language, 2008, pp. 11-20. Berlin: Springer-Verlag.

[10] D. Pape, L. Jesus, "Stop and fricative devoicing in European Portuguese, Italian and German", *Language and Speech* 58(2), 2015, pp. 224–246.

MEASUREMENT OF SPEECH INTELLIGIBILITY AFTER ORAL OR OROPHARYNGEAL CANCER BY AN AUTOMATIC SPEECH RECOGNITION SYSTEM

M. Balaguer[1,2], L. Gelin[1], V. Woisard[2,3], J. Farinas[1], J. Pinquier[1]

[1] IRIT, CNRS, Université Toulouse III, Toulouse, France
[2] Hôpital Larrey, CHU Toulouse, Toulouse, France
[3] Laboratoire Octogone-Lordat, Université Toulouse II, Toulouse, France

mathieu.balaguer@irit.fr, lucile.gelin@irit.fr, woisard.v@chu-toulouse.fr, jerome.farinas@irit.fr, julien.pinquier@irit.fr

Abstract:
Background: Speech intelligibility alteration is a frequent consequence of oral/oropharyngeal cancer. The development of automatic speech recognition (ASR) systems could overcome the limitations of perceptual speech assessment.
Objective: To predict speech intelligibility after treatment of oral or oropharyngeal cancer using scores from an ASR system.
Methods: Spontaneous speech of patients was recorded during a semi-structured interview. Six experts evaluated the subjects' intelligibility perceptually. An ASR system (TDNNf-HMM) trained on healthy adult speech and adapted to phoneme recognition was also used. Automatic scores were computed: phonemic scores, confidence scores. LASSO regression was used to select the parameters from the ASR system that best predicted intelligibility.
Results: Spontaneous speech of 25 patients was recorded. LASSO regression led to retain 3 parameters: number of sonants recognized per second, proportion of occlusives, and average confidence score of fricatives. These three parameters present a strong correlation (rs=0.91) with the perceptual score (expert panel). This automatically predicted score is stable and reliable (5-block cross-validation: rs= 0.90).
Conclusion: The use of ASR systems in the measurement of intelligibility in ENT oncology is promising. An optimization of these systems for pathological speech would open new perspectives for the determination of fine low-level speech deficits to adapt therapeutic objectives.
Keywords: **Speech, Automatic analysis, Oncology**

I. INTRODUCTION

Oral or oropharyngeal cancer alter speech abilities [1], in particular speech intelligibility. Intelligibility can be defined as the degree of accuracy with which the acoustic speech signal produced by a speaker is decoded by a listener in terms of "low-level" units (i.e., phonemes, phoneme groups, or syllables) [2].

Speech disorders are one indicator of intelligibility, and are mainly measured perceptually in clinical assessment [3]. Therapists quantify intelligibility using a variety of measurement tools, such as visual analog scales, Likert scale measures, or by measuring an error rate after transcription [2]. However, this standard perceptual evaluation has many limitations, particularly concerning its reliability. This measure is indeed judge-dependent, due to expertise effects or differences in internal referents [4]. Intra-individual variability effects are also involved: the same judge may assign different scores depending on the assessment context, the mental availability or habituation to pathological speech [5].

To overcome these biases, new tools for automatic instrumental speech assessment are being developed. They aim at extracting from the speech signal parameters for characterizing impairments [6]. These automatic and acoustic tools measure the quality of acoustic-phonetic decoding in a controlled speech context, such as text reading [7]. But few are applicable to spontaneous speech, due to a lack of a reference to which to compare the patient's speech – automatic alignment requiring prior manual transcription is too constraining to be applicable. Yet, this production context is the closest to the daily speech production [8] and needs to be investigated.

The objective is to predict speech intelligibility after treatment of oral or oropharyngeal cancer using scores from an automatic speech recognition system.

II. METHODS

This study is a cross-sectional observational study.

The study protocol was approved by the Committee for the Protection of Persons (CPP: Ouest IV, 19/02/2020, reference 11/20_3) within the framework of the ANR RUGBI project (https://www.irit.fr/rugbi, grant ANR-18-CE45-0008).

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

A. Participants

Patients coming for consultation or hospitalization in an ENT-oriented rehabilitation service or in an ENT consultation were included. Inclusion criteria were: being of legal age (at least 18 years old) and having been treated for oral or oropharyngeal cancer (surgical treatment and/or radiotherapy and/or chemotherapy, all tumors sizes) for at least six months (chronic and stable nature of the disorder). Exclusion criteria were: fatigable patients, associated pathology potentially responsible for speech or fluency disorders (e.g., stuttering, speech disorder from neurologic disease

B. Speech recordings

All subjects were recorded in a non-anechoic room, to be as close as possible to the usual clinical evaluations. No external or internal noise (such as air conditioning or ventilation. was to be perceptible in order not to disturb the quality of the recording. The speech samples were recorded on a ZOOM H4N Pro digital recorder (48 kHz sampling rate, 16-bit resolution, mono). The headset microphone (Thomann T.Bone HC 444 TWS) was placed 6 cm from the subject's mouth, positioned frontally below the level of the lower lip and at the level of the right labial commissure. For processing, the audio files were resampled to 16 kHz. The use of a Voice Activity Detector (WebRTC-VAD: https://github.com/wiseman/py-webrtcvad) was then used to isolate the subject's speech segments, excluding the examiner's speech segments.

To get a sample of spontaneous speech, the subjects were recorded during a semi-structured interview.

C. Speech analysis

A panel of expert listeners experienced in the evaluation of speech disorders was recruited to obtain a reference measure of intelligibility: one phoniatric physician and five speech therapists practicing in an ENT/oncology department.

The experts had to listen to the recording of the interview and to quantify the intelligibility on a scale from 0 (unintelligible) to 10 (totally intelligible). The baseline perceptual intelligibility score was the average of the scores given by the 6 judges.

The subjects' speech segments – determined by the Voice Activity Detector – were given as input to a TDNNf-HMM (factorized Time-Delay Neural Network - Hidden Markov Model [9]) ASR system. The model used in this study [10] was developed using the Kaldi toolkit [11] and adapted for phoneme recognition (Phone Error Rate=23.5% on a typical adult corpus [10]). The system was trained using the Common Voice online database: in French, the training corpus includes 148.9 hours of read text recordings, by 1,276 speakers. For decoding, in each 25 ms frame (with a 10 ms step), the phone closest to the acoustic features carried by the signal will be retained and associated with the corresponding phoneme (among 33 French phonemes). A confidence score is also associated to each recognized phoneme using a Minimum Bayes Risk method [12]. WIP (Word Insertion Penalty) and LMWT (Language Model Weights) have been set to their minimum value (WIP=0; LMWT=7) to obtain a raw output.

For each subject, 16 scores were calculated based on the system outputs (see details in Table 1).

D. Statistical analysis

The analyses were carried out using Stata 16.1 software (StataCorp. 2019. Stata Statistical Software: Release 16. College Station, TX: StataCorp LLC.).

Due to the size of the study (n<30), the statistical tests used were nonparametric. In all analyses, a level of significance at 5% was chosen. For descriptive analysis, perceptual intelligibility and automatic scores were described by mean and median as indicators of central tendency, and by standard deviation, interquartile range, minimum and maximum values as indicators of dispersion. Correlations between intelligibility and automatic scores were analyzed using Spearman's correlation coefficients. Finally, the predictive process of automatic parameter selection was performed using LASSO regression (penalized regression).

III. RESULTS

A. Participants

Twenty-five patients were included (median age 67 years, IQR 12; 15 males, 10 females; oral cavity 14, oropharynx 10, both locations 1). 57.9% of patients were treated for a large tumor (T3 or T4). Surgical treatment was performed in 88% of cases (radiotherapy: 96%, chemotherapy: 60%, surgery and radiotherapy: 84%). The median time since the end of treatment was 40 months (range: 6-564 months).

B. Perceptual assessment of intelligibility (reference score)

Mean intelligibility was 6.87 (median: 7.17, range: 1.17-10). Inter-judge agreement was strong among the 6 expert judges: ICC=0.82 [0.72, 0.91].

C. Parameters from the ASR system: automatic scores

The 22 automatic scores were extracted for each subject (Table 1).

Table 1: Details of scores for the 22 automatic parameters from the ASR system

Parameter	Mean	SD	Me-dian	IQR	Min. value	Max. value
Total of different phonemes recognized (*difphon*)	4.55	1.56	4.78	2.40	1.12	7.49
Number of phonemes recognized per second						
Total phonemes (*sumphons*)	29.20	5.57	32.00	3.00	5.00	32.00
Consonants (*csns*)	2.23	0.89	2.34	1.27	0.17	4.05
Occlusives (*occs*)	0.58	0.41	0.56	0.62	0.00	1.51
Fricatives (*fris*)	0.80	0.29	0.85	0.31	0.00	1.15
Sonants (*sonants*)	0.96	0.38	1.00	0.48	0.17	1.62
Nonsonants (*nonsonants*)	1.38	0.60	1.33	0.76	0.00	2.61
Vowels (*vows*)	2.22	0.65	2.33	1.06	0.96	3.32
Semi-consonants (*semicsns*)	0.11	0.08	0.11	0.11	0.00	0.36
Proportion of phonemes recognized among consonants						
Occlusives (*propocc*)	0.23	0.12	0.27	0.20	0.00	0.37
Fricatives (*propfri*)	0.36	0.14	0.34	0.16	0.00	0.78
Sonants (*propsonant*)	0.46	0.14	0.42	0.11	0.23	1.00
Nonsonants (*propnsonant*)	0.59	0.14	0.62	0.13	0.00	0.78
Proportion of phonemes recognized among vowels						
Nasal vowels (*propvnasal*)	0.18	0.10	0.17	0.09	0.05	0.44
Proportion of phonemes recognized among all phonemes						
Vowels (*propvow*)	0.51	0.09	0.49	0.04	0.43	0.85
Nasal phonemes (*propnasal*)	0.19	0.06	0.19	0.06	0.06	0.37
Confidence scores						
Overall (*conf*)	0.84	0.02	0.84	0.03	0.78	0.89
Consonants (*confc*)	0.87	0.04	0.88	0.03	0.76	0.93
Occlusives (*confo*)	0.87	0.07	0.90	0.09	0.72	0.95
Fricatives (*conff*)	0.88	0.04	0.88	0.05	0.79	0.93
Vowels (*confv*)	0.80	0.03	0.80	0.02	0.77	0.91
Semiconso-nants (*confs*)	0.76	0.04	0.76	0.04	0.65	0.84

D. Parameters selection

Spearman's correlation coefficients are given as absolute values. Eight parameters (36%) show a high correlation with the baseline intelligibility score (rs≥0.70). Seven parameters (32%) showed moderate correlation (0.50≥rs>0.70). Details are shown in Fig. 1.

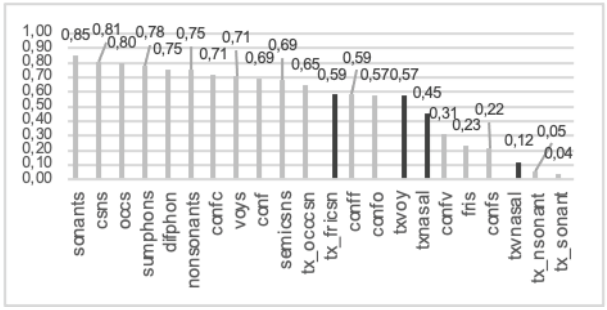


Fig. 1: Spearman's correlation coefficients between perceptual intelligibility and the 22 automatic scores (light grey: positive correlation coefficients, dark grey: negative ones).

Among the 22 automatic parameters, the LASSO regression allowed to select four parameters: the proportion of occlusives among consonants (*propocc*), the number of sonants per second (*sonants*), the average confidence score on fricatives (*conff*) and the number of occlusives per second (*occs*). An analysis of multicollinearity led to remove of the '*occs*' parameter (variance inflation factor = 7.17). The regression performed on the three remaining parameters explained 82.4% of the variance in intelligibility (R2), for a root mean squared error of 1.21. The predicted intelligibility is calculated as follows (1):

$$\text{intelligibility} = -0.073 + 4.982*sonants + 6.188*propocc + 0.851*conff \quad (1)$$

The correlation between the perceptual intelligibility and the intelligibility predicted by the automatic parameters is rs=0.91 (p<0.001) (Fig. 2).

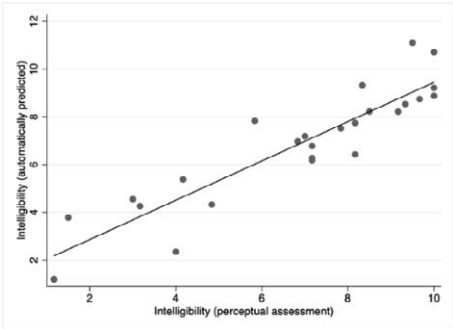


Fig. 2: Scatter plot between perceptual intelligibility and intelligibility by the three retained parameters

Cross-validation shows a strong correlation between the reference score and (i) intelligibility predicted by 5-block cross-validation (rs=0.90, p<0.001), (ii) intelligibility predicted by leave-one-out cross-validation (rs=0.90, p<0.001).

IV. Discussion

Intelligibility can be effectively predicted using three parameters from an ASR analysis of oral/oropharyngeal cancer speech. However, the results of this study can be considered preliminary due to the small sample size of subjects (n=25). The increase of the sample size would allow to conclude more strongly about the generalization and stability of these results.

The ASR system used is trained on typical (i.e., non-pathological) speech. Indeed, we wanted to measure a gap between healthy and pathological speech by targeting indicators of speech intelligibility. But one can wonder if training the system on pathological speech would allow to obtain more adapted acoustic models. In that case, if the acoustic models determined are more efficient (with a low Phone Error Rate in particular), the automatic scores calculated on the system outputs could perhaps allow to highlight finer deficits. Large corpora are necessary to train acoustic models that are relatively more stable given the pathological character of the speech. As no large French cancer speech corpus exists to date, transfer learning techniques can be used to adapt typical speech models to new corpora on relatively few data [13]. Specifically, it would be possible to adapt the current speech recognition system on other unused speech tasks in our corpus, such as sentences or text reading and pseudoword repetitions. Optimizing the quality of speech recognition could also involve the use of promising new ASR systems: the Listen, Attend and Spell (LAS) architectures [14], or Transformers [15]. These systems have been adapted to non-typical speech by Gelin [10], in this case children's speech. Their adaptation to oncologic speech would be relevant to study their performance.

ASR systems have multiple advantages in clinical evaluation: they are applicable to spontaneous speech, the scores are reliable, the required equipment is inexpensive, and the evaluation is fast. Thus, it remains relevant to explore the contributions of ASR for pathological speech analysis.

V. Conclusion

The use of ASR systems to assess intelligibility in ENT oncology is promising. An increase in sample size and analyses on optimization of these systems for pathological speech would open new perspectives for the determination of low-level speech deficits to adapt therapeutic objectives.

REFERENCES

[1] PA. Borggreven, IM. Verdonck-De Leeuw, MJ. Muller et al., "Quality of life and functional status in patients with cancer of the oral cavity and oropharynx: Pretreatment values of a prospective study", *European Archives of Oto-Rhino-Laryngology*, vol 264, pp. 651–657, 2007.

[2] KC. Hustad, "The Relationship Between Listener Comprehension and Intelligibility Scores for Speakers With Dysarthria", *Journal of Speech, Language, and Hearing Research*, vol. 51, pp. 562-573, 2008.

[3] T. Pommée, M. Balaguer, J. Mauclair, J. Pinquier, V. Woisard, "Assessment of adult speech disorders: current situation and needs in French-speaking clinical practice", *Logopedics Phoniatrics Vocology*, pp. 1–15, 2021.

[4] C. Kuo, K. Tjaden, "Acoustic variation during passage reading for speakers with dysarthria and healthy controls", *Journal of Communication Disorders*, vol 62, pp. 30–44, 2016.

[5] S. Fex, "Perceptual evaluation", *Journal of Voice*, vol 6, pp. 155-158, 1992.

[6] C. Middag, JP. Martens, G. Van Nuffelen, M. De Bodt, "Automated Intelligibility Assessment of Pathological Speech Using Phonological Features", *EURASIP Journal on Advances in Signal Processing*, 2009.

[7] M. Balaguer, T. Pommée, J. Farinas, J. Pinquier, V. Woisard, R. Speyer, "Effects of oral and oropharyngeal cancer on speech intelligibility using acoustic analysis: Systematic review", *Head and Neck*, vol 42, pp. 111-130, 2020.

[8] S. Knuijt, JG. Kalf, BGM. van Engelen, BJM. de Swart, ACH. Geurts, "The Radboud Dysarthria Assessment: Development and Clinimetric Evaluation", *Folia Phoniatrica et Logopaedica*, vol 69, pp. 143-153, 2017.

[9] D. Povey, G. Cheng, Y. Wang, et al., "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks", *Interspeech ISCA*, pp. 3743-3747, 2018.

[10] L. Gelin, M. Daniel, J. Pinquier, T. Pellegrini, "End-to-end acoustic modelling for phone recognition of young readers", *Available from: lalilo.com*, 2021.

[11] D. Povey, A. Ghoshal, G. Boulianne et al., "The Kaldi Speech Recognition Toolkit", *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[12] H. Xu, D. Povey, L. Mangu, J. Zhu, "Minimum Bayes Risk decoding and system combination based on a recursion for edit distance", *Computer Speech and Language*, vol 25, pp. 802-828, 2011.

[13] D. Wang, TF. Zheng, "Transfer learning for speech and language processing", *IEEE APSIPA*, pp. 1225-1237, 2015.

[14] W. Chan, N. Jaitly, Q. Le, O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition", *IEEE ICASSP*, pp. 4960-4964, 2016.

[15] L. Lu, C. Liu, J. Li, Y. Gong, "Exploring Transformers for Large-Scale Speech Recognition", *Interspeech ISCA*, pp. 5041-5045, 2020.

ANALYZING THE INTERACTION BETWEEN THE READER'S VOICE AND THE LINGUISTIC STRUCTURE OF THE TEXT: A PRELIMINARY STUDY

Benedetta Iavarone[1,2], Maria Sole Morelli[3], Dominique Brunato[2], Shadi Ghiasi[4], Enzo Pasquale Scilingo[4], Nicola Vanello[4], Felice Dell'Orletta[2], Alberto Greco[4]

[1] Scuola Normale Superiore, Pisa, Italy; [2] Istituto di Linguistica Computazionale "Antonio Zampolli", Pisa, Italy; [3] Fondazione Toscana Gabriele Monasterio, Pisa, Italy; [4] Dipartimento di Ingegneria dell'Informazione, Research Center "E. Piaggio", University of Pisa, Pisa, Italy

benedetta.iavarone@sns.it, msmorelli@monasterio.it, dominique.brunato@ilc.cnr.it, shadi.ghiasi@centropiaggio.unipi.it, enzo.scilingo@unipi.it, nicola.vanello@unipi.it, felice.dellorletta@ilc.cnr.it, alberto.greco@unipi.it

Abstract: **In this study, we present a preliminary analysis of the relationship between the linguistic profile of a text and the voice properties of the reader aiming to improve the speech-based emotion recognition systems. To this aim, we recorded the speech signals from a group of 32 healthy volunteers reading aloud neutral and affective texts and used the BioVoice toolbox to compute some of the main speech features. The selected texts were analyzed to quantify their lexical, morpho-syntactic, and syntactic content. Correlation and Support Vector Regressor analyses between linguistic and speech features have shown a significant modulation of some voice acoustic properties performed by the linguistic structure of the text. Particularly, a significant effect was shown on some specific speech features often used for the assessment of human emotional state (e.g., F0). This suggests that the lexical, morpho-syntactic, and syntactic properties could play an important role in the emotional dynamics of a person.**

Keywords: **Speech analysis, linguistic profile, emotions, Support Vector Regressor**

I. INTRODUCTION

Human speech is the result of fine control of up to eighty muscles from respiratory, laryngeal, pharyngeal, palatal, and orofacial groups [1]. Such control is a complex process that involves both somatic and autonomic nervous systems (ANS) activity. This latter is the main responsible for the regulation of bodily functions and is the primary mechanism of emotional regulation [2]. Alterations in the respiratory activity induced by the ANS manifest changes in the emotional state of the speaker by influencing the voice spectrum characteristics such as the fundamental frequency (F0 - the frequency of vibration of the vocal folds), and its formants (F1, F2, F3 - resonance frequencies of the vocal tract) [3]. Hence, speech processing represents one of the most promising tools in the affective computing field for a non-invasive assessment of the speaker's emotional state [4]. Indeed, voice signal analysis has been successfully used to explore several psychological dimensions of the speaker: emotion [5], mood [6], stress [7, 8], and personality [9] have been widely studied. To effectively characterize the affective prosody, several previous studies have developed and applied analytic methods to measure changes in pitch, loudness, speech rate, and pause [10]. However, the use of these features to infer the emotional state of a speaker remains an extremely complex task. One important and still little studied source of complexity could be the interaction between the speaker's hidden emotional state and the linguistic and semantic properties of what the speaker is saying. The combination of such linguistic and speech information in computational models could improve the accuracy of inferring the speaker's emotional state. Indeed, a text is characterized by many levels of information (linguistic, lexical, stylistic). By annotating these levels, it is possible to extract many features modeling the lexical, grammatical, and semantic phenomena to construct a linguistic profile that characterizes language variations within and across texts [11]. The linguistic profile has been used for different applications, such as registry and genre variation [12], or the study of psycholinguistic phenomena. In [13], the authors have shown that linguistic features can be effectively used to predict the human perception of sentence complexity, intended as processing difficulty of the language. Linguistic aspects and their effect on human processing effort and perception of complexity were studied also in [14], where the authors demonstrate that linguistic aspects from context play an important role in the perception of complexity and cognitive processing effort. Recently, Singh et al. [15] have proposed a deep learning hierarchical model for emotion recognition, combining text analysis computed by ELMo v2 with prosody, voice quality, and spectral features. However, formal modeling of the relationship between prosodic and linguistic features has not been investigated yet. In this preliminary study, we aim at studying whether the acoustic features, commonly used to characterize

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

speech production prosody, are significantly influenced by the linguistic structure of the pronounced text. To this aim, we analyzed speech signals and linguistic profiles of texts with different levels of arousal and valence. We apply correlation and regression methods to understand how the linguistic profile and structure of the texts interact with the speech production of the same texts.

II. METHODS

A group of 33 healthy volunteers was enrolled in the study (17 females), aged between 26.6 and 30.0. None of them suffered from heart diseases, mental disorders, or phobias. Each participant gave their written informed consent, and the study was approved by the Ethical Committee of the University of Pisa. We selected four texts, two describing different medieval tortures and two describing text types and writing styles. Based on the topics covered, two texts were classified as high arousal and negative valence, whereas the other two were neutral. Moreover, before starting the experiment, a group of 10 subjects, other than those enrolled in this study, evaluated the texts in terms of arousal and value, confirming the arousal and valence levels supposed apriori based on the reading topic. Each participant was asked to read aloud one neutral and one affective text, randomly chosen [16]. All texts have similar lengths to make the duration of the reading similar among subjects. The speech signal and other physiological signals such as electrocardiogram and electrodermal activity (not considered in this study) were recorded during the reading task.

A. Linguistic analysis

The texts were divided into sentences, using the full stop as a splitting criterion, i.e., identifying a sentence as the part of text between two full stops. After the splitting, neutral texts contained a total of 25 sentences, with an average sentence length of 28 tokens; affective texts contained a total of 40 sentences, with an average sentence length of 21 tokens. Each sentence was analyzed from a linguistic point of view and represented as a vector of ~140 features, a subset of the ones described in [11] that model a wide range of properties extracted from different levels of linguistic annotation. The features capture on one hand complex information like the syntactic phenomena (subordination, structure, and length of dependency relations, structure of the verbal predicates) or morpho-syntactic phenomena (distribution of grammatical categories across the text, aspects about the verb conjugation), on the other hand, they capture raw properties, like the length of the text and its components (sentences and words). The features can be grouped based on the linguistic aspects they describe and are further discussed below.

(1) Raw Text Properties. Features on the length of the text and of the sentences and the words that are in it; **(2) Lexical Variety**. Features on how varied the vocabulary of a text is, determined as the percentage of diverse and nonrepeated words over the total number of words; **(3) Morpho-syntactic information**. Features on: *(i)* the distribution in the text of grammatical categories (e.g., adjectives, nouns, determiners, pronouns); *(ii)* the ratio of content words (nouns, verbs, adjectives, and adverbs) over the total number of words in a text; *(iii)* the inflectional morphology, i.e., the distribution, for verbs and auxiliaries, of a set of inflectional features (e.g., mood, tense); **(4) Verbal Predicate Structure**. Features on: *(i)* the distribution of verbal heads, i.e., the average number of propositions (main or subordinate) co-occurring in a sentence; *(ii)* the distribution of verbal roots, i.e., the percentage of verbal roots out of the total of sentence roots; *(iii)* verb arity, i.e., the average number of instantiated dependency links sharing the same verbal head; **(5) Global and local parsed tree structures**. Features on: *(i)* the average depth of the syntactic tree, i.e., the average of the longest dependency link in a sentence. *(ii)* the average number of tokens per clause, where the number of clauses is the ratio between the number of tokens in a sentence and the number of verbal or copular heads; *(iii)* length of dependency links, i.e., the number of words occurring between the syntactic head and its dependent; *(iv)* the average depth of complement chains (a list of consecutive complements); *(v)* the order of the subject and the object in a sentence; **(6) Syntactic relations**. Features on the percentage distribution of 37 universal dependency relations; **(7) Subordination phenomena**. Features on: *(i)* the distribution of main clauses vs. subordinate clauses; *(ii)* the distribution of subordinates in post-verbal and preverbal position; *(iii)* the average number of subordinates recursively embedded in the top subordinate clause.

B. Speech signal processing

To analyze the speech time series and extract from each sentence acoustic parameters, we used the BioVoice toolbox [17]. The toolbox detected first only voiced parts of each segment. Then, F0, F1, F2, and F3 were calculated. In each voiced frame, F0 is estimated with a two-step procedure: first, Simple Inverse Filter Tracking (SIFT) was applied to signal time windows of fixed length related to the F0 range; secondly, F0 is adaptively estimated on signal frames of variable length inversely proportional to F0, through the Average Magnitude Difference Function (AMDF) within the range provided by the SIFT [18]. To extract formants values, Autoregressive Power Spectral Density (AR PSD) was considered. Furthermore, in each sentence, the total time duration of reading, the overall voiced duration, and the average voice duration were extracted.

C. Statistical analysis and modeling of the features

Before running the analyses, we scaled the frequency features of the voice in each sentence, as they are subject-dependent. For each subject and each frequency feature (F0, F1, F2, and F3), we computed F_i^{scaled}, as $F_i^{scaled} = F_i / \overline{Fi}_{neu}$ where F_i represents the frequency feature of interest (in neutral or emotional test in each sentence) and $\overline{Fi}_{neu}$ the mean of the frequency of the corresponding neutral texts, computed for all time duration. As a first analysis, we examined the relationship between linguistic features and speech features. In this way we could understand which linguistic aspects of the text are most related to speech production, discovering the underlying interaction between linguistic structure and speech. To do so, we correlated each linguistic feature with every speech one, using Spearman's correlation coefficient. We selected all pairwise correlations that had a correlation coefficient different from zero and a p-value < 0.05. Afterward, we tested the predictive strength of the linguistic profile. We implemented a regression model to predict acoustic parameters, using as input to the model the linguistic features. We employed a Support Vector Regressor (SVR) implemented with a Radial Basis Function (RBF) kernel and standard parameters. To account for within-subject repetitions, we used leave-one-out cross-validation, training the model on all subjects minus one, and testing on the left-out subject. The baseline was calculated by running the model with only the length of sentences as input feature.

III. RESULTS

Table I shows a summarized representation of the correlation results between speech frequency features and linguistic features. Linguistic features are grouped according to their function and the linguistic aspect they describe. We report the percentage of subjects for which the features in the group were significantly correlated with acoustic features; when two percentages are presented, they indicate the minimum and the maximum number of subjects for which the different linguistic features of the group were significant. Overall, linguistic features within the same group were significant for a similar or the same number of subjects. As expected, acoustic features that reflect the length of the sentences (Mean and Signal Duration) were always correlated with linguistic features that encode aspects of sentence length, for most subjects. We found significant correlations for a high number of subjects for F0 and F3 and many linguistic aspects, while F1 and F2 were the least correlated with linguistic features. The highest correlations were found with features regarding subordination phenomena and the structure of the parsed tree, especially for F3, with up to 70% of subjects showing a significant correlation. Most linguistic features that show significant correlations are related to different aspects of language complexity, such as the length of sentences, syntactic structures (e.g., longer dependency links), or the verbal morphology (e.g., a past verbal tense may be perceived as more complex than the present tense). In Table II, we report the results for the prediction of the acoustic features using the SVR model.

TABLE I
CORRELATION SUMMARY RESULTS

Type of feature	F0	F1	F2	F3	Mean Duration	Signal Duration
raw text properties						
number of tokens	27%	3%	3%	61%	70%	97%
inflectional morphology						
auxiliary and verb form	21-33%	6-12%	<6%	33-39%	45-52%	97%
auxiliary and verb mood	3-33%	<12%	<6%	18-39%	33-61%	97%
auxiliary and verb person	27-36%	9-12%	<6%	36-46%	52-73%	97%
auxiliary and verb tense	9-33%	<12%	<3%	30-39%	45-61%	97%
parsed tree structure						
syntactic length links	33%	12%	3%	42%	61-64%	97%
subordinates chains	49-55%	27-33%	15-18%	67-70%	88-94%	97%
preposition distribution	49-52%	27-30%	15-18%	67-70%	88-91%	97%
pre- post- verbal object	46-49%	27%	15%	61-67%	88%	97%
pre- post- verbal subject	46%	21-27%	15%	61%	88%	97%
syntactic relations						
dependencies dist.	33-46%	12-24%	3-12%	39-67%	61-88%	97%
subordination						
embedded subordin. dist.	52-55%	27-30%	15-18%	67-70%	91-94%	97%
pre- post- verbal subord.	55-61%	30%	21%	67-70%	94%	97%
principals and subord.	55%	33%	15-21%	67-70%	94%	97%
verb edges	61-64%	30-36%	21-24%	70%	94%	97%
verb head and root	33%	12%	3-9%	42-46%	61-73%	96%

TABLE II
REGRESSION RESULTS FOR THE PREDICTION OF LINGUISTIC FEATURES

	% significant subjects	*mean correlation*	*correlation variance*	*baseline*
F0	15%	**0.4032**	0.0027	0.3622
F1	61%	**0.5419**	0.0181	-0.0272
F2	97%	**0.5424**	0.0089	0.0524
F3	27%	**0.4593**	0.0061	0.3264
Mean duration	91%	**0.5836**	0.0123	0.4399
Signal duration	100%	**0.9559**	0.0008	0.9447

To evaluate the goodness of the model, we correlated the model's predictions with the actual values of the features that we predicted, calculating the mean Spearman's correlation and its variance over all subjects. Percentages show the number of subjects for which the predictions were significantly correlated. Our predicting model always performed better than the baseline. The robustness of the model is confirmed by the low variance, indicating that the acoustic values predicted are consistent among the different subjects. The prediction of mean and signal duration was significant for almost every subject. This was expected, as these features are directly linked to the length of the sentences, a feature that the model could see in input. The predictions of F1 and F2 were significant for many subjects (>60%). Contrary to what was seen previously in the correlation analysis, where F0 and F3 were obtained significant results for a high number of subjects, when predicting them with the SVR their predictions are significant for a low number of subjects.

IV. Discussion and conclusion

In this preliminary study, we combine the analysis of the linguistic profile of neutral and emotional texts with

the speech analysis of the reader. We assumed that the speech signal reflected the emotional state induced by the task and assessed by the SAM. Correlation and regression methods were used to understand how the linguistic profile and structure of the texts interact with speech production. We found a statistically significant relationship between some of the linguistic properties of the text, regarding their syntactic structure, subordination phenomena within the texts and the verbal predicate structure, and the speech features that describes some prosodic aspects of speech often related to the human emotional state (e.g., F0, F3). This could suggest a double possible interpretation: on the one hand, it could suggest that the linguistic structure of the pronounced sentence may be a confounding factor that masks the actual contribution of prosodic features in the estimation of the emotional state. On the other hand, the linguistic structure itself could have a direct influence on the emotional state of the subject. This last hypothesis has already been supported by some studies that have combined the features derived from voice processing with some linguistic features to feed classifiers for the recognition of the emotional state [15, 19]. However, in these studies, the encoding of the text considers the lexical and contextual aspects of language but does not consider other important features considered in our study such as morpho-syntactic or syntactic ones. Indeed, these features could have a strong impact on the emotional state of an individual, because they are related to a variety of psycholinguistic phenomena and could affect the cognitive load and processing difficulty of the language user. Future studies will investigate the selected linguistic features to estimate their actual effect on emotional state prediction. Moreover, we will consider other physiological parameters such as electrocardiogram and electrodermal activity recorded during the reading task to evaluate their correlation with voice and linguistic parameters in affective reading.

REFERENCES

[1] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, Discrete time processing of speech signals, 2000.

[2] A. L. Callara, L. Sebastiani, N. Vanello, E. P. Scilingo, and A. Greco, “Parasympathetic-sympathetic causal interactions assessed by time-varying multivariate autoregressive modeling of electrodermal activity and heartrate variability,” IEEE Transactions on Biomedical Engineering, 2021.

[3] Z. Zhang, “Mechanics of human voice production and control.” The Journal of the Acoustical Society of America, vol. 140, no. 4, p. 2614, 2016.

[4] C. S. Hopkins, R. J. Ratley, D. S. Benincasa, and J. J. Grieco, “Evaluation of voice stress analysis technology,” in Proceedings of the 38th annual Hawaii international conference on system sciences. IEEE, 2005, pp. 20b– 20b.

[5] S. G. Koolagudi and K. S. Rao, “Emotion recognition from speech: a review,” International journal of speech technology, vol. 15, no. 2, pp. 99–117, 2012.

[6] N. Cummins, et al, “A review of depression and suicide risk assessment using speech analysis,” Speech Communication, vol. 71, pp. 10–49, 2015.

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

[8] R. Fernandez and R. W. Picard, “Modeling drivers’ speech under stress,” Speech communication, vol. 40, no. 1-2, pp. 145–159, 2003.

[9] A. Guidi, C. Gentili, E. P. Scilingo, and N. Vanello, “Analysis of speech features and personality traits,” Biomedical Signal Processing and Control, vol. 51, pp. 1–7, 2019.

[10] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” Pattern Recognition, vol. 44, no. 3, pp. 572–587, mar 2011. [11] D. Brunato, A. Cimino, F. Dell’Orletta, G. Venturi, and S. Montemagni, “Profiling-ud: a tool for linguistic profiling of texts,” in Proceedings of The 12th Language Resources and Evaluation Conference, 2020, pp. 7145– 7151.

[12] S. Argamon, “Computational register analysis and synthesis,” Register Studies, Forthcoming, 2019.

[13] D. Brunato, et al., “Is this sentence difficult? do you agree?” in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2690–2699.

[14] B. Iavarone, D. Brunato, and F. Dell’Orletta, “Sentence complexity in context,” in Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, 2021, pp. 186–199.

[15] P. Singh, R. Srivastava, K. Rana, and V. Kumar, “A multimodal hierarchical approach to speech emotion recognition from audio and text,” Knowledge-Based Systems, vol. 229, p. 107316, 2021.

[16] S. Ghiasi, G. Valenza, M. S. Morelli, M. Bianchi, E. P. Scilingo, and A. Greco, “The role of haptic stimuli on affective reading: a pilot study,” in 2019 41st Annual EMBC. IEEE, 2019, pp. 4938–4941.

[17] M. S. Morelli, S. Orlandi, and C. Manfredi, “Biovoice: A multipurpose tool for voice analysis,” Biomedical Signal Processing and Control, vol. 64, p. 102302, 2021.

[18] C. Manfredi, L. Bocchi, and G. Cantarella, “A multipurpose user-friendly tool for voice analysis: Application to pathological adult voices,” Biomedical Signal Processing and Control, vol. 4, no. 3, pp. 212–220, jul 2009.

[19] B. T. Atmaja and M. Akagi, “Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using svm,” Speech Communication, vol. 126, pp. 9–21, 2021

SESSION III
NEUROLOGICAL DISORDERS

A LONGITUDINAL STUDY OF VOICE TREMOR IN INTELLECTUALLY IMPAIRED AUTISTIC PERSONS

V. Rodellar-Biarge[1] and M. Jodra-Chuan[2,3]

[1]Neurospeech Laboratory - Center for Biomedical Technology - Universidad Politécnica de Madrid Campus de Montegancedo - Pozuelo de Alarcón -28233 Madrid (Spain)
[2]Asociación Nuevo Horizonte
Comunidad de Madrid 43 - Las Rozas - 28231 Madrid (Spain)
[3]Departamento de Personalidad, Evaluación y Psicología Clínica, Facultad de Educación, Universidad Complutense de Madrid, C/ Rector Royo Villanova 1, 28040 Madrid (Spain)
mariavictoria.rodellar@upm.es - majodra@ucm.es

Abstract: **Autism Spectrum Disorders (ASD) are a group of disorders of neurobiological origin that affect the development of the person, producing alterations in their cognitive process, and in the way they relate to their environment. The diagnosis lies in assessments based on observable behaviors. There is an intensive research to find markers related with biochemical and clinical data, some studies relying on markers found in voice acoustics. Most of these studies are based on the analysis of prosody. Our work follows a different approach and focuses on the analysis of the glottal source tremor parameters that are present also in the phonation of people suffering from neurological diseases such as Parkinson's Disease, among others. This work presents a prospective study of the physiological, neurological, flutter and other tremors in the voice of a man and a woman with autism and intellectual disability through a longitudinal study over a period of 2 years.**

Keywords: **Autism, Intellectual Disability, Tremor in voice, Biomarkers**

I. INTRODUCTION

The Fifth Edition of Diagnostic and Statistical Manual of Mental Disorders [1] from the American Psychological Association encloses various subtypes of pervasive developmental disorders into one category named as Autism Spectrum Disorder (ASD). ASD is a neurobiological disorder that affects social interaction, communication, creativity and imagination of individuals who suffer from it. The severity level of ASD is qualified according to the indicators: intellectual disability (verbal and nonverbal), language alterations (isolated words, sentences, fluent speech), medical and genetic markers and other neurological indicators (mental and behavioral disorders, depression, tics, self-aggressions, sleep and feeding alterations, etc.). An individual to be diagnosed of Intellectual Disability must have an IQ score near to or below 70, and other clinical features, as well as significant impairments in the skills needed to live in an independent and responsible manner, compared to other same-age individuals [1].

ASD affects approximately to 2% of children in the United States [2] and around 1% of children in Europe and the 33% of them suffering also from intellectual disability. ASD appears to affect men 3 to 4 times more than women [3]. The exact causes that may produce the symptom remain still unknown. There is no cure for it, so the daily routines for people suffering ASD must be oriented to ensure a certain level of quality of life.

Some authors highlight a certain parallelism between the cognitive features characteristic of ASD and cognitive aging processes in the general population [4], which points to early cognitive aging in autism [5]. One of the greatest difficulties in diagnosing ASD or early aging in these individuals is the lack of biometric tests. Currently, the diagnosis lies in assessments based on observable behaviors such as gestures, voice or social relationships, among others. Receiving treatment, ideally before the age of three, can greatly improve the development of a child suffering ASD, but difficulties in making an objective and accurate diagnosis may prevent children from receiving early care. Therefore, it is a great challenge to find markers for ASD. Markers would allow making diagnoses based on objective tests, classifying the severity of the syndrome, monitoring the response to a therapy, predicting the evolution of the syndrome, adjusting treatments in the aging stage, etc.

There are different research works in the literature that approach the search for quantitative markers in the voice for ASD [6]. The characteristics and methods used in the studies are very diverse. And it seems that determining a set of parameters validated as markers of ASD has not yet been identified. Most of the works are

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

based on the study of prosody from the production of spontaneous or elicited speech, mainly analyzing the tone, the volume, the duration of the phrases and the silences, and the quality of the voice (jitter, shimmer, harmonic to noise ratio, etc.). Most of the studies are based on groups, whose individuals present heterogeneity in their clinical characteristics and are affected with different degrees of severity, but all of them have developed functional spoken language.

Our work follows a different approach and focuses on the analysis of the glottal source tremor parameters that are present also in the phonation of people suffering from neurological diseases such as Pakinson's Disease, among others [7]. This work presents a prospective study of voice tremor as a marker in people with autism and intellectual disability through a longitudinal follow-up during a period of time of 2 years.

II. METHODS

In the following subsections we will describe the characteristics of the participants, how the voice recording was performed, the signal analysis, and the tools used to obtain the results shown.

A. Participants

We will present a female and a male case from all the participants in our research. The severity of the participants' symptoms has been evaluated with the CARS [8] and DEX [9] tests. CARS test (Childhood Autism Rating Scale) is used to identify the severity of characteristic symptoms of ASD. And DEX (Dysfunction Executive) aims to assess executive dysfunction in daily life. The female case F-ARG 19740123, is 46 years old and, according to the CARS (41) and DEX (51) coefficients, presents severe symptoms of autism and a significant dis-executive impairment. The male case M-RTC19811108, is 40 years old and as the female case also presents severe symptoms of autism and a significant dis-executive impairment (CARS = 41, DEX = 44). They are native Spanish speakers, although they have severe difficulties in maintaining reciprocal social interaction through speech and make it impossible to analyze their prosody but they have the ability to repeat sounds following the instructions of their caregivers. The participants are dependents and are assisted by the Nuevo Horizonte Association. This research study has been approved by that institution with the authorization of the participants' tutors.

B. Data and parameters

The participants were asked to utter a sustained [a:] as long as possible. The key point in obtaining the recordings was the presence of their caretaker, who gave them instructions on how to proceed. All participants collaborated very well and made an effort to perform the exercise according to the instructions given, but the main difficulty in the recordings was obtaining samples of [a:] longer than 2 sec. The voices were recorded in a comfortable and quiet room and in a very relaxed atmosphere. The recording sessions with each participant have a maximum duration of 5 minutes, after this time they show signs of stress and fatigue.

ASD people does not easily endure to wear any external measurement device, they tend to remove and threw them away. So, we have chosen a recording system being the less intrusive to them. The recordings were made with the wireless cardiod microphone/transmitter Sennheiser SK 300 G2A, located 15 cm far from the mouth, a receptor Sennheiser EM 300 G2 and the Adobe Audition Software to manage data acquisition. Data was sampled at 44.1 kHz with 16-bit resolution in uncompressed *vaw* format. Recording sessions contained caregivers' speech, very short vowels, noise, screaming, cluttered signals, etc. In these recordings, the participant's clean vowels are identified, those that have a minimum duration of 500 ms are selected and each of them is saved in separate files.

The first recordings of this study date back to July 2019, and the following session took place in January 2020. The initial idea for this work was to carry out a longitudinal study with 6-month separate samples, but data collection had to be interrupted due to the covid-19 pandemic. Data acquisition could be continued in March 2021, having carried out one recording session per month until August 2021.

Parameter extraction is performed on a 400 ms fragment of the vowel under study. The selected fragment is the one with the greatest phonation stability, that is, less distortion in frequency and amplitude (jitter, shimmer). The parameters evaluated are physiological, neurological and flutter tremor. The physiological tremor lower band is limited to 2.5 Hz, due to the use of 400 ms windows in the calculations. Tremors above the flutter frequency band have been included in a parameter called Global tremor. The results are obtained using the BioMet®Phon tool [10].

III. RESULTS

The results were obtained from voice samples taken on 2019/07/23, 2020/01/29, 2021/03/26, 2021/04/23, 2021/05/28 and 2021/06/25, and they were compared with a normative database [10]. The normative parameter database contains the results of the voices of 50 women and 50 men who do not have any voice pathology or neurological alteration.

The results from the participants in the different recording sessions are shown below. In each of the sessions the number of vowels analyzed was variable, since the number of valid vowels found from the original recordings is not always the same. The p-value of each set of tremors is calculated to check how aligned the result is with the normative reference population, under the null hypothesis of equal means. The results are classified as Normative (N) if their p-value is >0.05, that is, the calculated parameters do not support the rejection of the null hypothesis with the distribution where no voice disorders or neurological alterations have been observed. If the value of p is <0.05, two cases are distinguished, for parameters that are located well above the maximum normative value, they are classified as hyper-normative (H+), and if they are below the minimum normative value as hypo-normative (H-). And finally, to compare the variation of the results among the different days, the parameters that exceed a threshold value, established at 3 times the normalized normative value of each parameter, are indicated.

A. Female F-ARG19740123

Table 1 summarizes the number of vowels analyzed each day, and the type of the results according to the p-value, and Table 2 highlights the parameters that are above the threshold of three times the normalized value.

In the first of the sessions (2019/07/11), only three vowels with a valid duration could be extracted to calculate the tremor parameters, And only the neurological tremor passes the established threshold, as it can be seen in Table 2. The next sampling session took place 6 months later (2020/01/29) and all the parameter results are outside the limits of the norm, and all the tremors pass the threshold. The mobility restrictions of people in Spain, due to the pandemic, forced the interruption of voice recordings for one year and two months, beginning to be retaken again on 2021/03/26. Five vowels from that session were studied, resulting in three sets of parameters close to the norm, and two outside it with opposite behaviors, only the flutter tremor passed the threshold. The rest of the voice samples analyzed are more recent and have been taken with a time difference of one month. In the results of 2021/04/23, there are two normative patterns (N), two hyper-normative (H+) and one hypo-normative pattern (H-). All tremors are over the threshold. The data from 2021/05/28 generate results where most of the parameters are within the norm (4 out of 5), and none of the parameters pass the threshold. In the last analyzed data from the session at 2021/06/25, most of the results are outside the norm, and all the tremors exceed the threshold. In the last two cases, H- type results are not observed.

Table 1. Results for female F-ARG19740123

DATE	N of vowels	(H+)	(H-)	N
2019/07/11	3	1	1	1
2020/01/29	4	2	2	-
2021/03/26	5	1	1	3
2021/04/23	5	2	1	2
2021/05/28	5	1	-	4
2021/06/25	10	6	-	4

Table 2. Tremors above the threshold for female F-ARG19740123

DATE	Physio.	Neuro.	Flutter	Global
2019/07/11		✓		✓
2020/01/29	✓	✓	✓	✓
2021/03/26			✓	
2021/04/23	✓	✓	✓	✓
2021/05/28				
2021/06/25	✓	✓	✓	✓

B. Male M-RTC19811108

The results obtained for the male participant are displayed in the same way as for the female ones, and the voice data has also been recorded on the same days as hers.

Globally analyzing the results of Table 3, which summarizes the number of vowels analyzed and their associated performance, the scarce presence of H- patterns is observed, and the H+ patterns dominate on all the days studied, with the sole exception of 2021/05/28 in which there is a greater number of patterns adjusted to the norm. It can be observed in Table 4 that the tremors exceed the pre-set threshold, with the exception of the flutter and global tremors in the first two days and in the day 2021/05/28 that presents a greater number of normal patterns.

Table 3. Results for male M-RTC19811108

DATE	N of vowels	(H+)	(H-)	N
2019/07/11	3	2	-	1
2020/01/29	2	1	1	-
2021/03/26	7	5	-	2
2021/04/23	11	7	1	3
2021/05/28	5	1	-	4
2021/06/25	8	6	-	2

Table 4. Tremors above the threshold for male M-RTC19811108

DATE	Physio.	Neuro.	Flutter	Global
2019/07/11	✓	✓		
2020/01/29	✓	✓		
2021/03/26	✓	✓	✓	✓
2021/04/23	✓	✓	✓	✓
2021/05/28				
2021/06/25	✓	✓	✓	✓

IV. Discussion

The data acquisition process was not easy due to the particularities presented by the group of people studied. This affects the accuracy of the results, which is not uniform because they are based on a different number of samples per day. Both the female and male participants studied presented similar degrees of autism and executive dysfunction. Both generated normative, hypo-normative and hyper-normative tremor results the same day, which are compatible with phonations corresponding to people older than their age. It is also observed that, when the number of patterns of the results outside the norm ((H +) + (H-)) is greater than the number of patterns in the norm, all the tremor parameters take high values, since they exceed the fixed threshold.

V. Conclusion

In order to make a uniform interpretation of the results, it is necessary to establish a recording protocol that enables the intake of a minimum number of valid vowels per day and per person. The tremor parameters, in a simple first analysis such as the one that has been carried out, seem to be valid markers to study these persons presenting very limited verbal communication. Joint studies of these parameters and the comorbidities associated with ASDs could allow us to understand some of the behaviors of these patients.

REFERENCES

[1] APA, American Psychiatric Association. Diagnostic and statistical manual of mental disorders (5th edition), 2013.

[2] M. J. Maenner, K. A. Shaw, J. Baio, et al., "Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years" - *Autism and Developmental Disabilities Monitoring Network*, 11 Sites, US, 2016. MMWR Surveill. Summ. Vol. 69 (No. SS-4), 2020, pp. 1-12.

[3] Y. Kim, B. Leventhal, Y. Koh, et al., "Prevalence of autism spectrum disorders in a total population sample", *Am. J Psychiatry*, Vol. 168, 2011, pp. 904-912.

[4] A. Roestorf, D. M. Bowler, M. K. Deserno, et al., "Older Adults with ASD: The Consequences of Aging." Insights from a series of special interest group meetings held at the International Society for Autism Research 2016–2017. *Research in Autism Spectrum Disorders 63*, (2019), pp. 3–12.

[5] Bowler, D. M., *Autism spectrum disorder. Psychological theory and research.* Chichester: Wiley, 2007.

[6] R. Fusaroli, A. Lambrechts, D. Bang, et al., "Is voice a marker for autism spectrum disorder? A systematic review and meta-analysis". *Autism Research*. Vol. 10 (3), 2017. pp. 384-407.

[7] J. Schoentgen, A. Kacha and F. Grenez, "Joint Analysis of Vocal Jitter and Tremor in Vowels Sustained by normorphonic and Parkinson Speakers". *Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA): 11th International Workshop,* 2019, pp. 37-40.

[8] E. Schopler, R. J. Reichler and B. Renner, *The Childhood Autism Rating Scale (CARS)*, Los Angeles, CA: Western Psychological Services, 1988.

[9] P. W. Burgess, N. Alderman, B. A. Wilson, et al., (1996). The Dysexecutive Questionnaire (DEX). In B. A. Wilson, N. Alderman, P. W. Burgess, H. Emslie, & J. J. Evans (Eds.), *Behavioral Assessment of the Dysexecutive Syndrome*, Bury St. Edmunds, UK: Thames Valley Test Company, 1996, pp. 18–19.

[10] P. Gómez, V. Rodellar, V. Nieto, et al., "A System to Monitor Phonation Quality in Clinics". *The Fifth International Conference on eHealth, Telemedicine and Social Medicine (eTELEMED),* 2013, pp. 253-258.

ANALYSIS OF CROSS DISORDER SEVERITY PREDICTION PROBLEMS BASED ON SPEECH FEATURES

Gábor Kiss[1], Miklós Gábriel Tulics[1], Attila Zoltán Jenei[1], Dávid Sztahó[1]

[1] Laboratory of Speech Acoustics, Budapest University of Technology and Economics, Department of Telecommunications and Media Informatics, Budapest, Hungary
kiss.gabor@vik.bme.hu, tulics.miklos@vik.bme.hu, jeneia@edu.bme.hu, sztaho.david@vik.bme.hu

Abstract: **In this research, we examined how a model that implements speech-based severity prediction of a given disorder performs in the case of subjects who do not suffer from that specific disorder but another speech-affecting disorder. Recent research claims that speech can be a promising biomarker in support of diagnosis for many diseases. Such tools are likely to appear in medical practice in the near future. However, most research only examines how accurately it is possible to distinguish between healthy and diseased individuals, so it is difficult to estimate how the models created in this way perform for other unknown speech affecting diseases. In the present research, three regression models using the Support Vector Regression machine learning method specific to a particular disease/disorder were created (depression, Parkinson's disease, and dysphonia) and examined how they perform for the other two disease/disorder types. Based on the results, it can be stated that disease-specific models can be used with limited success in supporting general diagnostics, so in the case of tools that can be applied in practice, the outlined problem is waiting to be solved.**

Keywords: **depression, Parkinson's disease, dysphonia, speech processing, regression**

I. INTRODUCTION

Speech may be an appropriate biomarker for many diseases/disorders for which, there is no rapid, non-invasive diagnostic method. An important area of research today is the development of systems to support speech-based diagnostic tools, and based on the results so far, it is possible to recognize depression [1], [2], Parkinson's disease [1], schizophrenia [2], amyotrophic lateral sclerosis [3], and other dysphonic disorders [1], etc. Most often, recognition of a specific disorder is examined among healthy samples [2], [3], in fewer cases, the possibility of distinguishing several disorders at the same time is considered [1]. However, it is not clear how a model trained to recognize severity of a particular diseases/disorders performs on other unknown speech affecting disorders.

For most of the diseases studied, several speech features have been successfully demonstrated, which alters significantly with the effect of a given disease [2], [4], [5], [6]. However, a problem is that typically no speech feature alone has adequate specificity and sensitivity, so we are only able to create high-precision models using machine learning methods. Nonetheless, it is not possible to extract knowledge from such models that would make it easy to estimate how a particular model performs in case of unknown speech altering diseases. Another problem is the lack of a uniform protocol for what speech patterns should be used to create models, which can further complicate the assessment of the overall diagnostic capability of a given model.

Therefore, in the present research, we examine how disease-specific models (depression, Parkinson's disease, dysphonia) perform for other diseases/disorders using the same types of speech patterns. In this way, we want to estimate how difficult the problem outlined above can be.

II. METHODS

A. Database

Three speech databases were used in this study containing Hungarian read speech samples of healthy subjects and subjects suffering from depression (Hungarian Depressed Speech Database - HDSD), Parkinson's disease (Hungarian Parkinson's Speech Database - HPSD) and dysphonic speech (Hungarian Voice Disorder Speech Database - HVSD) [1]. In each database, each person read the same tale of about 1 minute in length and the severity of the given disorder was recorded for each subject in each database.

Beck Depression Inventory-II (BDI) [7] was used for depression severity description. The scale distinguishes 4 categories, 0-13 healthy, 14-19 mild depression, 20-28 moderate depression, and 29-63 major depression. In the database, the mean and standard deviation of the BDI score of depressed individuals was 27.08 (± 8.6), while that of healthy

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

controls was 4.47 (± 3.7). The HDSD database contains speech samples from 131 depressed and 107 healthy individuals, with a mean and standard deviation of the age of the individuals: 41.8 (+ -15.9) years. The age distributions of depressed and healthy individuals were similar.

Hoehn and Yahr (HY) scale [8] was used for Parkinson's disease severity estimation. The scale ranges from 0 to 4, with 0 being the healthy condition and 4 being the most severe. In the database, the mean and standard deviation of the HY score of Parkinson's diseased individuals was 2.62 (± 1.0), while the HY score of healthy controls was always 0. The HPSD database contains speech samples from 79 Parkinson-diseased and 32 healthy individuals, with a mean and standard deviation of the age: 64.8 (+ -9.2) years. The age distributions of Parkinson's diseased and healthy individuals were similar.

RBH scale [9] was used for dysphonia. The scale ranges from 0 to 3, with 0 being the healthy condition and 3 being the most severe. In the database, the mean and standard deviation of the RBH score of dysphonic speech disordered individuals was 1.83 (± 0.8), while the RBH score of healthy controls was always 0. The HVSD database contains speech samples from 245 patients with voice disorders and 193 healthy individuals, with a mean and standard deviation of the age: 51.2 (+ -13.4) years. The age distributions of dysphonic disordered and healthy individuals were similar.

B. Feature Extraction

The calculation of several acoustic-phonetic features requires that speech patterns be segmented at the phoneme level. Segmentation was performed by a forced alignment method [10]. A total of 70 features were extracted from a speech sample.

The following features were calculated: articulation rate; mean, range, standard deviation, and quantiles (1%, 5%, 10%, 25%); of intensity; mean, range, standard deviation, slope and quantiles (1%, 5%, 10%, 25%) of f0; the mean and standard deviation of the formant frequencies (F1 and F2) and their bandwidths (B1 and B2) calculated from the whole recording and only from the vowel E; mean and standard deviation of harmonicity to noise ratio (HNR), jitter, shimmer calculated from the whole recording and only from the vowel E, mean of 12 MFCC coefficients calculated from the whole recording and only from the vowel E, ratio of transients (RoT) [11], pause ratio and soft phonation index (SPI) [12].

C. Training and Testing

Three different regression models (depression model, Parkinson's model and dysphonic model) were trained for each disorder, the training was performed using epsilon Support Vector Regression [13]. The LibSVM implementation [14] was used. We optimized the input feature vector (using fast forward selection), the kernel and hyper parameters (using grid search) for each model using leave-one-out cross validation (LOOCV). In case of feature selection, a maximum of 20 features were specified as the stopping criterion. Linear and rbf kernels were tested, the cost and gamma (in case of rbf kernel) hyperparameters were tested between 2^{-10} and 2^{10}.

Then, each optimized model was tested on the samples from the other two database/disorder types and compared their severity scores with the original ones.

III. RESULTS

In Table 1, the self-evaluation of the optimized models with LOOCV is marked in italics, and these values are in the diagonal. We have marked in bold if the mean predicted score of the given group of the given database fell into the non-healthy category.
From the Table 1, it can be observed that the mean estimated value of each healthy control group falls into the healthy category based on each model, however, the mean score of the HPSD control group predicted by the depression model is close to the border of mild depression (BDI = 14). A possible reason for this may be that the average age of the individuals in the group was over 60.

Table 1. The average prediction results of the three optimized models (depression, Parkinson and dysphonic) on the three databases examined (HVSD, HDSD, HPSD)

		Predicted Scores		
		RBH	BDI	HY
HVSD	Dysphonic	***1.49 (±0.8)***	9.4 (±3.6)	0.73 (±0.6)
	Healthy Control	*0.29 (±0.3)*	10.7 (±4.5)	0.70 (±0.6)
HDSD	Depressed	0.35 (±0.3)	***22.3 (±7.0)***	**1.42 (±0.6)**
	Healthy Control	0.28 (±0.3)	*9.0 (±0.7)*	0.88 (±0.7)
HPSD	Parkinson's diseased	0.64 (±0.5)	**17.9 (±3.1)**	***2.20 (±0.8)***
	Healthy Control	0.66 (±0.2)	13.2 (±3.9)	*0.72 (±0.7)*

It can also be observed that in the case of depressed samples the model indicated on average mild Parkinson's disease, and in the case of patients with Parkinson's disease the average estimated value falls into the category of mild depression. Individuals with depression and Parkinson's disease were on average in the healthy category based on the dysphonic model, while individuals with dysphonia also fell into the healthy category on average based on both depression and Parkinson's models.

IV. DISCUSSION

Based on the results, it can be stated that subjects with dysphonic speech can be considered healthy according to both models that predict the severity of depression and the severity of Parkinson's disease. Their mean severity scores did not show a significant difference compared to the healthy group. Subjects with Parkinson's disease scored significantly higher on the depression scale and were rated as mildly depressed on average. The same can be said for subjects with depression, they scored significantly higher on the Parkinson's scale and were rated by the model as mild severity on the Hoehn and Yahr scale on average. It is important to note that although depression is a common accompanying symptom of Parkinson's disease, the subjects studied did not suffer from depression, and it can be stated with certainty that individuals with depression did not have Parkinson's disease either. Presumably, both depression and Parkinson's disease altered some of speech parameters used for the models in a similar way.

It is important to note that while the age of the subjects in the depressed (HDSD) and dysphonic (HVSD) databases was similar, the mean age in the Parkinson's disease database (HPSD) was considered to be significantly higher. This difference apparently caused a significant difference only in the case of the depression model, since in this case the average estimated value of healthy individuals was close to the limit of mild depression. This may be due to the fact that in the case of the elderly there is a decline in the speech rate, a narrowing of the dynamics of speech, which are typical features of the speech of depressed persons.

V. CONCLUSION

In the present research, we examined how models based on speech signal processing trained to predict specific diseases/disorders (depression, Parkinson's disease, and dysphonia) perform in case of other speech-altering diseases/disorders. The opportunity to conduct the study was provided by the fact that we had three databases (HDSD for depression, HVSD for dysphonia and HPSD for Parkinson's disease) containing speech samples from the same text that are read by individuals with a particular disease/disorder and healthy controls.

Three models were trained based on the three databases using the Support Vector Regression machine learning method. Each model was evaluated on the database used to create it using LOOCV, as well as on the other two databases. During the evaluation, we compared the predicted average scores of the healthy and non-healthy groups based on the scale used by the model.

Analyzing the results, we found that depressed individuals were predicted as mildly Parkinson's on average, while Parkinson's individuals were predicted as mildly depressed on average. In contrast, individuals with dysphonia did not show an average higher score for depressed or Parkinson's models, and the dysphonic model correctly estimated individuals with depression and Parkinson's disease to be non-dysphonic on average.

Another interesting fact was that for elderly controls in the Parkinson's Database (HPSD), the depression model estimated higher scores, although their mean values remained correctly below the border of mild depression.

Based on the research, it can be concluded that in the case of a system that can be used in practice, an important requirement may be not only to be able to distinguish between healthy and specific disorder / disorder group, but it is also necessary to know the error habits of a given model in the case of disorders affect speech and unknown to the given model. This phenomenon must be solved in some form.

ACKNOWLEDGMENT

Project no. K128568 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] D. Sztahó, G. Kiss, M. G. Tulics, B. Hajduska-Dér, and K. Vicsi. "Automatic discrimination of several types of speech pathologies," *IEEE In 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD),* pp. 1-6, 2019.

[2] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review." *Laryngoscope*

Investigative Otolaryngology, vol. 5(1), pp. 96-116, 2020.
[3] A. KwangHoon, M. J. Kim, K. Teplansky, J. R. Green, T. F. Campbell, Y. Yunusova, D. Heitzman, and J. Wang. "Automatic Early Detection of Amyotrophic Lateral Sclerosis from Intelligible Speech Using Convolutional Neural Networks." *In Interspeech*, pp. 1913-1917. 2018.
[4] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri. „A review of depression and suicide risk assessment using speech analysis." *Speech Communication,* vol. 71, pp. 10–49, 2015.
[5] J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, J. Vargas-Bonilla, S. Skodda, J. Rusz, and E. Nöth, "Voiced/unvoiced transitions in speech as a potential bio-marker to detect Parkinson's disease," in *INTERSPEECH*, pp. 95-99, 2015.
[6] Y. Zhang, and J. J. Jiang, "Acoustic analyses of sustained and running voices from patients with laryngeal pathologies" Journal of Voice 22, pp. 1–9, 2008.
[7] A. T. Beck, R. A. Steer, R. Ball, and W. F. Ranieri, "Comparison of beck depression inventories -IA and -II in psychiatric outpatients," *Journal of Personality Assessment,* vol. 67, pp. 588–597, 1996.
[8] M. M. Hoehn, and M. D. Yahr, "Parkinsonism onset, progression, and mortality," *Neurology*, vol. 17, pp. 427–427, 1967.
[9] J. Wendler, A. Rauhut, and H. Kruger. "Classification of voice qualities" in: *Journal of Phonetics*, vol. 14.3-4, pp. 483-488, 1986.
[10] D. Sztahó, G. Kiss, L. Czap, and K. Vicsi, „A computer-assisted prosody pronunciation teaching system.," in: *WOCCI*, pp. 45-49, 2014.
[11] G. Kiss, and K. Vicsi, "Physiological and cognitive status monitoring on the base of acoustic-phonetic speech parameters," in: *International Conference on Statistical Language and Speech Processing*, Springer. pp. 120–131, 2014.
[12] M. G. Tulics, and K. Vicsi, "Phonetic-class based correlation analysis for severity of dysphonia," in: *8th IEEE Conference on Cognitive Infocommunications (CogInfoCom2017),* pp. 21–26, 2017.
[13] C. Cortes, and V. Vapnik, "Support vector machine", *Machine Learning*, vol. 20(3), pp. 273-297, 1995.
[14] C. C. Chang, and C. J. Lin, "LIBSVM : a library for support vector machines." *ACM Transactions on Intelligent Systems and Technology*, vol. 2(3), pp. 1-27, 2011.

SPEECH SIGNAL ANALYSIS AS AN AID TO CLINICAL DIAGNOSIS AND ASSESSMENT OF MENTAL HEALTH DISORDERS

E. Bruno[1,4], L. Weiner[2,3], E. Martz[2,3], A. Greco[4,5], N. Vanello[4,5]

[1]Institute of Computational Linguistic ILC-CNR Pisa, Pisa, Italy
[2]Department of Psychiatry, University Hospital of Strasbourg, Strasbourg, France
[3]Laboratoire De Psychologie Des Cognitions, University of Strasbourg, Strasbourg, France
[4]Dipartimento di Ingegneria dell'Informazione, University of Pisa, Pisa, Italy
[5]Research Center E. Piaggio, University of Pisa, Pisa, Italy

,
ester.bruno@ilc.cnr.it, weiner@unistra.fr, emilie.martz@etu.unistra.fr, alberto.greco@unipi.it, nicola.vanello@unipi.it

Abstract: **In this paper, we propose to use specific speech tasks, along with speech processing and machine learning methods, to support clinical decision in mental health. Specifically, we will focus on classification of relevant Attention Deficit Hyperactivity Disorder (ADHD) subtypes. Both supervised and unsupervised classifiers will be explored. Speech features will be derived from Verbal fluency tests (VFT). The results show good performances of the supervised approach, highlighting the fact that significant information is carried by the speech signal. On the other side, unsupervised classifier results are not in a good agreement with clinician scoring. The results are discussed in the light of possible benefits of developing both approaches within clinical research.**

Keywords: **speech analysis, Attention Deficit Hyperactivity Disorder, Verbal fluency tests, SVM-RFE, k-means**

I. Introduction

The use of speech analysis to support clinical decision is gaining an increasing interest. Since the development of approaches to speech and voice analysis for the evaluation of the emotional and mood state of the speaker[1, 2], applications in mental health research have been proposed [2, 3, 4].
This study focuses on the design of a tool to aid clinicians in the diagnosis and monitoring Attention Deficit Hyperactivity Disorder (ADHD), based on speech analysis and machine learning methods. This tool could help identifying necessary interventions or modulating therapy. ADHD is a developmental and neurological disorder that persists into adulthood for the majority of cases. Three types of ADHD can be identified: predominantly inattentive, predominantly hyperactive-impulsive and combined type. Inattentive subjects have trouble focusing their attention and concentrating. They may not listen well to directions and miss important details, they seem absent-minded and lose track of their things. Hyperactive subjects are fidgety, restless and easily bored. They are constantly in motion, have difficulty performing quiet activities and they often interrupt conversations or others' activities. Individuals with combined-type ADHD display a mixture of all the symptoms outlined above. These deficits in social interactions present a central problem causing social, occupational and emotional disadvantages [5]. Symptoms of ADHD are treated with pharmacological treatments combined with behavioural interventions. The greatest difficulty for clinicians is to identify the ADHD subtypes, given that ADHD is often comorbid with other disorders, such as bipolar disorder. The combined form in particular shares a number of symptoms with bipolar disorder, e.g., mood dysregulation. This makes the differential diagnosis in clinical settings particularly difficult [6]. The identification of ADHD subtypes has important consequences for care due to the different treatment options that can be used according to the clinical presentation of patients.
The analysis of the voice turns out to be an excellent tool to measure he mood dysregulation characteristic of ADHD. Several studies have been performed using voice features to characterize subjects suffer-

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

ing from depression and bipolar disorders [1, 7]. However, very few studies have investigated the utility of voice features for the classification of ADHD subtypes. This is an important first step before checking whether and how voice features could aid the differential diagnosis of ADHD relative to bipolar disorder in clinical settings.
The present study aims at exploring whether speech features can be used to classify ADHD. Starting from speech signals recorded during verbal fluency tests (VFT), speech features were extracted and used to train two classifiers. Both unsupervised and supervised classifiers are proposed. While the former is trained using clinical labels by the physicians, the latter aims at highlighting possible patient grouping from speech without any *a-priori* information. Both approaches could support physicians in formulating a diagnosis and monitor patient status, also when the subject is not hospitalized.

II. Methods

Fifty-five ADHD patients (29 females, age 18 - 57 years, M= 34.94, SD= 11.11), were recruited from inpatient and outpatient psychiatry clinics at the University Hospital of Strasbourg. This study was approved by the Regional Ethics Committee of Eastern France. ADHD diagnosis was established by psychiatrists based on the DIVA 2.0 [8]. The diagnosis was retained if patients present at least 5 inattentive and/or 5 hyperactive symptoms. Among our group, 13 patients were classified as inattentive, only one as hyperactive and 39 were classified as combined.
Voice signals were recorded during verbal fluency tests (VFT) in a quite and low reverberation room using Audacity software (fs=44100 Hz, 24 bit resolution PCM). An high quality microphone was connected to a laptop and kept approximately 60 cm away from the subject. In VFT, patients were instructed to produce words according to specified rules, such as phonemic or semantic criteria, through the continuous association of words following a cue word or simply free word generation in absence of a specified criterion. 260 audio signals were obtained in total belonging to 55 different subjects.
Once all the tasks were recorded prosodic features and spectral features were extracted and investigated. Prosodic features describe how it changes shape during vocalization, such as the speed at which the vocal cords move, called the fundamental frequency (F0), and the energy of the voice..
Features extraction was performed with BioVoice, a multi-purpose software tool developed under Matlab® at the Biomedical Engineering Lab, Firenze University [9]. BioVoice first implements the selection of voiced/unvoiced (V/UV) audio segments and then all the features of interests are extracted from each voiced segment. In the time domain, the number and length of voiced segments, the number and length of pause segments, percentage of voiced segments and other information are extracted. Speech fundamental frequency (F0), formant frequencies (F1, F2, F3), noise level (Normalized Noise Energy) and jitter were estimated. Moreover, statistical descriptor of F0 and first three formants such as mean, median, standard deviation, maximum and minimum values were estimated.
A second set of features describing the prosodic behaviour in each word were estimated. Specifically, this set of features describe the F0 contour using an approach borrowed by Taylor's tilt intonational model [2, 10]. The F0 contour was estimated using the Camacho's SWIPE' algorithm [11]. Spectral features related to the Long Term Average Spectrum, LTAS, were also estimated [12].
Subsequently two classifiers have been trained using the above mentioned features, to distinguish subjects belonging to the two groups. Specifically, both a supervised and unsupervised approach have been tested. While in the supervised approach the classification by the clinicians was used to train the classifier, in the unsupervised approach the clinical labels were only used a-posteriori to assess the classifier agreement with the physician scoring. As a supervised approach, a Support Vector Machine exploiting Recursive Features Selection (SVM-RFE) has been used. SVM-RFE is an algorithm that combines SVMs with a backward variable selection. It selects the most accurate features subset that gives the best classification of the subjects in terms of accuracy [13]. For the evaluation of the algorithm, the leave-one-subject-out (LOSO) cross validation was used, for nearly unbiased estimation of the out-of-sample error.
The second approach was an unsupervised classifier performed using K-means clustering. First dimensionality reduction was performed using PCA. The percentage of total variance explained was used to find the number of components required to explain at least 70% variability. After dimensionality reduction step, K-means clustering was carried out.
Without exploiting *a-priori* knowledge, the K-means clustering performs a partition of the dataset into k predefined distinct non-overlapping subgroups in which each observation belongs to only one group. The criterion is based on a squared euclidean distance. The efficiency of the algorithm was verified by comparing the labels found with the la-

bels provided by the clinicians. The comparison was made by constructing the confusion matrix and summarised by classification accuracy measures such as accuracy, F1 score and Matthews correlation coefficient (MCC). Both supervised and unsupervised method have been applied in order to classify two classes in the ADHD group, namely inattentive and combined patients.

III. Results

A. Unsupervised classifier results

Results in terms of confusion matrix and statistical parameters of the results are shown in Table I and Table V respectively. Inattentive patients were correctly classified in 10 cases and misclassified in 3. Combined patients were classified in agreement with clinicians diagnosis in 17 cases out of 32. The resulting accuracy and F1 score were respectively 0.60 and 0.53, and the MCC was 0.27.

TABLE I: Confusion Matrix from the unsupervised classifier results

		Predicted		
		Inattentive	Combined	Total
Actual	Inattentive	10	3	13
	Combined	15	17	32
	Total	25	20	45

TABLE II: Unsupervised classifier performance scores.

Accuracy	F1score	MCC	Recall	Precision
60%	52.63%	27.41%	76.92%	53.13%

B. Supervised classifier results

The highest accuracy, of about 89%, was achieved with a set of 9 features. Specifically, three Tilt-related features, namely derpos, derneg and Tilt, three features describing LTAS shape, namely $\text{LTAS}_{ratiomax}$, $\text{LTAS}_{ratiomedian}$, slope, and three features from BioVoice, namely Maximum Pause Duration ($\text{PauseDuration}_{max}$), F2_{max},T0 F0_{min}.
Fig. 1 shows the accuracy trend of the SVM-RFE learning algorithm as a function of selected features, which increase in number at each step (according to the RFE algorithm ranking), while in Table III are shown the accuracy values with the various subsets of features. An accuracy of 80% is achieved with the first 5 features. The maximum is reached with a number of features equal to 9.

Results in terms of confusion matrix are shown in Table IV. Among the inattentive patients, a correct classification was achieved in 9 out of 13 patients. Combined presentation of ADHD was correctly classified in 31 cases and misclassified in 1.

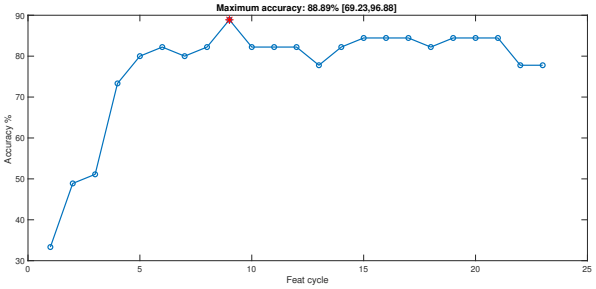


Figure 1: Accuracy trend of the SVM-RFE algorithm as a function of selected features. Inattentive and combined ADHD.

TABLE III: Accuracy values for each subset of features for supervised classifier.

Features subset	Accuracy %
derpos	33.33
derpos, PauseDur_{max}	48.89
derpos, PauseDur_{max}, derneg	51.11
derpos, PauseDur_{max}, derneg, Tilt	73.33
derpos, PauseDur_{max}, derneg, Tilt, F2_{max}	80
derpos, PauseDur_{max}, derneg, Tilt, F2_{max}, $\text{LTAS}_{ratiomax}$	82.22
derpos, PauseDur_{max}, derneg, Tilt, F2_{max}, $\text{LTAS}_{ratiomax}$, $\text{LTAS}_{ratiomed}$	80
derpos, PauseDur_{max}, derneg, Tilt, F2_{max}, $\text{LTAS}_{ratiomax}$, $\text{LTAS}_{ratiomed}$, slope	82.22
derpos, PauseDur_{max}, derneg, Tilt, F2_{max}, $\text{LTAS}_{ratiomax}$, $\text{LTAS}_{ratiomed}$, slope, T0 F0_{min}	88.89

TABLE IV: Confusion Matrix from the supervised classifier results.

		Predicted		
		Inattentive	Combined	Total
Actual	Inattentive	9	4	13
	Combined	1	31	32
	Total	10	35	45

TABLE V: Supervised classifier performance scores.

Accuracy	F1score	MCC	Recall	Precision
88.89%	78.26%	72.1%	69.23%	90%

IV. Discussion

The results obtained with the supervised classifier indicate that speech signals acquired from VFT contain relevant information about ADHD type. Unsupervised classifier results are not in agreement with clinical scoring. In both cases, larger classification error occurred for the combined class. This result is in agreement with the clinician's difficulty in recognizing mixed symptoms.
Given the low number of samples and the high number of features, risk of overfitting was faced in the supervised model. For this reason, SVM with a proper recursive feature elimination scheme was adopted. Specifically, RFE reduces the problem dimensionality by selecting the features which maximize the accuracy, thus mitigating the risk of overfitting [13]. The analysis of classification accuracy, as a function of number of feature, indicates that a lower number of features could be selected to further reduce the

overfitting risk.
Although promising, the results have been found on a low number of subjects, so we have to stress that caution must be taken in generalizing our findings. We believe that the development of both supervised and unsupervised approaches for the classification of ADHD could lead to an improvement of information for the clinicians. The good results by using the supervised classifier seem to indicate that this approach could be used to aid clinician diagnosis. However, supervised classifiers exploit the diagnosis by the physician even if diagnosis might be also prone to classification error. Unsupervised approaches could be less biased by the *a-priori* information by the clinician. Nonetheless, they should be feed by a proper selection of features, possibly obtained using an experimental paradigm able to highlight the differences among the patients. For this reason, physicians should be deeply involved in the critical analysis of automatic or semi-automatic methods results. This could allow identifying possible specific clinical characteristics or pushing the researcher to further explore the speech features of the subjects that were classified both in agreement and disagreement with the clinicians.

V. Conclusion

In this work, a classification of ADHD patients has been carried out exploiting speech signals acquired using VFT. The goal was to identify inattentive and combined subtypes, since the latter need different clinical treatment. The supervised approach allowed to obtain good classification results and a showed a greater ability to classify patients according to clinician diagnosis with respect to the unsupervised classifier. Studies with larger samples are needed to further investigate the relationship between speech features and classification results in ADHD and to mitigate a possible risk of overfitting. Future developments will concern the critical discussion of classification performances of both approaches with the clinicians and the the possible added value of unsupervised learning machine classification. Finally, it could be particularly relevant to determine whether voice features acquired with VFT could aid the distinction between ADHD, especially the combined subtype, and bipolar disorder.

References

[1] K. Scherer, Vocal communication of emotion: A review of research paradigms, Speech Commun. 40 (2003) 227–256.

[2] A. Guidi, N. Vanello, G. Bertschy, C. Gentili, L. Landini, E. Scilingo, Automatic analysis of speech f0 contour for the characterization of mood changes in bipolar patients, Biomedical Signal Processing and Control 17 (2015) 29–37.

[3] M. Cannizzaro, B. Harel, N. Reilly, P. Chappell, P. Snyder, Voice acoustical measurement of the severity of major depression, Brain and Cognition 56 (2004) 30–35.

[4] Å. Nilsonne, J. Sundberg, S. Ternström, A. Askenfelt, Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression., The Journal of the Acoustical Society of America 83 2 (1988) 716–28.

[5] L. Weiner, N. Perroud, S. Weibel, Attention deficit hyperactivity disorder and borderline personality disorder in adults: A review of their links and risks, Neuropsychiatric Disease and Treatment 15 (2019) 3115 – 3129.

[6] E. Martz, G. Bertschy, C. Kraemer, S. Weibel, L. Weiner, Beyond motor hyperactivity: Racing thoughts are an integral symptom of adult attention deficit hyperactivity disorder, Psychiatry Research 301 (2021) 113988.

[7] L. Weiner, A. Guidi, N. Doignon-Camus, A. Giersch, G. Bertschy, N. Vanello, Vocal features obtained through automated methods in verbal fluency tasks can aid the identification of mixed episodes in bipolar disorder, Translational Psychiatry 11 (1) (2021) 415.

[8] J. Kooij, Adult adhd: Diagnostic assessment and treatment, 2012.

[9] M. Morelli, S. Orlandi, C. Manfredi, Biovoice: A multipurpose tool for voice analysis, Biomed. Signal Process. Control. 64 (2021) 102302.

[10] P. Taylor, Analysis and synthesis of intonation using the tilt model., The Journal of the Acoustical Society of America 107 3 (2000) 1697–714.

[11] A. Camacho, J. Harris, A sawtooth waveform inspired pitch estimator for speech and music., The Journal of the Acoustical Society of America 124 3 (2008) 1638–52.

[12] G. Andrea, J. Schoentgen, G. Bertschy, C. Gentili, L. L. Landini, E. Scilingo, N. Vanello, Voice quality in patients suffering from bipolar disease, in: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015, pp. 6106–6109.

[13] K. Yan, D. Zhang, Feature selection and analysis on correlated gas sensor data with recursive feature elimination, Sensors and Actuators B-chemical 212 (2015) 353–363.

MODELING DYSFUNCTIONS IN THE COORDINATION OF VOICE AND SUPRAGLOTTAL ARTICULATION IN NEUROGENIC SPEECH DISORDERS

Bernd J. Kröger

Department of Phoniatrics Pedaudiology and Communication Disorders,
RWTH Aachen University, Aachen, Germany

Abstract:
Neurogenic speech disorders like apraxia of speech or dysarthria show various symptoms in speech and vocalizations. Many of these symptoms can be simulated using a neural model of speech production which includes components for linguistic planning, motor planning, motor programming, and articulatory execution. The execution module (articulatory-acoustic synthesizer) comprises a supra-laryngeal, laryngeal, and sub-laryngeal part and generates normal as well as disordered vocal fold and articulator movements and it generates normal as well as disordered phonation and speech signals.

The concept of gestures as target-directed articulator movements (gestures) is of central importance in our approach. In this paper we concentrate on the simulation of dyscoordinating and of over- and undershooting articulatory and phonatory gestures. The resulting simulated acoustic signals will be compared to natural acoustic signals of normal and disordered speech and vocalizations.

Keywords: **Neurogenic speech disorders, speech gestures, coordination of gestures, articulatory overshoot, articulatory undershoot**

I. INTRODUCTION

Apraxia of speech is defined as a deficit in planning of speech while *dysarthia* is defined as a deficit in motor programming and neuromuscular execution [1]. Both types of speech disorders affect the control of the supra-laryngeal as well as for the laryngeal and sub-laryngeal domain (articulation, phonation, respiration) and these speech disorders affect the segmental level, i.e., lead to distortions of speech sounds, as well as to a distortion of intonation and of syllabic stress patterns. Deficits in speech motor (apraxia of speech) result in deficits in temporal coordination of gestures within and between all three domains (articulation, phonation, respiration) as well as in deficits in correct implementation of the movement target for each single gesture. Planning deficits are mainly due to neural dysfunctions in premotor areas and motor cortex. Deficits in speech motor programming and execution (dysarthria) affect the realization of each gesture by distortions in gesture control and in gesture execution. That leads to imprecise realizations of gestures with respect to gesture duration but mainly with respect to target reaching. Programming, execution, and control deficits are mainly due to neural dysfunctions in motor neurons, basal ganglia, and/or cerebellum.

In this paper we will concentrate on selected symptoms of different speech disorders. Patients suffering from *apraxia of speech* (planning deficits resulting from dysfunctions at different cortical locations) show symptoms like groping, speech sound distortions, articulation errors in producing complex syllables, slow speech rate, and syllable segregation [1]. In case of dysarthria, we need to separate different subtypes [2, 3]. Patients suffering from *ataxic dysartria* (control deficits resulting from cerebellar dysfunctions) show slow and irregular articulatory movement rates and high variability in syllable intensity level. Patient suffering from *flaccid dysarthria* (lower motor neuron damage) show symptoms like breathy voice, short phrases, increased nasal resonance resulting from imperfect closure of the velopharyngeal port and imprecise articulation. *Spastic dysarthria* (bilateral damage of upper motor neurons) leads to symptoms like strained voice and slow articulation resulting from too high muscle tonus. *Hypokinetic dysarthria* (control deficit resulting from basal ganglia dysfunctions) leads to low movement amplitudes while *hyperkinetic dysarthria* (same) leads to involuntary strong and imprecise movements, which not necessarily result from high articulatory effort.

The concept of *speech gestures* [4, 5] allows to explain the speech and voicing symptoms mentioned above by checking the temporal coordination of gestures within a syllable as well as by introducing the idea of gesture target overshoot and gesture target undershoot. Gestures can be defined for the supra-laryngeal system (*vocalic gestures, consonantal gestures,* and *velopharyngeal gestures*, Kröger & Birkholz 2007, p. 181ff) as well as for the laryngeal and sub-laryngeal system. In case of the laryngeal (glottal) system we can differentiate *glottal gestures* controlling vocal fold tension and glottal gestures controlling the positioning of the arytenoids. The later are glottal opening gestures for

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

producing unvoiced speech sounds, glottal closing gestures for producing phonation, and glottal tight closing gestures for producing glottal stop sounds (ibid.). In the case of the sub-laryngeal system, *pulmonary gestures* can be defined. The goal of these gestures is to control subglottal pressure as well as the time span for which a certain degree of subglottal pressure can be hold and for which a certain amount of airflow can be guaranteed to enable phonation as well as secondary sound source excitation.

Gestures always define *target-directed articulator movements*. The goal of each gesture is to reach an acoustically or perceptually relevant target state. In case of articulatory gestures, the target defines a spatial positioning of articulators within the vocal tract, e.g., for reaching vocalic tract shapes or for reaching consonantal constrictions or closures. In the case of glottal gestures, a target is defined as the positioning of the vocal folds or as a certain degree of vocal fold tension. In the case of pulmonary gestures, a target is defined dynamically as the dynamic change in lung volume which leads to the generation of a specific level of subglottal pressure.

II. METHODS

A. Description of the model

The model comprises a neural control component and an articulatory-acoustic model (Fig. 1). A compete linguistic description of the utterance, i.e., a narrow phonological transcription (linguistic input) is transformed into a gesture score (motor plan). The specification of the gesture score, i.e., the temporal coordination of all gestures of an utterance is called *motor planning* which takes place in the premotor area of the brain. The specification of each gesture with respect to the resulting neuromuscular activity is called *motor programming* and leads to a specific neural activation pattern for each syllable at the level of the primary motor cortex. The *execution* of gestures or motor programs is performed by the neuromuscular units of all articulators which lead to defined articulator movements for all articulators of all model components, i.e., for the movements controlling the pulmonary system (lung volume), for the movements controlling the vocal fold positioning, and for the movements controlling the lower jaw, tongue body, tongue tip, velum, and lips. A more detailed discussion concerning the separation of motor planning, motor programming and execution can be found in [1].

Movements of many articulators directly result from neuromuscular activations generated by the neural control component. But in the case of the vocal folds the control component only determines the (rest-)positioning of the folds for phonation or for producing voiceless sounds, while the vocal fold vibration patterns is initiated and controlled by aerodynamic states. The same holds for vocal tract articulators like lips, tongue tip and uvula in case of trills (/B/, /r/, and /R/).

While the brain locations for planning and activating motor programs are cortical, and while the execution of motor programs is mainly done via a direct feedforward motor neuron pathway, *somatosensory feedback signals,* i.e., *tactile and proprioceptive signals* are processed by the basal ganglia-thalamus complex as well as by the cerebellum for controlling and for eventually correcting motor programs and for altering motor programs and motor plans in case of articulatory distortions or in case of changes in the production system due to aging or disorders (feedback processing pathway in Fig. 1 and see [1]).

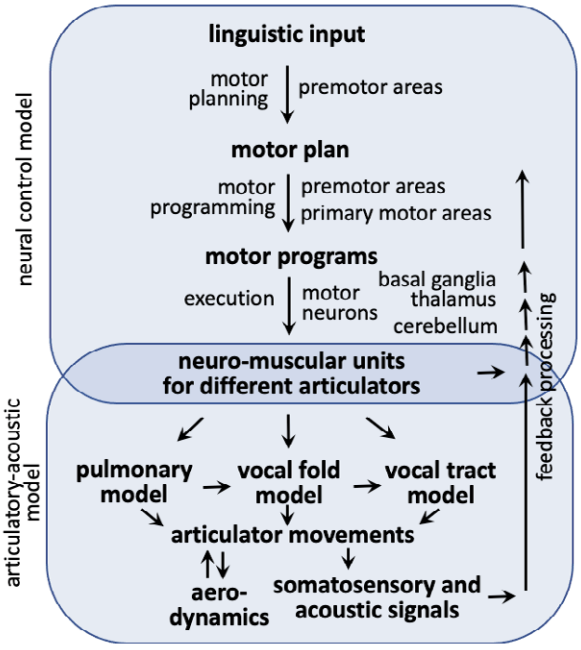


Figure 1: The production model

The articulatory-acoustic model comprises a pulmonary model for generating subglottal pressure and airflow, a self-oscillating vocal fold model for generating vocal fold oscillations for phonation and a vocal tract model for generating vocal tract shapes as function of time. The acoustic glottal source signal in modified in the vocal tract and is radiated it from the lips and nostrils [6, 7].

The motor plan of an utterance is specified as gesture score. A gesture score for the utterance or word (example: [pani]) is given in Fig. 2. The gestures are ordered in six tiers and the gesture targets are named for each gesture: (i) the targets of *vocalic gestures* describes the global form of the vocal tract (global tract form gestures: low tongue body -> /a/; high front location of tongue body -> /i/; high back location of tongue body -> /u/; the labial part of vocalic gestures, i.e., rounded or spread lips, is not displayed in Fig. 2); (ii) the targets of *consonantal gestures* describe the formation of a local

constriction or closure within the vocal tract (local tract constriction gestures: labial closing gestures -> /b/, /p/, /m/; apical closing gestures -> /d/, /t/, /n/, velar closing gestures -> /g/, /k/, /N/; apical near closing gestures -> /s/, /z/, etc.; see [4]).

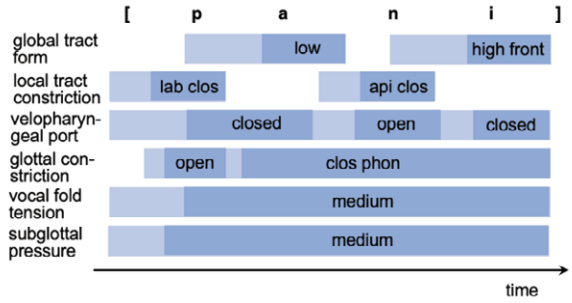


Figure 2: Gesture score of [pani]

(iii) *velopharyngeal opening vs. closing gestures* realize nasal vs. oral speech sounds (tight velopharyngeal closure in case of obstruents, i.e., in case of plosives and fricatives for guaranteeing a pressure built-up in the oral cavity during oral closure); (iv) *glottal opening vs. closing gestures* realize voiceless vs. voiced sounds (tight glottal closure to produce a glottal stop); (v) the target of a *vocal fold tension gesture* defines a F0-target within the intonation contour of an utterance (targets: low, medium, high tension of vocal folds); (vi) the target of a *pulmonary gesture* is holding a specific level of subglottal pressure over the whole time interval of an utterance (targets: low, medium, high in order to realize a soft, normal, or loud voice).

The light blue bars (including the dark blue portions) in Fig. 2 indicate the duration of activation for each gesture. The light blue time interval marks the *movement phase* of a gesture, while the dark blue time interval marks the period in which the gesture reached its target (*target phase*). In the case of vocalic tract-shaping gestures the movement phase is mainly hidden behind a local consonantal tract constriction. In the case of consonantal tract constriction gestures the movement phase occurs within the target phases of vowels and thus allows the perception of place of articulation by the appearance of audible formant transitions.

The gesture targets define (i) the main characteristics of the speech sounds like vocalic formant pattern (vocalic gestures), manner and place of articulation (consonantal gestures), nasal or oral realization of a speech sound (velopharyngeal gestures), voiced or unvoiced realization of a speech sound (glottal gestures), or they define (ii) important supraglottal features of an utterance like current F0-level (vocal fold tension gesture), current loudness or stress level (pulmonary gesture).

A *normal realization, an undershot, overshoot, and a corrected overshoot realization of a gesture* is shown in Fig. 3. Target over- or undershoot can be defined if gesture targets have spatial dimensions (vocalic tract shapes, consonantal closures or constrictions, degree of opening of velopharyngeal port or of glottal constriction) or if targets are defined in the acoustic or aerodynamic domain as frequency value or as pressure level. Target overshoot can be corrected during gesture execution by reversing the movement direction at a certain point in time (Fig. 3, bottom). In case of undershoot the duration of gesture activation interval (of gesture movement phase) needs to be extended or the articulator velocity must be increased (not shown in Fig. 2).

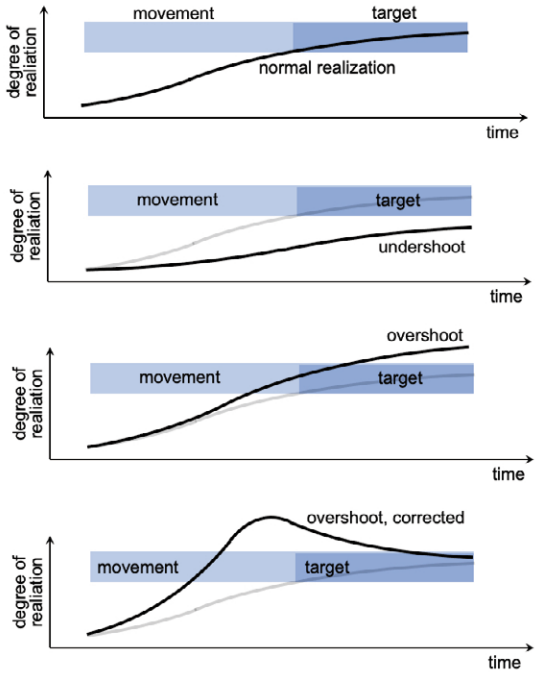


Figure 3: normal gesture, undershoot gesture, and corrected overshoot gesture

B. Simulation experiments

Five types of simulation experiment are executed: Simulation of undershoot and overshoot in case of (i) phonatory gestures (glottal and/or pulmonary gestures), (ii) vocalic gestures, (iii) consonantal gestures, and (iv) velopharyngeal gestures. Simulation of dyscoordination for (v) glottal relative to consonantal gestures.

III. RESULTS

Qualitative results for over- and undershoot of single gestures as generated by our simulation model are listed here for different types of gestures: (i) *Glottal gestures*: Undershoot and overshoot in glottal adduction (rest position of vocal folds for phonation is too wide or too narrow) was studied in the context of simple vocalic syllables like a sustained [a::]. Undershoot (rest position

is too wide) perceptually leads to breathy voice quality. Overshoot in glottal adduction (vocal folds are strongly adducted; high medial compression) leads to harsh and strained voice quality and phonation may stop. Thus, overshoot in glottal adduction gestures forces the model to overshoot the pulmonary gesture (increase subglottal pressure) in order to maintain phonation. (ii) *Vocalic gestures*: Undershoot and overshoot was studied in babbling sequences like [bababa], and [sasasa]. Undershoot results perceptually in a too central schwa-like vowel quality. Speech sounds effortless and underarticulated. In contrast, overshoot in our model leads to static and less coarticulated speech but all vowels sound clearly articulated. (iii) *Consonantal gestures*: Undershoot and overshoot was studied in the same babbling sequeunces (see above). Undershoot leads to short and imprecise productions of consonants. In few cases no closure or constriction is produced and the consonant is acoustically not present. Overshoot leads to very long constrictions or closures. Speech now sounds overarticulated. (iv) *Velopharyngeal gestures*: Undershoot and overshoot was studied in the same babbling phrases (see above). Undershoot perceptually leads to nasalized speech. Plosives and fricatives are acoustically less present, because the pressure built-up in the oral cavity is imperfect.

One experiment was conducted to study (v) *dyscoordination of consonantal and phonatory gestures* in the case of the syllable [ba]. In normal speech a phonatory gesture (glottal closing gesture) reaches its target region synchronously with the vowel (see syllable [pa] in Fig. 2: the target phase of the phonatory gesture (clos phon) starts after consonantal release of [p]). But in the case of a preceding voiced consonant (e.g., [ba]), the phonatory gesture reaches its target region earlier: normally during consonantal closure. If the glottal gesture in coordination with a pulmonary gesture now is shifted to even more earlier points in time, we get an inadequate *pre-phonation effect*, which can be transcribed as [@ba].

IV. DISCUSSION AND CONCLUSIONS

A first qualitative auditory evaluation of synthesized samples of over- and undershoot for different types of gestures as well as of temporal dyscoordination of articulatory and phonatory gestures allows an association of some of these mechanisms with types of neurogenic speech disorders. (i) Pre-phonation resulting from dysfunctions in temporal coordination of articulatory and phonatory gestures occurs in apraxia of speech. (ii) Undershoot of gestures leading to soft speech, monotonous intonation, and reduced intelligibility of speech sounds occur in hypokinetic speech. (iii) It is difficult to associate overshoot phenomena synthesized in our model with hyperkinetic speech samples. More research is needed here. (iv) It is difficult to associate under- or overshoot phenomena with ataxic dysarthria. Complex syllables often are suppressed (produced fast and slurred) in natural data while simple syllables are articulated in a normal way. That results from articulatory reorganization affecting the whole motor plan of a syllable. (v) The same applies to spastic dysarthria. If a gesture target cannot be reached in its normal time interval because movements are too slow, reorganization of the motor plan takes place and lead to an increase in duration of the movement phase of gestures and subsequently to an increase in syllable durations. (vi) In contrast, in case of flaccid dysarthria the patient does not try to reach targets because of his experience about his motor constraints (his inabilities in target reaching). The patient stays with gesture undershoot.

While this preliminary study shows the capability of our model in explaining some basic types of articulatory and phonatory settings occurring in different types of neurogenic speech disorders, a more detailed evaluation of the generated speech samples is needed for a more detailed comparison with natural speech samples.

REFERENCES

[1] A. van der Merwe, A., "New perspectives on speech motor planning and programming in the context of the four- level model and its implications for understanding the pathophysiology underlying apraxia of speech and other motor speech disorders," *Aphasiology*, vol. 35, pp. 397-423, 2021.

[2] R.D. Kent, J.F. Kent, J.R. Duffy, J.E. Thomas, G. Weismer, and S. Stuntebeck, "Ataxic Dysartria," *J. Speech Lang. Hear. Res.*, vol. 43, pp. 1275-1289, 2000.

[3] F.L. Darley, A.E. Aronson, and J.R. Brown, "Differential diagnostic patterns of dysarthria," *J. Speech Hear. Res.*, vol. 12, pp. 246-269, 1969.

[4] B.J. Kröger, and P. Birkholz, "A gesture-based concept for speech movement control in articulatory speech synthesis," in *Verbal and Nonverbal Communication Behaviours* LNAI 4775, A. Esposito, M. Faundez-Zanuy, E. Keller, and M. Marinaro Eds. Berlin, Heidelberg: Springer, 2007, pp. 174-189.

[5] C.P. Browman, and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, pp. 155-180, 1992.

[6] B.J. Kröger, T. Bekolay, and C. Eliasmith, "Modeling speech production using the Neural Engineering Framework," *Proceedings of CogInfoCom 2014* Vetri sul Mare, Italy, 2014, pp. 203-208.

[7] B.J. Kröger, "On the production mechanisms of the singer's formant," *Proceedings of the 23rd International Congress on Acoustics* Aachen, Germany, 2019, pp. 4568-4575.

ANALYSIS OF VOCAL PATTERNS AS A DIAGNOSTIC TOOL IN PATIENTS WITH GENETIC SYNDROMES

Lorenzo Frassineti[1,2], Alice Zucconi[3], Federico Calà[4], Elisabetta Sforza[3], Roberta Onesimo[3], Chiara Leoni[3], Mario Rigante[3], Claudia Manfredi[1*] and Giuseppe Zampino [3,5*]
[1] Department of Information Engineering, Università degli Studi di Firenze, Firenze, Italy
[2] Department of Medical Biotechnologies, Università degli Studi di Siena, Siena, Italy
[3] Fondazione Policlinico Universitario A. Gemelli IRCCS, Roma Italy
[4] School of Engineering, Università degli Studi di Firenze, Firenze, Italy
[5] Università Cattolica del Sacro Cuore, Roma, Italy
*Claudia Manfredi and Giuseppe Zampino jointly coordinated this work
lorenzo.frassineti@student.unisi.it, ali.zucconi@gmail.com, federico.cala@stud.unifi.it, giuseppe.zampino@unicatt.it, claudia.manfredi@unifi.it

Abstract: **Acoustical analysis is widely used in the diagnosis of speech disorders related to several pathologies and helps in defining the severity of their clinical pictures. Recently it was proved that some genetic syndromes may have a specific language phenotype. In this work we apply acoustical analysis to the discrimination between four genetic syndromes: Down, Noonan, Costello and Smith-Magenis. The analysis is performed with Praat and BioVoice tools. Several estimated acoustical features are applied as input to machine-learning models. Though preliminary, the results are encouraging: the acoustical analysis of the sustained vowel /a/ give an average accuracy > 50% with both tools. Our findings confirm that for some syndromes a specific "vocal phenotype" exists that might support the clinician in highlighting syndrome's characteristics not yet studied.**
Keywords: **Language Phenotype, BioVoice, Praat, Genetic Syndrome, Costello Syndrome, Noonan Syndrome, Smith Magenis Syndrome, Down Syndrome.**

I. INTRODUCTION

Genetic syndromes have been extensively studied for a better definition of their clinical manifestation, natural history and etiopathogenetic mechanisms. Nevertheless, some relevant but still unexplored aspects of these multisystemic conditions are not yet fully exploited, one of them being the characterization of vocal production. Genetic factors play a pivotal role not only in the determination of distinct phenotypes and neurobehavioral profiles, but also in establishing voice patterns with recognizable sound characteristics. Therefore, perceptual and acoustical analysis of voice could be helpful for the evaluation of specific voice characteristics as a non-invasive approach to the assessment of genetic syndromes [1]. More than 240 genetic syndromes have distinctive abnormalities of voice quality, significant enough to be considered as diagnostic indicators [2]. For some genetic syndromes the existence of a specific language phenotype obtained by acoustical analysis was already discussed in the literature. For example, young subjects affected by Down Syndrome may have differences concerning tremor, biomechanical behaviour and vibration of the vocal folds as compared to normative subjects [3]. For the Smith-Magenis Syndrome, acoustical and biomechanical analysis was recently performed to detect possible differences between pathological subjects and control groups [4]. Also, for the Cornelia de Lange Syndrome, anomalies in speech such as high levels of speech impairment were found [5]. For the Noonan Syndrome some preliminary evaluation was made with acoustical and biomechanical analysis to explore different aspects of the syndrome [6]. These findings might contribute to the differential diagnosis between Noonan Syndrome and some RASopathies [7] that share several aspects with them, such as the Costello Syndrome [8]. Indeed the Costello Syndrome may have specific acoustical characteristics due to the craniofacial anomalies often related to this syndrome that could alter the process of phonation and articulation [9]. Finally, acoustical analysis could be helpful for an early intervention in patients with speech impairments, to improve their communication skills and reduce speech deficits [10]. Based on the above mentioned evidences, some genetic abnormalities of a recognizable phenotype are expected to determine a specific vocal phenotype. Therefore, vocal characterization could represent a useful tool in the diagnostic process and in defining the severity of some clinical pictures [4].

To this aim, machine-learning methods and supervised classifiers are applied here to acoustical parameters estimated with two analysis tools: Praat and BioVoice [13, 14]. Being based on non-invasive and easily administered tests, this approach could be helpful for obtaining additional features useful for diagnosis and for the automatic classification of

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

different syndromes. The paper is organized as follows: in Section II the dataset and machine-learning experiment are described. In Section III the main results obtained are presented. Section IV is devoted to the discussion of results, limits and possible future developments. Conclusions are reported in Section V.

II. Material and Methods

Data were collected at the Università Cattolica del Sacro Cuore, (Roma), Faculty of Medicine and Surgery. Machine-learning methods are applied to several acoustical parameters estimated from the vocal emissions of a set of 72 subjects (36 male and 36 female, age range 4-33 years, mean 14±7 years), affected by 5 different genetic syndromes. Specifically, the dataset consists of: 22 subjects with Down syndrome (DS); 17 with Noonan syndrome (NS); 19 with Costello Syndrome (CS); 10 with Smith-Magenis syndrome (SMS) and 4 with Cornelia de Lange syndrome (CdLS). However, the CdLS syndrome was excluded from the analysis due to the small number of subjects in this class. The vocal samples come from a previous study based on the SIFEL protocol [11], [12]. After a training phase of the subject, the recorded audio files consist of the vowel /a/ sustained for at least 4 seconds. Recordings were obtained using a portable DAT (Digital Audio Tape) in a controlled environment (environmental noise < 40dB), with the microphone set at 15 centimetres from the subject's lips and with an angle of 45°. The sampling rate was 44100 Hz. Moreover, in the same sessions, the Italian word /aiuole/ (flower beds) as well as the vowels /i/, /u/ /o/ and /e/ were recorded. However, in this work we did not perform the acoustical analysis of these data with BioVoice, because some of them were corrupted or no more available. Only the acoustical analysis previously performed by Praat [11, 13] was available. The quasi-stationary central part of each sustained vowel (about 3s of duration) was manually extracted by an expert, disregarding onset and offset [11].

For the acoustical analysis and classification we considered here both the previously collected dataset of parameters estimated with Praat and new estimates obtained with the BioVoice tool [14, 15]. Only the sustained vowel /a/ was considered. With Praat, the following 34 acoustical parameters were taken into account: mean, standard error, coefficient of variation, maximum and minimum of the fundamental frequency F0; Jitter (local, absolute, Relative Average Perturbation, DDP and PPQ5, where PPQ is Period Perturbation Quotient); Shimmer (%, dB, APQ3, APQ5, APQ11, DDA, where APQ is the Amplitude Perturbation Quotient); mean Noise to Harmonic Ratio (NHR); mean Harmonic to Noise Ratio (HNR); the first four formants (F1, F2, F3 and F4); four clinical features: gender, age, weight and body mass index.

With BioVoice we extracted 24 acoustical features. Analysis is performed distinguishing between infants (<14 years) and adults [14] and in the case of adults between male and female. The 24 acoustical parameters from BioVoice are: maximum, minimum, mean, median and standard deviation for F0 and formants F1, F2 and F3; $T0_{min}$ and $T0_{max}$ for F0; jitter; Normalized Noise Energy (NNE). As before, the four clinical features: gender, age, weight and body mass index (BMI) were also included. In a first step, we compared the acoustical parameters in common between BioVoice and Praat. Then, we used those parameters considering separately each syndrome subgroup. All features except gender (0=male, 1=female) were normalized to zero mean and unit variance and the corresponding feature matrix was applied as input to the following supervised classifiers: k-nearest neighbours (KNN), support vector machine (SVM) and ensemble methods (we considered RUSBoost, AdaBoost and Random Forest). These methods are implemented under MATLAB 2020b computing environment [16]. K-fold cross validation (k=5) and Bayesian Optimization were applied for the selection of the hyper-parameters of the models. The optimization was performed considering the highest global Accuracy as validation metric (i.e. the average Accuracy between the four classes). To improve the classifier's performance the ReliefF algorithm [16] was used as feature selection method. During the model selection process we also varied the number of input features for the classifiers. All the experiments were repeated 5 times, to take into account possible variations of the performance due to the random selection of the subjects during cross-validation. We did not find significant differences in the performances (<5% Accuracy). Finally, we performed the same experiment on the Praat dataset, considering also features from the vowels /a/, /i/ and /u/. In this case the features given by the formant ratios between vowels were added (e.g., $F1_{[a]}/F1_{[u]}$) [13]. As said before, this analysis could not be performed with BioVoice due to missing data.

III. Results

Table 1 shows the comparison between Praat and BioVoice concerning the vowel /a/. We used a two-sample t-test with level of significance $\alpha=0.05$. We checked the hypothesis of normality by Shapiro-Wilk Test (level of significance $\alpha=0.05$). Table 2 shows the True Positive Rate (TPR) and the False Negative Rate (FNR) for the four genetic syndromes.

With BioVoice the 10 features obtained for the best model were: $T0_{maxF0}$ /a/, gender, age, median F3 /a/, BMI, min F1 /a/, $T0_{minF0}$ /a/, min F0, jitter and weight. The best model for BioVoice was a KNN with a

Global Accuracy of 53.1%. Instead with Praat the best model was made of 15 features: gender, mean F1 /a/, age, mean F2 /a/, BMI, max F0 /a/, min F0 /a/, weight, mean F0 /a/. median F0 /a/, Shimmer /a/ APQ11, Shimmer /a/ APQ5, Shimmer local /a/, mean F4 /a/, Shimmer /a/ DDA. The best model with Praat was a KNN with 52.9% of Global Accuracy.

The features used after the selection process are listed in descending order according to their relevance.

Table 1 – Vowel /a/ - Comparison between BioVoice and Praat on the 4 syndromes. Statistically significant differences are highlighted in bold.

	Syndrome (p-value)			
Feature	**DS**	**NS**	**CS**	**SMS**
Median F0 /a/	0.91	0.74	0.99	0.77
Mean F0 /a/	0.80	0.80	0.95	0.66
Min F0 /a/	**0.01**	0.05	**p<0.01**	0.13
Max F0 /a/	**p<0.01**	0.44	**0.02**	0,16
Mean F1 /a/	0.55	0.43	0.92	0.56
Mean F2 /a/	**p<0.01**	**p<0.01**	**0.03**	0.11
Mean F3 /a/	**p<0.01**	0.12	0.23	**p<0.01**

Table 2 – Vowel /a/ - Comparison between BioVoice and Praat - Results of k-fold cross validation.

Genetic Syndrome	**BioVoice**		**Praat**	
	TPR	**FNR**	**TPR**	**FNR**
DS	61.9%	38.1%	63.6%	36.4%
NS	26.7%	73.3%	17.6%	82.4%
CS	68.4%	31.6%	73.7%	26.3%
SMS	55.6%	44.4%	40.0%	60.0%

Table 3 shows the results obtained for the four genetic syndromes considering all the available Praat features for vowels /a/, /u/ and /i/.

Table 3 - Vowels /a/, /i/ and /u/ - KNN's Multiclass confusion matrix with Praat parameters. Main diagonal: TPR for each class. Other values: FNR for a single class.

	Predicted Class			
True Class	**DS**	**NS**	**CS**	**SMS**
DS	68.2%	13.6%	18.2%	0%
NS	17.6%	64.7%	17.6%	0%
CS	31.6%	5.3%	63.2%	0%
SMS	20.0%	10.0%	10.0%	60.0%

The best model was a KNN with Global accuracy 64.7%. In this case, the following 15 features were selected: mean F1 /a/, age, gender, formant ratio $F1_{[a]}/F1_{[u]}$, max F0 /a/, mean F2 /a/, Shimmer APQ11 /a/, mean F0 /a/, median F0 /a/, min F0 /a/, Shimmer /a/ (dB), BMI, Shimmer APQ5 /a/, weight, Shimmer /a/ (local).

IV. Discussion

This work presents preliminary results concerning the discrimination among some genetic syndromes: Down Syndrome, Noonan Syndrome, Costello Syndrome and Smith-Magenis Syndrome. The analysis was performed with acoustical parameters estimated on the sustained vowel /a/ with BioVoice and Praat and applying machine-learning models. The aim of this work was the definition of a proper language phenotype able to distinguish the genetic syndromes considered. The results shown in Table 2 and 3 confirm a possible relationship between genetic syndromes and their specific acoustical characteristics. The results obtained with BioVoice and Praat are comparable. Statistical analysis highlights some differences between the two tools as far as the estimation of formants F2 and F3 for some syndromes is concerned (Table 1, p-values <0.05). This might be related to different techniques for formants estimation implemented in the two tools, as discussed in [14]. Moreover, differences between BioVoice and Praat exist concerning F0 max and min. This could be due to different ranges for F0 estimation defined by the two software tools. We remark that with BioVoice the selection of the frequency range for adults (male or female), infants and newborns is automatically made by BioVoice, while Praat requires some skill of the user to manually set the best frequency range. However, the results shown in Table 2 are preliminary, suggesting that the analysis of the vowel /a/ alone might not be enough for defining a vocal phenotype (TPRs<50%). This is confirmed in Table 3, where the acoustical analysis of vowels /i/ and /u/ performed with Praat was added for all the syndromes, giving Accuracy>50%. In particular, the formant ratio $F1_{[a]}/F1_{[u]}$ was classified as one of the most relevant features by the ReliefF algorithm. This result suggests that a multi-vowel analysis might add more information than a single vowel analysis and should be preferred for the characterization of these genetic syndromes. Our results also confirm evidences previously found for some genetic syndromes. Indeed, for DS, NS and SMS acoustical analysis was already proved useful to find differences between pathological and control groups [3, 4, 6]. Table 3 also shows that SMS has the lowest false negative rate (0%), confirming that acoustical analysis can provide characteristics strictly related to the pathology [4]. Our results suggest that acoustical analysis could be useful also for CS. Indeed, as shown in Table 3, the false negative rates between CS and NS were 5.3% and

17.6% respectively, thus acoustical analysis might be useful to discriminate between these two syndromes.

Our results are preliminary and further study is required to confirm them. First, the number of subjects was poor, thus more cases must be recruited especially for SMS and CdLS. Moreover, we did not perform a comparison between pathological subjects and control cases. This will be done in future work, also taking into account previous studies that already presented such differences for some genetic syndromes [3,4,6]. Considering the promising results obtained, further studies will be made to investigate if some of the acoustical features could be specific of a single genetic syndrome. The acoustical analysis of vowels /i/ and /u/ made with the Praat dataset was found useful, therefore we are planning to perform the same analysis with BioVoice on the same recordings, when available, and/or new ones. Another limit of the work presented here is the wide age range of the subjects, also due to the low number of cases in some syndromes (e.g. CdLS or SMS). If other subjects will be available, a more detailed analysis at different age ranges will be made. If successful, acoustical analysis may be included in the process of differential diagnosis as a completely non-invasive approach to detect specific acoustical characteristics related to speech or phonation impairment for several genetic syndromes, along with e.g. the analysis of facial characteristics and expressions [17].

V. Conclusions

The work presented here is a first step towards the analysis and disentangle of the complex mosaics behind the detection of "voice" phenotypes related to some genetic syndromes. Preliminary results suggest that acoustical parameters and supervised classifiers might provide additional information about genetic syndromes through the characterization of voice. Future work will be devoted to the definition of a protocol for data recording and will concern a larger number of subjects and syndromes, as well as different supervised classifiers and feature selection approaches.

Acknowledgements

Partially funded by the Italian Ministry of Foreign Affairs and Scientific Cooperation (MAECI) MX18MO14.

References

[1] Villafuerte-Gonzalez, R., et al., Acoustic analysis of voice in children with cleft palate and velopharyngeal insufficiency. Int J Pediatr. Otorhinolaryngol, 2015. 79(7): 1073-6.

[2] Hamosh, A., et al., Online Mendelian Inheritance in Man (OMIM), a knowledge base of human genes and genetic disorders. Nucleic acids research, 2005. 33(Database issue): D514-D517.

[3] Hidalgo-De la Guía, et al. (2021). Specificities of phonation biomechanics in Down syndrome children. Biomedical Signal Processing and Control, 63, 102219.

[4] Garayzábal-Heinze, E, et al. (2020). Voice characteristics in smith–magenis syndrome: an acoustic study of laryngeal biomechanics. Languages, 5(3), 31.

[5] Moore, M. V. (1970). Speech, hearing, and language in de Lange syndrome. Journal of Speech and Hearing Disorders, 35(1), 66-69.

[6] Lazzaro, G., Zampino G., et al. (2020). Defining language disorders in children and adolescents with Noonan Syndrome. Molecular genetics & genomic medicine, 8(4), e1069.

[7] Myers, A., et al. (2014). Perinatal features of the RASopathies: Noonan syndrome, cardiofaciocutaneous syndrome and Costello syndrome. American journal of medical genetics Part A, 164(11), 2814-2821.

[8] Zampino, G., et al. (1993). Costello syndrome: further clinical delineation, natural history, genetic definition, and nosology. American journal of medical genetics, 47(2), 176-183.

[9] Mori, M., et al. (1996). Elastic fiber degeneration in Costello syndrome. American journal of medical genetics, 61(4), 304-309.

[10] Moura, C. P., et al. (2008). Voice parameters in children with Down syndrome. Journal of Voice, 22(1), 34-42.

[11] Zucconi A., (2018). Analisi della voce dei bambini con sindromi genetiche: verso l'identificazione di un "fonotipo". [Master Thesis] Università Cattolica del Sacro Cuore, Faculty of Medicine and Surgery.

[12] Ricci Maccarini A, et al. Relazione Ufficiale del XXXVI Congresso Nazionale SIFEL. Acta Phon Lat 2002.

[13] Paul Boersma & David Weenink (2018): Praat: doing phonetics by computer [Computer program].Version 6.0.37, retrieved 28 August 2021 from http://www.praat.org/

[14] Morelli, M. S., Orlandi, S., & Manfredi, C. (2021). BioVoice: A multipurpose tool for voice analysis. Biomedical Signal Processing and Control, 64, 102302.

[15] Manfredi, C., et al. (2015). Automatic assessment of acoustic parameters of the singing voice: application to professional western operatic and jazz singers. Journal of Voice, 29(4), 517-e1.

[16] MATLAB and Statistics and Machine Learning Toolbox Release 2020b. The MathWorks, Inc., Natick, Massachusetts, United States.

[17] Bandini, A., ... & Manfredi, C. (2016). Markerless analysis of articulatory movements in patients with Parkinson's disease. Journal of Voice, 30(6), 766-e1.

SESSION IV
BIOMECHANICS

FITTING A BIOMECHANICAL MODEL OF THE FOLDS TO OSCILLATORY PATTERNS WITH AP AND LR ASYMMETRIES OBSERVED IN HIGH SPEED VIDEO DATA

C. Drioli[1], P. Aichinger[2]

[1] Department of Mathematics, Computer Science and Physics, University of Udine, Italy.
[2] Medical University of Vienna, Department of Otorhinolaryngology, Division of Phoniatrics-Logopedics, Vienna, Austria.
carlo.drioli@uniud.it, philipp.aichinger@meduniwien.ac.at

Abstract: **We discuss a numerical biomechanical model based on lumped and distributed elements, able to represent (L-R) asymmetries and anterior-posterior (A-P) phase differences in vocal cord oscillations, and we introduce a pitch-synchronous procedure to fit the model parameters to observed high-speed visual data. The model allows direct control over L-R unbalancing and the amount of phase delay between folds oscillations at the posterior and anterior part of the glottis, and the fitting relies on a cost function built upon a set of glottal area waveform (GAW) parameters extracted from high-speed videoendoscopic data. The pitch-synchronous procedure is assessed by addressing the time-varying tuning of the fundamental frequency of the model, to keep synchronization with the observed oscillation, and the reproduction of GAW parameter trajectories observed in high-speed videoendoscopic data.**

Keywords: **High-speed video analysis, vocal folds dynamical modelling, voice quality characterization, voice disorders.**

I. INTRODUCTION

Voice source analysis based on high-speed video recording of the vocal folds during sustained phonation has become a widespread diagnostic tool, and today a variety of imaging techniques are available, that are able to perform automated tracking and analysis of relevant glottal cues, such as folds edge position or glottal area. Moreover, reliable glottal models of different accuracy and complexity are today available that mimic the underlying dynamics of the folds [1], [2]. Recent research discussing connections between biomechanical modeling of the folds and high-speed videoendoscopic or videokymographic techniques can be found in [3], [4], [5]. This connection becomes even more interesting when considering that several attempts to fit models of fold oscillations to videoendoscopic data have proven successful [6], [7], [8], [9]. In this contribution we discuss the fitting to visual data of a biomechanical model proposed recently, able to reproduce the sagittal phase differences observed in vocal fold oscillations [10], [5]. The proposed fitting procedure is applied to a dynamic glottal source model in which the fold displacement along the vertical and the sagittal dimensions is modelled using delay lines. The fold model in use provides direct control over the amount of phase delay between folds' oscillations at the posterior and anterior part of the glottis, i.e., the sagittal axis, and at the superior and inferior part of the glottis, i.e., the vertical axis. The fitting procedure is assessed by addressing the time-aligned reproduction of GAWs and hemi-GAWs parameters computed from high-speed videoendoscopic data, in which sagittal phase differences are observed.

II. METHOD

In what follows, we first briefly describe the dynamic glottal source model, and then we illustrate the procedure that adapts the model parameters to match the GAW parameters extracted from high speed video data.

A. BIOMECHANICAL MODEL

The vocal cords model consists in a couple of single mass-spring systems, one for each cord, with stiffness k, damping r and mass m, interacting with a flow model component based on Bernoulli's law. The model has been described in details in [5], [10], and here its properties are only briefly recalled. Fold displacement x at the entrance of the glottis is the result of the force due to driving lung pressure and flow contribution at inferior glottal area, whereas the cord displacement along the vertical axis is modeled through a distributed element introducing a delay of the displacement of the fold from the bottom to the top. The propagation of the displacement along the sagittal axis is represented by a propagation line introducing a delay $\tau_{sag}(y)$, y being the sagittal position. The model is sketched in Figure 1.

In this investigation, the model is used to generate oscillatory patterns of the folds, which in turn can be converted into glottal area flow (GAW) patterns. The direct control of delays, masses, and spring parameters allows to obtain various L-R and A-P asymmetric

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

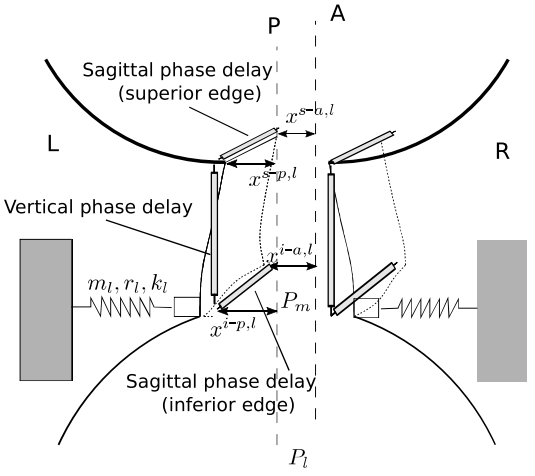


Fig. 1. Schematic view of the model: the vertical (inferior-superior) and sagittal (anterior-posterior) phase differences of the fold displacement are modelled using three propagation lines for each fold.

vibratory patterns, and to provide a compact mean to compare asymmetric patterns, these were characterized by a set of GAW-related parameters. 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 $hGAW^P$) as the time-varying area of the anterior and the posterior half of the glottis that satisfy $hGAW^A + hGAW^P = GAW$. A schematic representation of A-P hGAWs and L-R hGAWs is shown in Fig. 2. In each cycle, the instants corresponding to maximum excursions, i.e., T_c^R, T_c^L, T_c^A, T_c^P, are also defined. Finally, timing differences in the L-R and the A-P direction are defined as $\Delta T_c^{LR} = T_c^R - T_c^L$, and $\Delta T_c^{AP} = T_c^A - T_c^P$.

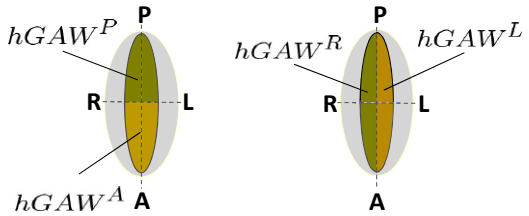


Fig. 2. Schematic representation of A-P hGAWs (left), and L-R hGAWs (right).

B. FITTING PROCEDURE

In [10], we showed that the model discussed so far is able to replicate asymmetry measures derived from the peak analysis of the L-R and A-P hemi-GAWs from HSV data. The analysis was referred to average behaviour with respect to a number of high-speed video frames, corresponding to approximately 10 to 15 glottal pulses, in steady oscillatory conditions. The tuning procedure, based on a LS optimization, was not concerned in maintaining pulse-by-pulse synchronization between the model and the observed oscillatory pattern, nor in matching the observed, possibly time-varying, GAW parameters on a pulse-by-pulse basis. Here, we address instead the problem of tuning the parameters of the biomechanical vocal folds model with a pitch-synchronous algorithm, so that the GAW parameters generated by the model replicate the GAW parameters observed. During the process, the mass-spring system parameters responsible for the edge motion of each fold are adjusted so to synchronize the model to the data, the delays $\tau_{sag}(y)$ are adjusted to reproduce A-P phase differences, and unbalancing of mass-spring system tuning is used to reproduce L-R asymmetries. As the model adopted here provides direct control over the amount of phase delay between folds oscillations at the posterior and anterior side of the glottis while keeping the oscillatory stability, it turns out to be possible to effectively tune this specific class of dynamical model of the folds through iterative search or gradient descent optimization algorithms.

Based on these consideratiosn, a pitch-synchronous parameter optimization scheme was designed, as illustrated in Fig. 3.

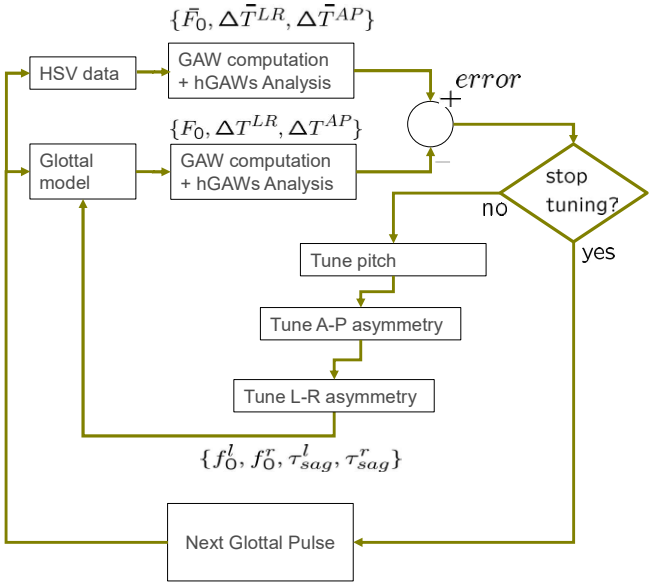


Fig. 3. Automatic pitch-synchronous parametric tuning

At each new glottal pulse, a GAW/hGAW analysis is performed on the HSV frames in the pulse time interval, and similarly the corresponding GAW parameters are computed on the GAW resulting from the glottal model simulation. Based on the error terms

$err_{F0} = (F_0 - \overline{F_0})^2$, $err_{LR} = (\Delta T^{LR} - \overline{\Delta T^{LR}})^2$, and $err_{LR} = (\Delta T^{AP} - \overline{\Delta T^{AP}})^2$, summed into the error criterion $err = err_{F0} + err_{LR} + err_{AP}$, a parameter optimization algorithm performs an iterative search during which the glottal model is required to generate a new version of the glottal pulse based on the parameter set provided by the search process for that target pulse. At each iteration, the error terms related to the three parameters are minimized one after the other through a gradient descent algorithm. When the total error err is considered acceptable, the GAW signal analysis and model tuning is performed on the time window related to the next glottal pulse.

III. FITTING THE MODEL TO OBSERVED PATTERNS: RESULTS

The proposed fitting procedure is assessed by tuning the biomechanical model parameters to replicate the GAW parameters observed in a selection of high speed videoendoscopic data. Note that with this setting, the tuning does not attempt to replicate the oscillatory patterns observed, i.e. the GAW and hGAW shape at each pulse, and an alternative approach would be to perform the fitting on the GAW waveform itself. This will be the subject of future investigations.

A set of recordings from the Laryngeal High-Speed Video Database of Pathological and Non-Pathological Voices described in [11] are used, in which several examples of patterns with A-P phase differences are observed. The following results are referred to a pilot experiment on snippet S_2 of the dataset, in which a the vibration pattern is characterized by L-R differences, as well as A-P phase differences. The fundamental frequency of oscillation is approximately 120 Hz.

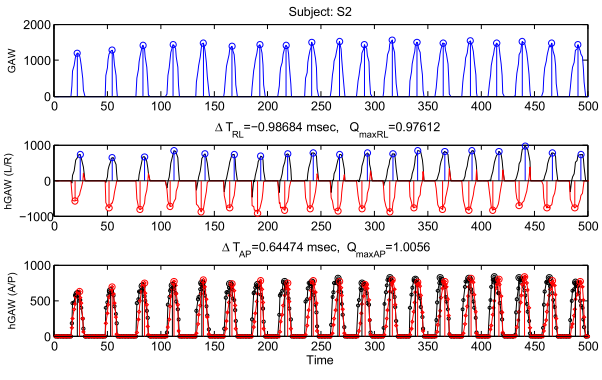


Fig. 4. The GAW analysis on snippet S_2.

Figure 4 shows the GAW analysis on the dataset snippet S_2, and Figure 5 shows the GAW generated by the model when tuning is performed on the oscillation frequency of the folds to match the GAW pulses timing.

The performance of the tuning is presented in terms of the root mean squared difference (RMSE) between

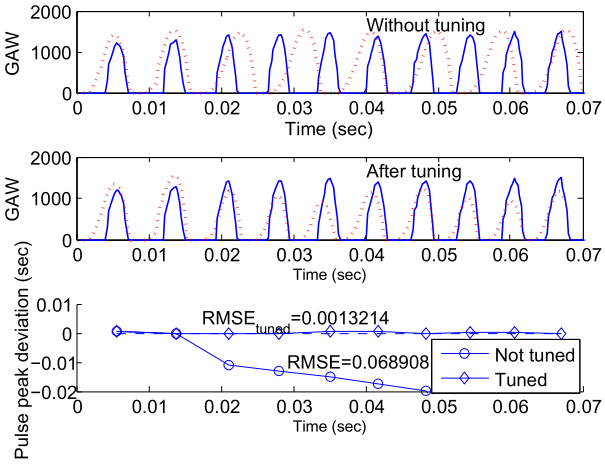


Fig. 5. Tuning of the oscillation frequency of the model. Upper plot: numerical simulation with empirical setting and no tuning; Center plot: simulation with pitch-synchronous tuning; lower plot: deviation of the pulse peaks in the two cases, and RMSE error.

the glottal area waveform (GAW) parameters computed from the high-speed videoendoscopic data and the GAW parameters computed form the vocal fold model simulation after the fitting.

In Figures 6 and 7, the effect of tuning the unbalancing parameter and the sagittal phase delay is illustrated. The average error (RMSE) computed on 10 GAW pulses is also provided in the plots.

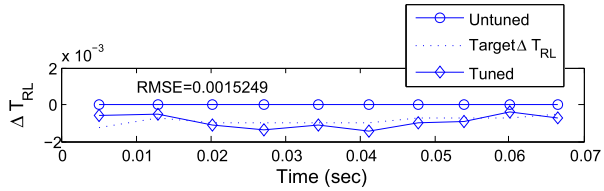


Fig. 6. Tuning of the oscillation frequency of the model. Deviation of the pulse peaks in the two cases, and RMSE error.

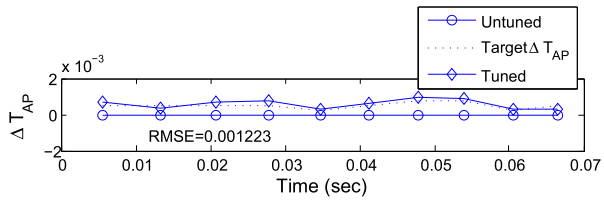


Fig. 7. Tuning of the oscillation frequency of the model. Deviation of the pulse peaks in the two cases, and RMSE error.

It can be seen from these examples how the tuning of the model permits to adapt the characteristics of each single pulse with respect to the desired GAW tuning, unbalancing, and AP phase delay properties.

IV. CONCLUSIONS

We discussed a pitch-synchronous adaptation procedure that allows to fit a lumped and distributed-elements vocal fold model to vocal folds oscillatory cues. Specifically, we addressed the reproduction of the GAW and hGAW parameters computed from oscillatory patterns observed in high-speed video recordings of the folds, including vertical and longitudinal phase differences and left-right fold mass unbalancing. The procedure was assessed by numerical simulations and parameter tuning on a small set of recorded HSVs, and the results referred to a selected snippet are shown. These include asymmetry measures derived from the peak analysis of the L-R and A-P hemi-GAWs and compared to those obtained from the HSV data. The comparisons suggest that it is possible to automatically tune the model parameters and to reproduce L-R asymmetries and A-P phase differences. These differences will be achieved by the procedure through left and right mass unbalancing, and longitudinal propagation delay tuning.

Future work is foreseen to enhance the cost function of the fitting procedure. This is desirable in order to design the fitting focusing not only on the GAW parameters observed but also on the GAW and hGAW shape at each pulse, so to address the fitting of the GAW waveform itself.

V. ACKNOWLEDGEMENTS

This work was supported by the Austrian Science Fund (FWF): KLI 722-B30.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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

[5] C. Drioli and P. Aichinger, "Modelling Longitudinal Phase Differences in a Lumped and Distributed Elements Vocal Fold Model," in *Proc. Conf. Mod. Anal. Voc. Emiss. Biomed. App. (MAVEBA)*, Firenze, Italy, 2019.

[6] R. Schwarz, M. Döllinger, T. Wurzbacher, U. Eysholdt, and J. Lohscheller, "Spatio-temporal quantification of vocal fold vibrations using high-speed videoendoscopy and a biomechanical model," *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 2717–2732, 2008.

[7] V. Devaraj and P. Aichinger, "Modelling of amplitude modulated vocal fry glottal area waveforms using an analysis-by-synthesis approach," *Applied Sciences*, vol. 11, no. 5, 2021.

[8] C. Drioli and G. L. Foresti, "Fitting a biomechanical model of the folds to high-speed video data through bayesian estimation," *Informatics in Medicine Unlocked*, vol. 20, p. 100373, 2020.

[9] 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.

[10] C. Drioli and P. Aichinger, "Modelling sagittal and vertical phase differences in a lumped and distributed elements vocal fold model," *Biomedical Signal Processing and Control*, vol. 64, p. 102309, 2021.

[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.

ARTIFICIAL HIGH-SPEED VIDEOS OF NORMAL AND DYSPHONIC VOCAL FOLD VIBRATION

P. Aichinger[1], S. P. Kumar[2], H. Lehoux[3], J. G. Švec[3,4]

[1] Division of Phoniatrics-Logopedics, Department of Otorhinolaryngology, Medical University of Vienna, Austria
[2] Centre for Healthcare Technologies, Department of Biomedical Engineering, Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, Chennai, India
[3] Voice Research Laboratory, Department of Experimental Physics, Faculty of Science, Palacky University, Olomouc, Czech Republic
[4] Voice and Hearing Centre Prague, Medical Healthcom, Ltd., Prague, Czech Republic
philipp.aichinger@meduniwien.ac.at, pravinkumars@ssn.edu.in, hugo.lehoux@upol.cz, jan.svec@upol.cz

Abstract: **Laryngeal high-speed videoendoscopy has been recognized as a valuable modality for scientific investigations of vocal fold vibrations. Its advantages over standard clinical stroboscopic imaging include its ability to provide detailed insights into divergences from the cyclicity of the vocal fold vibrations, which are characteristic for a significant subset of dysphonic voices. However, laryngeal high-speed videoendoscopy is not well established in the clinical care of disordered voices, partly because the interpretation of vibration patterns goes beyond the established clinical knowledge acquired from stroboscopic videos. A particular gap of knowledge exists in the understanding of how kinematic vocal fold parameters relate to patterns of vocal fold vibration, and how irregularities look like in high-speed videos. We aim at exploring these aspects using a computer model that takes kinematic parameters as inputs and synthesizes high-speed videos. The presented videos show zipper-like vocal fold vibrations, pressed phonation, voice onset, constant and time-varying left-right and anterior-posterior phase differences, and left-right frequency differences (diplophonia).**

Keywords: **Laryngeal high-speed videos, dysphonia, voice quality characterization**

I. INTRODUCTION

Vocal fold (VF) vibration kinematics reflect vocal health status and are a key element connecting voice physiology with voice acoustics and perception. The clinical standard for visualizing kinematics of VF vibrations is laryngeal stroboscopy, while laryngeal high-speed videolaryngoscopy is more frequently used in detailed scientific research on VF vibrations.

A recently proposed kinematic model for synthesizing single-line kymograms has used sinusoids for a number of surface points of the VFs' frontal cross section [1]. Sinusoids were combined for lateral and vertical movement, which enabled a circular motion in the frontal plane as well as the simulation of the mucosal waves that travel upwards on the medial VF surfaces and continue laterally on the top VF surfaces. The kymogram synthesizer was generalized in the past to be capable of simulating time-constant left-to-right phase differences, and it was validated by fitting simulated to clinical kymograms [2].

We propose synthesizing laryngeal high-speed videos (LHSVs) by stacking a number of artificial kymograms obtained with a further advancement of the synthesizer proposed earlier [1]. We use kymograms that are sampled at 256 equidistant sagittal positions. The synthesizer is advanced here to be capable of more general left-right differences. We also propose a few rules regarding the variation of kinematic parameters across sagittal positions, to simulate anterior-posterior differences. Most notably, we control the vibration at a few spatially separated supporting points (left, right, and anteriorly, midsagittal, posteriorly), while ensuring spatial continuity of the parameters along the sagittal axis during interpolation. As a result, provided control parameters help developing an intuition about their relation to the VF vibration patterns. In addition, these parameters enable the generation of different VF vibration patterns including transients, e.g., voice onset and offset, as well as anterior-posterior and left-right vibration asymmetry in amplitude, frequency, and phase.

II. METHODS

First, the model for generating single-line kymograms proposed earlier is described concisely. More detailed explanations regarding newly proposed left-right differences are also given. Second, we extend the model to enable the generation of artificial LHSVs by stacking a number of single-line kymograms. We

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

enable anterior-posterior differences of kinematic parameters such that the model is capable of producing a few qualitative key features that are observed in LHSVs of normophonia and dysphonia.

A. Kinematic model for single-line kymograms

Figure 1 illustrates the reference geometry of the VFs, the VF vibration, and the model of local illumination introduced in [1]. The reference geometry of the VF surfaces through a frontal cross section is initialized according to the M5 model [3]. The reference geometry is the VFs' contours at zero amplitude of vibration, i.e., the pre-phonatory shape also referred to as the 'previbratory position'. The M5 parameter that we vary in this study is the glottal halfwidth w_h, which reflects the adductory adjustment of the vocal folds. To simulate vibrations, surface points of the VFs are moved circularly. The vibration amplitudes, i.e., the circles' radii, are imposed on the lower and the upper margin separately. At the bottom VF surface, the amplitudes of the points below the upheaval point Z are set to 0. At the top surface, the amplitude decays gradually towards lateral, resulting in damping of the outward travelling mucosal waves. Phase differences between individual surface points are imposed in order to simulate the mucosal waves, i.e., surface waves that start at the subglottal upheaval point Z and travel via the medial surface to the top surface, and laterally from there.

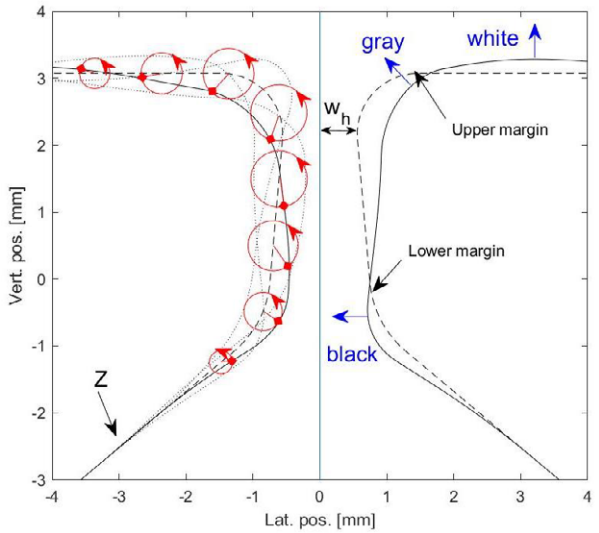


Figure 1: Contours of the vocal folds in a frontal cross section.

Kymographic images are obtained from vibrating VF contours using a local illumination model based on diffuse reflection as proposed in [1]. In particular, the light intensity across the VF surfaces is proportional to the distance to the light source, as well as the slope of the surface, i.e., its declination with regard to the direction of light incidence.

The kinematic VF parameters are allowed to be different for the left and the right VF. First, to simulate diplophonia, i.e., biphonation in which the left and the right VFs vibrate at different rates, we use different vibration frequencies for the left and the right VF. Typically, due to coupling via the common airstream through the glottis as well as collision, the frequencies are small integer multiples of a common cycle frequency, e.g., 3/4, or 4/5 [4].

Second, we distinguish between constant and time-varying phase differences. To simulate a time-constant delay of one VF with regard to the other, a time-constant left-right phase difference is imposed as proposed in [2]. This results in paramedian collision of the vocal folds, i.e., collision occurring at the right or the left of the midline. To simulate irregular VF vibration, a time-variant left-right phase difference is imposed. The phase difference is allowed to vary from one pulse to the next. The variation is imposed by randomly shifting times of individual pulses, as proposed in [5]. Time shifts are white Gaussian random numbers, which differ between the left and the right VFs. This results in a phase distortion and a jitter that differs between the two VFs.

B. Combining single-line kymograms to synthetic laryngoscopic videos

For each artificial LHSV, 256 kymograms generated at equidistant sagittal positions are stacked. The first kymogram is located at the anterior end of the VFs, i.e., the anterior commissure, and the last one at the posterior end, i.e., at the vocal processes. The length of the VFs is 15 mm in all presented simulations. Variation of kinematic parameters across sagittal positions is described as follows.

The posterior glottal halfwidth w_h^{post} controls the opening of the reference glottal opening at the posterior end of the VFs. At the anterior end, the glottal halfwidth is negative to compensate for the upper margin radius of the M5 model, which makes the top surface of the VFs plain where the VFs are connected. At sagittal positions in between the anterior and the posterior ends, the glottal halfwidth is linearly interpolated.

Vibration amplitudes A are controlled separately for the midsagittal position and posterior end, while they are 0 at the anterior end. The vibration amplitude is assumed to be maximal at the midsagittal position, and smaller at the posterior end.

An anterior-posterior phase difference Φ_s enables the generation of waves that travel in a sagittal direction.

We distinguish between constant and time-varying sagittal phase differences, as is done for the left-right phase differences.

Time-constant sagittal phase differences were already simulated in the past using delay lines [6]. Time-varying phase differences enable the simulation of irregular VF vibrations. The phase difference is allowed to vary from one glottal pulse to the next, resulting in a jitter that is different across sagittal positions.

III. RESULTS

Figure 2 shows two video frames of zipper-like VF vibrations and selected control parameters (amplitudes and halfwidth) across sagittal positions. Frames are shown for times of maximal and minimal opening. Zipper-like vibrations come with a chink that is caused by a large posterior halfwidth w_h^{post}. The most anterior point of collision *C* moves back and forth cyclically, like the zipper of a jacket. Also shown are midsagittal and posterior amplitudes of the upper and lower margins (A_u^{mid}, A_l^{mid}, A^{post}).

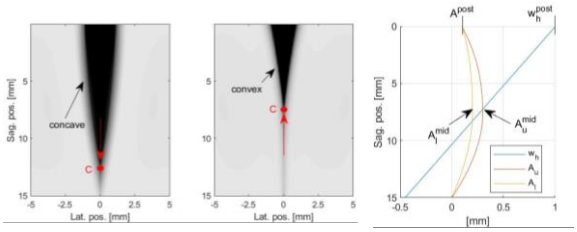


Figure 2: Zipper-like vocal fold vibration.

Figure 3 illustrates a video of pressed phonation, which arises from small halfwidths and large amplitudes. The frames show maximal opening (left), maximal closure (right), and a frame in between (middle). Mucosal waves that travel laterally are visible.

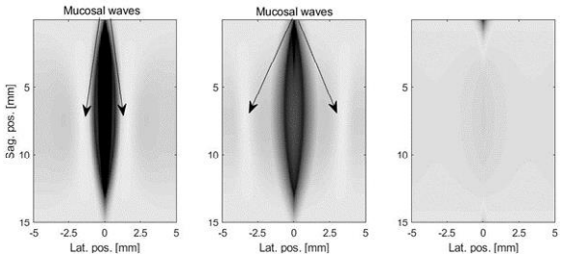


Figure 3: Pressed phonation.

Figure 4 shows selected control parameters and a kymogram of a breathy voice onset. The glottal halfwidth decreases over time while the vibration amplitudes increase smoothly. The upper plot indicates the increase of the posterior halfwidth w_h^{post+} approaching 0, and an amplitude factor A* approaching 1.

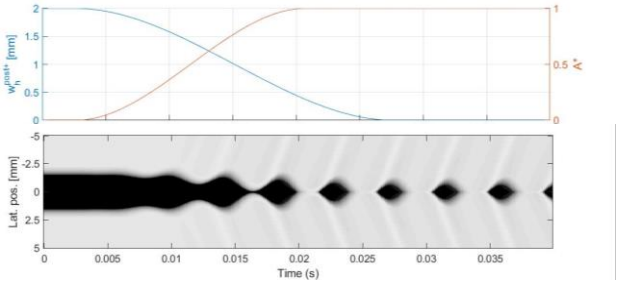


Figure 4: Breathy voice onset.

Figure 5 illustrates time-constant left-right phase difference. The video frames show maximal lateral deflections of the left and the right vocal folds, and a frame in between. The phonovibrogram is also shown.

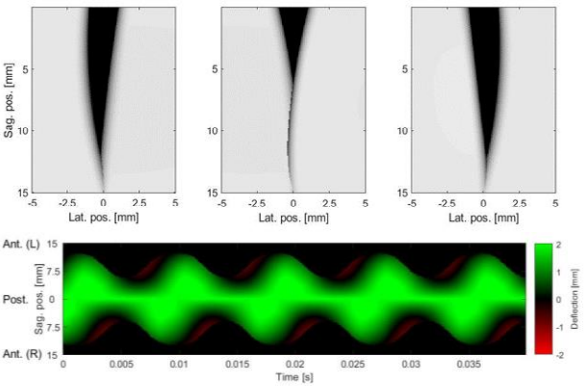


Figure 5: Time-constant left-right phase difference.

Figure 6 illustrates time-varying left-right phase difference. The phase distortion of the left and the right VF are reported individually (top, Φ_L^{dist}, Φ_R^{dist}). In times during which the distortion of the left VF is larger, the left VF's vibration is delayed, and vice versa (L: late, E: early).

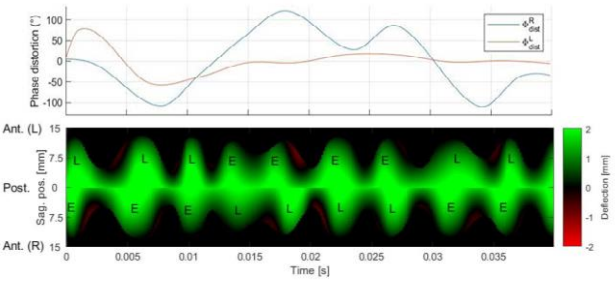


Figure 6: Time-varying left-right phase differences.

Figure 7 illustrates time-constant anterior-posterior phase difference. Three subsequent frames of the open phase are shown. The anterior vibration is delayed with regard to the posterior vibration, resulting in a wave travelling anteriorly.

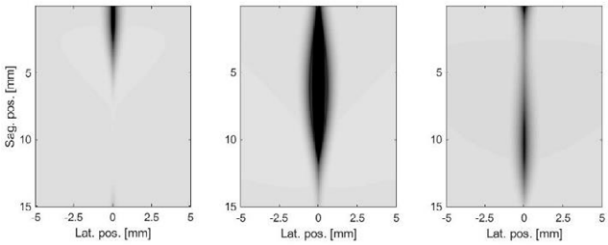


Figure 7: Time-constant anterior-posterior phase differences.

Figure 8 illustrates time-varying anterior-posterior phase differences. Shown are the time-varying phase distortion for the posterior and anterior end (Φ_P^{dist}, Φ_A^{dist}), a multi-line kymogram, and a phonovibrogram. Vertical lines in the kymograms indicate times of maximal midsagittal deflection. Anterior and posterior maxima are early (E), late (L), or in time (I), which may switch from one pulse to the next.

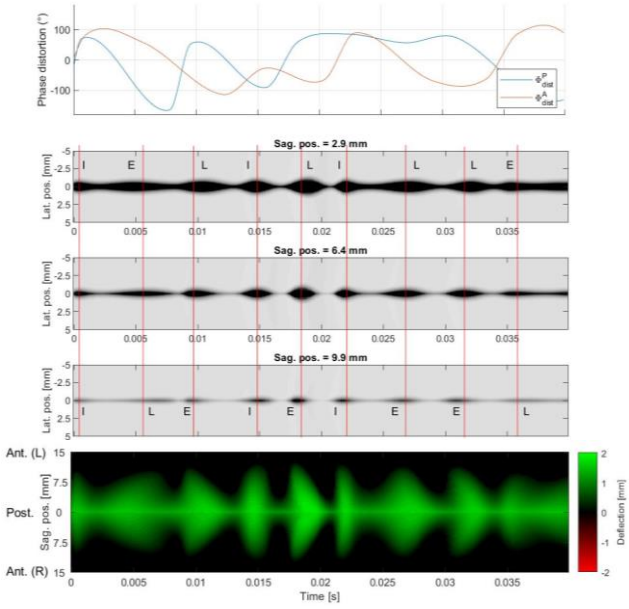


Figure 8: Time-varying anterior-posterior phase differences.

IV. Discussion and Conclusion

A kinematic model of VF vibrations is combined with a model of light reflection, enabling the creation of artificial LHSVs. The model of light reflection and the kinematic model for one frontal cross section proposed in the past is extended to reflect sagittal differences. Artificial LHSVs are generated by stacking several individual kymograms obtained at equidistant sagittal positions. Informal comparisons with clinical LHSVs appear to be promising, given that the vibration patterns observed in the artificial videos look visually similar to patterns seen in clinical videos. Formal comparisons with natural data are subject to future research.

V. Acknowledgements

This work was supported by the Austrian Science Fund (FWF): KLI 722-B30, by the Technology Agency of the Czech Republic project no. TH04010422 "VKG 3.0" and by the student project of the Palacký University Olomouc IGA_PrF_2021_017.

References

[1] S. Kumar and J. Švec, “Kinematic model for simulating mucosal wave phenomena on vocal folds,” Biomed. Signal Process. Control, vol. 49, pp. 328–337, 2019.

[2] S. Bulusu, S. Kumar, J. Švec, and P. Aichinger, “Fitting synthetic to clinical kymographic images for deriving kinematic vocal fold parameters: Application to left-right vibratory phase differences,” Biomed. Signal Process. Control, vol. 63, p. 102253, 2021.

[3] R. C. Scherer, D. Shinwari, K. J. De Witt, C. Zhang, B. R. Kucinschi, and A. A. Afjeh, “Intraglottal pressure profiles for a symmetric and oblique glottis with a divergence angle of 10 degrees,” J. Acoust. Soc. Am., vol. 109, no. 4, pp. 1616–1630, 2001

[4] 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, no. 4, pp. 2036–2046, 2015.

[5] P. Aichinger and F. Pernkopf, “Synthesis and Analysis-by-Synthesis of Modulated Diplophonic Glottal Area Waveforms,” IEEE/ACM Trans. Audio Speech Lang. Process., vol. 29, pp. 914–926, 2021.

[6] C. Drioli and P. Aichinger, “Modelling sagittal and vertical phase differences in a lumped and distributed elements vocal fold model,” Biomed. Signal Process. Control, vol. 64, p. 102309, 2021.

COMPARING DIFFERENT VOCAL EFFORT INDEXES TO THE LARYNGEAL TISSUES ACCELERATION

Filippo Sanjust[a], Renata Sisto[a], Teresa Botti[a], Luigi Cerini[a], Raffaele Mariconte[b]

[a] INAIL Research, Department of Occupational and Environmental Medicine, Epidemiology and Hygiene, Monteporzio Catone, Italy.
[b] INAIL Research, Certification and Verification Sector, Department of Technological Innovations and Safety of Plants, Products and Anthropic Settlements, Rome.

Abstract: **This work is aimed at confirming the validity of some of the indicators used to estimate the vocal effort of professional singers and at testing new parameters in the perspective to provide new aids to that workers who have voice as their main working tool. In particular, a specific software was developed for the analysis of wave files obtained from the subjects examined using vocal samples from opera-musical workers. Vocal effort of six professional singers of an Opera Institution has been first evaluated by means of the methodology of the APM (Ambulatory Phonation Monitor) commercial system. The comparison between this method, based on the measure of the acceleration of the laryngeal tissues, and a method based on the analysis of the energy content of the harmonic components of the speech signal permitted a validation of the latter.**

Two different indexes were analyzed, the SPR (Singer Power ratio) and a new proposed indexed based on the analysis of the pitch.

Keywords : **lyric singers, vocal effort, pitch, vibrato.**

I. INTRODUCTION

Professional singers are a category of workers exposed to a high risk of developing professional disorders linked to the so-called vocal effort (VE) due to the prolonged use of their phonatory apparatus. The stress of their vocal tract is constantly exposed to during performances can produce long-term effects ranging from voice quality degradation to severe laryngeal diseases. There are many scientific studies that analyze vocal fatigue in singers and actors similarly to what has already been seen in other categories of workers exposed to vocal effort such as teachers [1-3]. The study on the vocal effort of teachers focused on the analysis of the fundamental frequency (f_0) or pitch, the duration of phonation and the average sound pressure level emitted at a certain distance during daily work [4].

Afterwards, to evaluate the vocal effort of the singers, other parameters were introduced such as the variation of f_0 (such as "vibrato" analysis), its standard deviation, the so-called SPR or Singing Power Ratio and others [5-7]. These parameters, obtained with particular methods of analysis of vocal emission, were then correlated with the psychophysical evaluation reported subjectively by the subjects themselves [8-10].

At present, one of the most interesting parameters well correlated to the laryngeal stress is the so called vocal effort dose. This parameter is mainly quantified by three different components [11, 12]: the time dose, or the percentage of phonation time, the cycle dose and the distance dose. In particular, the time dose represents the total time the vocal cords vibrated during the total time of the speech that was recorded.

The percentage of phonation time is the ratio, expressed as a percentage, between the phonation time and the total recording time.

The main methodology for evaluating VE by means of vocal effort dose is based on the use of particular dosimeters (such as APM, Ambulatory Phonation Monitor) capable of directly recording the energy dissipated by the phonatory apparatus in terms of acceleration of the larynx tissues through a system based on an accelerometric measurement that allows to reconstruct those parameters that quantify the vocal effort

Our aim is to make a comparison between the VE, evaluated by means of APM commercial system, and a method based on the analysis of the energy content of the harmonic components of the speech signal permitting a validation of the latter. This work is necessarily linked to a previous one [13] in which some of the authors had already examined similar parameters such as f_0, f_0 vibrato and the SPR, obtaining some first indications on the goodness of these indicators. Due to the few subjects making up the sample under examination, this work is a preliminary study waiting to be able to perform better statistics based on new and larger samples of subjects.

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,

published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

II. METHODS

Vocal effort of six professional singers of an Italian Opera Institution, the Teatro Regio in Turin, has been evaluated by means of the methodology of the APM commercial system. In particular the instrument used for the vocal effort measurements consisted of two complete Ambulatory Phonation Monitors model APM 3200 (Kay PENTAX), both equipped with model 3200 Ambulatory Phonation Monitor Version 1.4 software required for carrying out the measurement configuration, calibration, download and offline analysis of the data stored by the device.

The APM consists of a data-logging unit, battery powered and wearable in a pouch, connected by cable to an accelerometer (BU7135 Knowles Corporation) embedded in a silicone base. Each of the APMs is equipped with a microphone complete with a table base, necessary in the calibration phase, and the RS232 / USB connection cable for communication between the APM and the PC. The data-logging unit is able to acquire a time history of the phonation parameters, SAL (skin acceleration level) and fundamental speech frequency (f_0) with a time interval of 50 ms. By means of a previous calibration, carried out by the subject on whom the measurement is made, the measured SAL values are correlated to the level of the vocal sound emission SPL (sound pressure level). The accelerometer is placed on subject's neck and connected with the portable device as shown in Fig. 1.

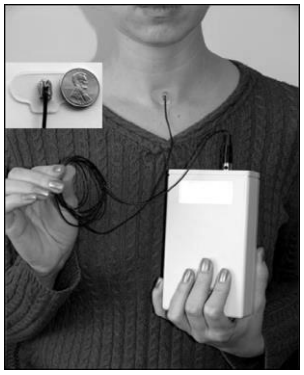


Fig. 1 Positioning of the accelerometer and dimensions of the data logger of the APM.

In order to correlate the accelerometric measure of SAL to SPL and, so, reconstruct vocal effort index, a calibration procedure was performed before the measurement. During this phase, the singer was positioned in such a way as to have his mouth in line with the calibration microphone, at a distance of about 15 cm from it, using the appropriate spacer. Afterwards, to further correlate APM VE values and the parameters under examination, we proceeded in the following way: the same singers to whom the APM dosimeter was applied were asked to perform some simple vocal exercises consisting of emitting short sentences, containing phonemes. These "vocalizes" were recorded before and after the vocal effort of the subject and were also analyzed using two different software. One is the free open source Praat software, used to extract the pitch trend or f_0 time history of the vowel 'a' emitted by the subject. The second is a virtual instrument developed in the LabView platform. With the use of this last virtual instrument it was possible to analyze the temporal trend of the pitch extracted previously with Praat. With this analysis one of the parameters was obtained, that is the total harmonic distortion (THD) of pitch trend, which reflects the greater or less precision of the so-called "vibrato" performed by the singer [14]. The precision in the emission of the vibrato is a parameter that is believed to reflect the strain of the singer's voice and, therefore, could represent an index of his vocal effort.

With the same LabView virtual instrument, a particular harmonic analysis of the recorded vocal signal was carried out from which the SPR parameter was obtained. Here we have preferred to adopt the definition used by Omori et al. [15] according to which the SPR is given by the difference in dB between the highest harmonic present in the range between 2-4 kHz and that between the range 0-2 kHz, but, as in our previous work [13], we extended these range to 0-2.5 kHz and 2.5-6 kHz.

III. RESULTS AND DISCUSSION

The measurements made with the APM on the six subjects after the speech activity of a normal working day gave the results reported in Table I.

Alongside the singer's types are: the estimated average sound pressure level, referred to 15 cm from the mouth ($SPL_{15\ cm}$ [dB]), the A-weighted equivalent continuous sound level at a distance of 1 m from the mouth referred to the entire measurement time ($L_{Aeq\ 1m}$ [dB(A)]) and, in the last column, the degree of effort as established by the standard UNI EN ISO 9921.

Table I: Vocal Effort indexes obtained with the APM.

Voice	**$SPL_{15\ cm}$ [dB]**	**$L_{Aeq\ 1m}$ [dB(A)]**	**Degree of effort according to UNI EN ISO 9921**
Soprano 1	75.3	61.3	normal
Soprano 2	70.3	55.9	relaxed
Mezzo Soprano	83.3	73.4	strong
Tenor	70.7	66.4	elevated
Alto	71.8	74.2	strong
Baritone	74.,6	74.4	strong

These indexes were related to the value of the SPR parameter described above and obtained by the LabView software following the analysis of the "wav" files of the vocalizations of the vowel 'a' carried out by the singers before and after the aforementioned singing activity. A similar correlation was made after the analysis carried out on the pitch files extracted with the Praat program providing the additional parameter under consideration, namely the TDH of the trend over time of f_0. All these results are reported in Table II.

Table II: Comparison between APM indexes and SPR and THD parameters.

Voice	**Degree of effort index**	**SPR pre [dB]**	**SPR post [dB]**	**THD pre %**	**THD post %**
Soprano 1	normal	-29.8	-29.5	44.1	134.0
Soprano 2	relaxed	-25.2	-25.5	33.7	22.6
Mezzo soprano	strong	-17.1	-17.4	31.2	42.1
Tenor	elev.	-23.7	-16.5	8.4*	95.6
Alto	strong	-25.1	-28.8	58.4	124.0
Baritone	strong	-20.7	-30.0	107.4	63.2

It can be basically noted that in absence of effort, that is with a relaxed or normal voice index, the value of SPR does not change, while, in the presence of effort with a strong or high voice index, the parameter, except for one case, increases in absolute value, as expected. As regards the SPR parameter, we note a trend that substantially confirms what has already been deduced in an our previous work [13], although a slightly different definition of the same parameter was used in other papers [15-19]. Although the statistical significance was not achieved, it can be seen in Fig.2 that the differences in SPR between pre and post exposure to vocal effort tend to be much larger, in absolute value when the effort is intense.

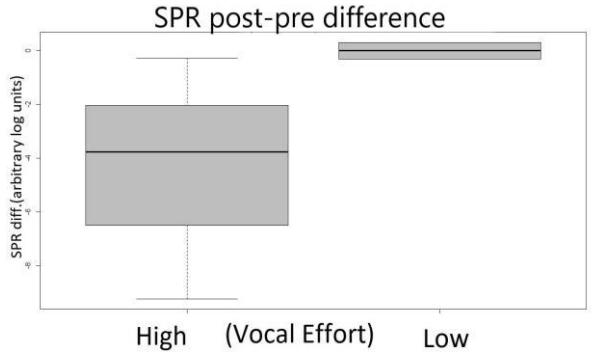


Fig.2 Boxplot representing the SPR difference

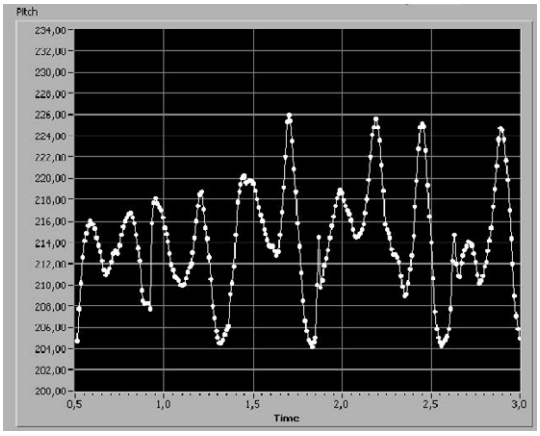


Fig.3 Frame of temporal trend of the pitch showing an excellent emission of "vibrato".

The behavior of the THD is much more difficult to be interpreted. This occurrence is probably due to the non-homogeneity of the vibrato emission mechanism used by the singers, as discussed later. It is interesting to note that in the case of the lowest THD value (Tenor) there is in correspondence an excellent vibrato emission, as shown in Fig. 3.

As regards the pitch TDH parameter , we have seen that this parameter has the right trend with the vocal effort when there is a sustained and stable emission of the vibrato itself.

Unfortunately, as we did not explicitly requested the singers to emit the vocal sung using vibrato, the result of this analysis is strongly influenced by this circumstance.

At the time of recording these vocalizations, we had not yet focused on the analysis of this new parameter, but we were based on a protocol used in a similar previous study [13]. We are planning to focus, in a future study, on a more accurate analysis of the vibrato, emitted with adequate vocalizations.

In fact, it is noted that where the vibrato is clearly present, the parameter under investigation follows the expected trend as function of the vocal effort index, i.e. a degradation of the "purity" of the vibrato, increasing with the effort, can be observed, which translates into an increase of the THD parameter of the pitch.

V. CONCLUSION.

The comparison between the vocal effort, evaluated by means of a method based on the measure of the acceleration of the laryngeal tissues, implemented on the APM commercial system, and a method based on the analysis of the energy content of the harmonic components of the speech signal is likely to permit a validation of the latter. Two different indexes were analyzed, the SPR and a new proposed index based on

the analysis of the pitch (THD of f_0). The first seems to follow the indication given by the accelerometric method, although, due the scarcity of the sample observed, the statistical significance was not achieved. The second parameter, although it is in principle promising, did not give good correlation with the APM parameters. This occurrence is probably due to the non-homogeneity of the vibrato emission as discussed. On the other hand, the energy content of the harmonic components of the speech signal is a simple analysis that could be implemented also on a smart phone that would permit speech signal recording of the subject under test. This procedure could permit an auto evaluation of the vocal effort level during the routine performance at workplace, making this method very simple to use in remote measurement campaigns that are essential in this period of covid-19 pandemic.

ACNOWLEDGEMENTS

We wish to thank the Artistic Direction and the Prevention and Protection Department of the Teatro Regio in Turin. We are also grateful to the singers who kindly co-operated in this research.

REFERENCES

[1] Titze IR., McCabe DJ., "Chant therapy for treating vocal fatigue among public school teachers: a preliminary study." *Am. J. Speech. Lang. Pathol*., 11, pp.356–69, 2002.

[2] Vilkman E., "Occupational safety and health aspects of voice and speech professions." *Folia Phoniatr. Logop*. 56, pp.220 –53, 2004.

[3] Sundberg J., *The Science of Singing Voice*. DeKalb, Illinois: North. Illinois Univ. Press, 1987.

[4] Manfredi C., Bocchi L., Cantarella G., "A Multipurpose User-Friendly Tool for Voice Analysis: Application to Pathological Adult Voices", *Biom. Signal Proc. & Control*, vol.4, pp. 212–220, 2009.

[5] Shipp T, Leanderson R., Sundberg J., "Some acoustic characteristic of vocal vibrato*." J. Res. in Singing*, 4, pp.18-25, 1980.

[6] Stemple JC., Stanley J., Lee L., "Objective measures of voice production in normal subjects following prolonged voice use." *J. Voice*, 9(2), pp.127–33, 1995.

[7] Ford Baldner E., Doll E., and Van Mersbergen M. R., "A Review of Measures of Vocal Effort With a Preliminary Study on the Establishment of a Vocal Effort Measure." *Journal of Voice*, 29(5), pp.530-41, 2015.

[8] Kitch JA., Oates J., "The perceptual features of vocal fatigue as self-reported by a group of actors and singers." *J. Voice*, 8(3), pp.207–14, 1994.

[9] Sapir S., Mathers-Schmidt B., Larson GW., "Singers' and non-singers vocal health, vocal behaviors, and attitudes toward voice and singing: indirect findings from a questionnaire", *Eur.J.Disord.Comm*., 3, pp.193–209, 1996.

[10] Carroll T., Nix J., Hunter E., Emerich K., Titze I., Abaza M., "Objective measurement of vocal fatigue in classical singers: A vocal dosimetry pilot study," *Otolaryng.Head & Neck Surg*., 135, pp. 595-602, 2006.

[11] Titze IR., Svec JG., Popolo PS., "Vocal dose measures: quantifying accumulated vibration exposure in vocal fold tissues." *J.Speech Lang.Hear.Res*., 46, pp.922–35, 2003.

[12] Švec J.G., Titze I.R., Popolo P.S., "Vocal Dosimetry: Theoretical and Practical Issues." AQL 2003 Hamburg: *Proceeding Papers for the Conference Advances in Quantitative Laryngology, Voice and Speech Research*, 2003.

[13] Sisto R., Pieroni A., Annesi D., Nataletti P., Sanjust F., Manfredi c., Venzi M., "Vocal effort in a group of singers of a National lyric orchestra" *Proceedings 6th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 179, Firenze, Italy, 2009.

[14] Półrolniczak E. and Kramarczyk M., "Analysis of the signal of singing using the vibrato parameter in the context of choir singers," *Journal of Electronic Science and Technology*, vol. 11, no. 4, pp.417–423, December 2013.

[15] Omori K., Kacker A., Carroll L. M., Riley W. D., and Blaugrund S.M., "Singing power ratio: Quantitative evaluation of singing voice quality," *Journal of Voice*, vol. 10, no. 3, pp.228 – 235, 1996.

[16] Bloothooft, G. & Plomp, R., "The sound level of the singer's formant in professional singing". *J. Acoust. Soc. Am*., 79(6), pp.2028-2033, 1986.

[17] Seidner, W., Schutte, H., Wendler, J., & Rauhut, A. "Dependence of the high singing formant on pitch and vowel in different voice types". *Proceedings ofthe Stockholm Music Acoustics Conference*. 1983

[18] Watts, C., Barnes-Burroughs, K., Estis, J., & Blanton, D.,"The singing power ratio as an objective measure of singing voice quality in untrained talented and nontalented singers.". *Journal of Voice*, 20(1), pp.82-88, 2006.

[19] Lundy D, Roy S, Casiano R, Xue J, Evans J. "Acoustic analysis of the singing and speaking voice in singing students."*Journal of Voice*, 14, pp.490–493, 2000.

VIBRATION MODE DECOMPOSITION FROM HIGH-SPEED IMAGING OF AUTO-OSCILLATING VOCAL FOLD REPLICAS WITHOUT AND WITH VERTICAL TILTING

A. Van Hirtum[1], X. Pelorson[1], I. Tokuda [2]

[1] LEGI, UMR CNRS 5519, Grenoble Alpes University, France
[2] Dep. Mech. Eng., Ritsumeikan Univ., Nojihigashi, Kusatsu, Shiga 525-8577, Japan

Abstract: **Vertical left-right level difference due to angular asymmetry characterises unilateral vocal fold paralysis. High-speed image sequences of three distinct auto-oscillating silicone vocal folds replicas are analysed simultaneously in both space and time with a dynamic vibration mode decomposition while imposing different degrees of angular asymmetry. From the modes eigenvalue spectra, it is found for all three replicas that the degree of angular asymmetry affects the decay of vibration modes. More in particular, for the assessed vocal fold replicas, increased mode decay is observed when the replica contains a stiff epithelium-like surface layer whereas it decreases otherwise. Spatial mode patterns near the glottal aperture reflect the mode order along the posterior-anterior direction and the imposed angular asymmetry reduces the spatial mode extent near the tilted vocal fold edge. Consequently, the quantified dynamic vibration mode properties, including the ones observed for higher order modes, are of potential interest for clinical studies involving high-speed vocal folds auto-oscillation imaging.**

Keywords: **High-speed vibration imaging, dynamic vibration mode analysis, mechanical vocal fold replica, vertical positioning asymmetry**

I. INTRODUCTION

Unilateral vocal fold paralysis (UVFP) is a common vocal fold (VF) pathology characterised by an air escape due to left-right VF asymmetries of the VF's shape, tension or/and positioning. The glottic insufficiency associated with UVFP is reported to lead to dysphonia as air leakage is often associated with breathy voice or vocal fatigue.

In recent work [1,2], three molded deformable multi-layer silicone VF replicas (two-layer M5, three-layer MRI and four-layer EPI) with different degree of complexity were used to study the effect of vertical tilting of a single VF on their auto-oscillation. Concretely, the right VF was kept in place whereas the posterior edge of the left VF was tilted in the medio-sagittal plane towards the superior direction. The resulting vertical tilting is parameterised by angular asymmetry angle α. Imposing angular asymmetry in the range from 0° up to 25° results in a vertical level difference up to a few millimeters as observed in patients suffering from UVFP. Imposed vertical tilting of a single VF causes left-right VF positioning asymmetry whereas left-right VF tension and shape symmetry are maintained. The use of three different VF replicas allowed to consider the impact of tilting for replicas with different shape and multi-layer composition as outlined in [1,2].

Analysis of the upstream pressure [1] showed a decrease of the oscillation frequency and an increase of the oscillation onset threshold pressure with increasing α was observed. The gradual loss of VF's contact with α, inducing increased glottal air leakage, was pointed out to catalyse these tendencies.

High-speed (HS) imaging of the vibrating VF's for different α was considered in [2]. Local image features, exploiting HS videokymographic (VK) line-scans, were quantified during steady state oscillation. Left-right vibration asymmetry parameters showed that the normal VF entrains the movement of the tilted VF. This left-right vibration asymmetry caused the mucosal wave velocity in the tilted VF to be lower than the one in the normal VF.

In this work, it is sought to further investigate sustained steady state VF vibration without ($\alpha = 0°$) and with ($\alpha > 0°$) angular asymmetry by analysing high-speed (HS) image sequences of the vibrating VF's. Instead of focusing on local spatial features or line scans as in previous work [2], it is sought to assess simultaneously temporal as well as spatial information of the global VF's auto-oscillation. A dynamic vibration mode decomposition (DMD) of the observed vibration snapshots is applied as other VF vibration mode decomposition methods such as the empirical eigenfunction analysis do not inform on the temporal mode dynamics. This allows to evaluate how the effect of angular asymmetry affects the spatial and temporal vibration mode features.

II. EXPERIMENTAL & ANALYSIS METHODS

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

The experimental setup and flow supply used to generate the FS interaction underlying the VF auto-oscillation is similar to the one described in [1,2]. The pressure difference driving the VF's auto-oscillation corresponds to the measured upstream pressure. All experiments are performed with a mean upstream pressure set just (< 50 Pa) above the oscillation onset. Angular asymmetry degrees are set to 0°, 4°, 10° and 20° for all replicas. For the M5 VF replica also 16° is considered. The upstream pressure varies between 400 Pa and 1500 Pa depending on the asymmetry degree as well as on the VF replica and the oscillation frequency ranges from 93 Hz up to 144 Hz. As detailed in [2] a single HS camera (frame rate 4 kHz) is placed in the medio-frontal plane in order to acquire instantaneous images of the vibrating VF's. In this plane, three viewing angles θ_{HS} are considered non-simultaneously resulting in top, diagonal and level views. Acquired instantaneous images are two-dimensional grayscale intensity matrices whose dimension is set by the camera resolution of 240 px × 320 px.

For each assessed condition, 1 second or 4000 subsequent snapshots of steady-state auto-oscillation are analysed. A DMD analysis is applied to each image sequence rearranged in the system matrix **X** [3]. DMD results in a temporal and spatial decomposition of the global VF auto-oscillation assuming a locally linear system dynamics **s**(t) with time t. The system dynamics is then obtained from the eigenvectors and eigenvalues of the system matrix which is matrix notation becomes

$$\mathbf{s}(t) = \mathbf{\Phi} \exp(\mathbf{\Omega} t)\mathbf{b},$$

with $\mathbf{\Omega}$ a diagonal matrix whose entries are the eigenvalues, $\mathbf{\Phi}$ is a matrix whose columns are the eigenvectors and $\mathbf{b}$ is a column vector of the coefficients of the first image in the eigenvector basis. The sought spatial analysis of the global auto-oscillation is provided by the DMD modes given by the eigenvectors. The sought temporal analysis is provided by the corresponding eigenvalues. For each spatial mode, its mode frequency is obtained from the imaginary part of the eigenvalue $\Im(\omega)$ whereas its growth rate is determined from the real part of the eigenvalue $\Re(\omega)$, which is smaller or equal to zero for stable modes. The DMD eigenvalue spectrum is thus obtained by plotting for each mode its frequency as a function of its growth rate. The amplitude of each mode is given by the associated coefficient in $\mathbf{b}$.

III. VIBRATION ANALYSIS RESULTS

A. Temporal DMD mode analysis: eigenvalue spectra

DMD eigenvalue spectra of stable vibration modes for top views of all three replicas and different asymmetry angles are plotted in Fig. 1.

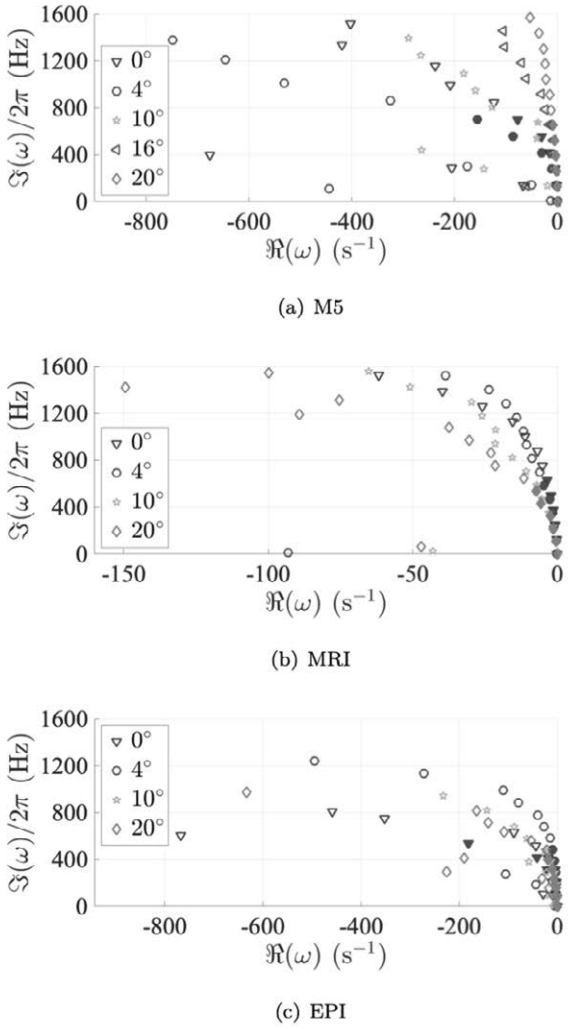


Fig. 1: DMD eigenvalue for top views and different asymmetry angles (symbols) for: a) M5, b) MRI and c) EPI. Filed symbols indicate dominating modes with oscillation frequency up to 750 Hz.

All spectral branches originate with a non-oscillation mean flow mode (j=1) associated with the zero eigenvalue at the origin. This non-oscillation mode expresses the influence of the mean flow along the inferior-superior direction on the VF's vibration pattern. Stable branches with least decay are identified as main spectral branches. Dominating oscillation modes (j = 2...6) on the main spectral branches with frequencies up to 750 Hz are indicated with filled symbols. Both the non-oscillation mode and most dominating oscillation modes decay slowly as their real parts approximate zero. The real part of the eigenvalue

decreases as the mode frequency increases, typically above 750 Hz for the main spectral branches, so that these modes decay mode rapidly and hence contribute less to observed temporal vibration patterns. It is noted that primary oscillation (PO mode, mode j = 2) frequencies obtained from the DMD eigenvalue spectra matches with previous reported values in [1,2]. Despite these similar tendencies for all VF replicas, Fig. 1 shows that the type/structure of VF replica affects the influence of angular asymmetry angle on the eigenvalue spectra. This is most obvious for the M5 replica compared to the MRI and EPI replicas.

Increasing the angular asymmetry from 0° to 4° either decreases (MRI/EPI) or increases (M5) the decay of high frequency modes. Therefore, in terms of the number of non-decaying stable oscillation modes, introducing a slight angular asymmetry either enriches (MRI/EPI) or strips (M5) the vibration pattern. Further increasing the angular asymmetry angle above 4° reveals the opposite tendency as mode decay either decreases (M5) or increases (MRI/EPI). So that in terms of the number of non-decaying stable oscillation modes, the vibration pattern either enriches (M5) or recedes (MRI/EPI). Given the geometry and multi-layer composition of the used VF replicas, it is suggested that observed eigenvalue spectral differences are due to the presence (MRI/EPI) or absence (M5) of a surface layer representing the epithelium.

As perfect angular symmetry (0°) is unlikely to occur in real life, the found robustness of the MRI/EPI VF replica to small angular asymmetries (4° in Fig. 1) supports the hypothesis that adding an epithelium-like layer alters silicone VF replicas from a vibratory point of view. In addition, it provides evidence of the importance of this layer for normal VF vibration. Larger angular asymmetries (> 4° in Fig. 1) mimic pathological VF vibration as all replicas become prone to glottal air leakage due to vertical level difference. Increased mode decay, only observed for the MRI/EPI replica and not for the M5 replica, is consistent with impaired oscillation and reduced vibration quality associated with glottal insuciency reported for human speakers. Thus, it is hypothesized that from a vibration point of view the presence of an epithelium-like layer results in more reasonable vibration patterns. From Fig. 1 is seen that the effect of angular asymmetry on mode decay is most obvious for higher frequency modes (typically > 750 Hz). Therefore, it might be of interest to investigate higher frequency vibration modes as potential clinical markers for voice pathology studies such as UVFP.

The described changes in mode decay rate with angular asymmetry can be partly explained by changes in structural damping. Structural damping is likely to increase (or decrease) as the duration of vocal fold contact along the inferior-superior direction inside each glottal cycle increases (or decreases). This duration was quantified for the assessed VF replicas in [2]. It follows that structural damping either increases (MRI/EPI) or decreases (M5) for angular asymmetry larger than 4° as vocal fold contact was found to increase (MRI/EPI) or decrease (M5).

B. Spatial DMD mode analysis

Modes j, indicated by filled symbols, are arranged (j = 1 . . . 6) according to increasing frequency.

Examples of non-oscillation or mean flow modes, corresponding to the zero eigenvalue at the origin of the eigenvalue spectra (mode j = 1), depend on the imposed asymmetry angle as the resulting glottal air leakage affects the mean glottal flow. Non-oscillation mode amplitudes b_1 are plotted in Fig. 2. Amplitudes are normalised by the amplitude of the PO mode (j = 2) associated with the lowest stable oscillation frequency denoted as $b_{2,0°}$. All values exceed unity so that the mean flow mode dominates the eigenvalue spectra for all assessed configurations. View angles other than top view (diagonal or level) allow to observe phenomena along the medio-sagittal plane, and hence along the main flow direction in the inferior-superior direction, more directly resulting in larger mean flow amplitudes.

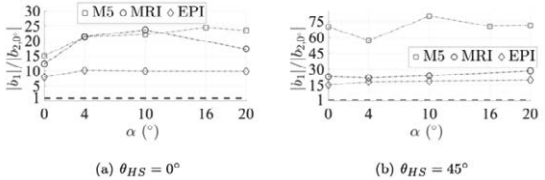


Fig. 2: Normalised non-oscillation mode (j=1) amplitude magnitude: a) top view (θ_{HS}=0°), b) diagonal (θ_{HS}=45°) view.

Mean flow mode amplitudes (j=1, Fig. 2), in particular for diagonal and level views, are most pronounced for the M5 replica, which quantifies the the large displacement in the flow direction experimentally observed for the M5 replica.

Normalised amplitudes of five oscillation modes (j = 2…6) for the MRI and EPI replicas from top and diagonal view angles are plotted in Fig. 3. The PO mode, whose frequency corresponds to the fundamental frequency characterising voiced speech utterances, has the largest amplitude for each angular

asymmetry angle. The plots confirm the decrease of the PO mode frequency with increasing angular asymmetry angle reported in [1,2]. Higher order post-PO modes obtained for j = 3 . . . 6, exhibit oscillation frequencies near the harmonics of the PO mode frequency f_2 for all angular asymmetry angles as $f_j/f_2 \sim$ j-1 holds. As for the mean flow mode, non-top view angles result in larger mode amplitudes as mode patterns develop along both the horizontal transverse and medio-sagittal plane. Overall mode amplitudes decrease with mode order j regardless of view angle or angular asymmetry angle.

Spatial oscillation mode patterns of the MRI replica (top and diagonal views) are illustrated in Fig. 4. Increasing mode order (j) changes the spatial mode with respect to its extent and with respect to the node pattern. In the vicinity of the glottal aperture (within the frames in Fig. 4) typical higher order mode patterns are observed as the number of nodes increases with j. For small angular asymmetry angles (4°) mode patterns develop similarly along the posterior-anterior direction and occupy both VF surfaces as tilting does not alters the elasticity of each VF. For large angular asymmetry angles (20°) spatial modes modes reflects the loss of full VF contact along the posterior-anterior direction. It is observed that a surface region extending from the posterior edge of the tilted VF is no longer part of the spatial mode pattern and that the area of the excluded surface region increases with j. Near the glottal aperture the spatial mode pattern for large asymmetry angles depends on the used view angle. For top view, the loss of VF contact limits the node pattern. For diagonal view, the node pattern becomes more apparent as the tilted VF is observed more directly.

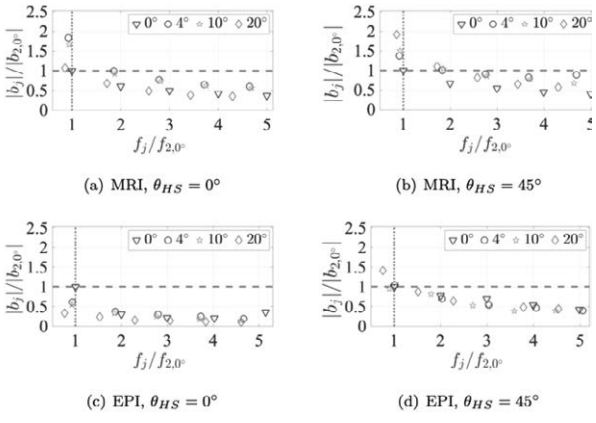


Fig. 3: Normalised oscillation mode amplitude (j=2…6) for top (θ_{HS}=0°) and diagonal (θ_{HS}=45°) views for MRI and EPI replicas.

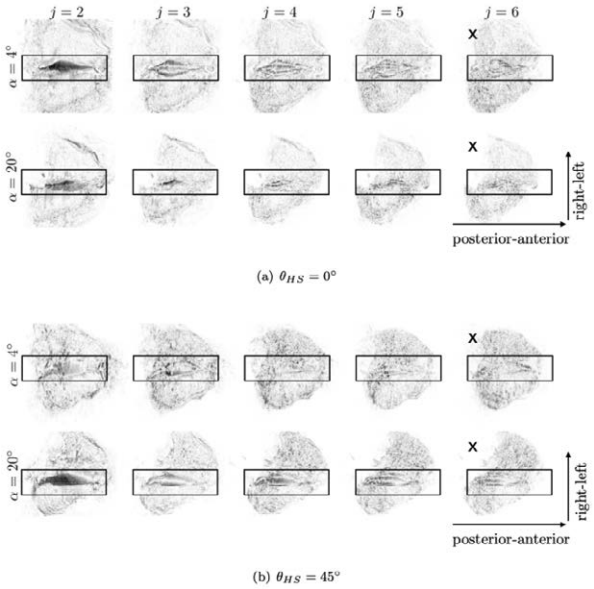


Fig. 4: Spatial oscillation mode patterns scaled between 0 (white, invariant) and 1 (black, most variant) for the MRI replica and asymmetry angles 4° and 20° for top (θ_{HS}=0°) and diagonal (θ_{HS}=45°) views. The glottal aperture is framed. Posterior side of the left VF is tilted as indicated (×) for j=6.

V. Conclusion

Dynamic vibration mode decomposition in space and time of HS images of auto-oscillating silicone vocal folds replicas without and with vertical tilting is assessed. Results encourage to further investigate the potential interest of DMD mode properties, and in particular higher order vibration modes, as potential clinical HS image markers, e.g. for UVFP.

ACKNOWLEDGEMENTS
Full3DTalkingHead (ANR-20-CE23-0008-03).

REFERENCES

[1] A. Bouvet, I. Tokuda, X. Pelorson, A. Van Hirtum, “Influence of level dif- ference due to vocal folds angular asymmetry on auto-oscillating replicas,” *JASA,* vol. 147, pp. 1136-1145, 2020.
[2] A. Bouvet, I. Tokuda, X. Pelorson, A. Van Hirtum, “Imaging of auto-oscillating vocal folds replicas with left-right level difference due to angular asymmetry,” *Biomed Signal Process Control,* vol. 63, pp. 1-12, 2021.
[3] J. Kutz, S. Brunton, B. Brunton, J. Proctor, “Dynamic mode decomposition: data-driven modeling of complex systems,” Eds. Philadelphia: SIAM, 2016, pp. 1-233.

QUANTIFYING THE ELASTICITY OF MECHANICAL VOCAL FOLDS REPLICAS

M. Ahmad, X. Pelorson, A. Van Hirtum

[1] LEGI, UMR CNRS 5519, Grenoble Alpes University, France
mohammad.ahmad//xavier.pelorson//annemie.vanhirtum @ univ-grenoble-alpes.fr

Abstract: **The mechanical properties of two types of deformable mechanical vocal folds replicas are considered. On one hand, the linear elasticity of multi-layered silicone vocal fold replicas with constant elasticity is considered. On the other hand, the mechanical properties of pressurised latex tube replicas with variable elasticity are assessed. In order to quantify the elasticity of these VF's and their influence on the fluid-structure interaction underlying sustained auto-oscillation, experimental as well as model results are presented. This way this work contributes to customised VF replicas with predicted and quantified elasticity.**

Keywords: **Deformable mechanical vocal folds replicas, Young's modulus estimation, Small strain range, Fluid-structure interaction**

I. INTRODUCTION

For human speech sound production, and particularly for phonation or voiced sound production, the presence of two apposed vocal folds (VF) within the larynx, illustrated in Fig. 1, is crucial. Indeed, the fluid-structure interaction between airflow coming from the lungs and the deformable VF tissues on each side of the glottal constriction can result in sustained VF auto-oscillation which is the major sound source for voiced speech sounds. As a consequence, structural properties of healthy as well as pathological VF's influence the fluid-structure interaction and therefore the voiced sound source and potentially the quality of voiced speech sounds.

Physical studies of the fluid-structure interaction underlying the VF auto-oscillation have a long tradition relying on simplifications of the anatomical VF structure as this approach enhances the reproducibility, quantifiability, controllability and hence interpretability of experimental results.

In this work, the elasticity of two types of deformable vocal folds replicas is considered. A first type of deformable VF approximations focuses on maintaining, up to some degree, the anatomical multi-layered structure so that each layer has an appropriate, but constant elasticity. These replicas are obtained as an overlap of molding layers composed of different silicone mixtures. This way the influence of the degree of anatomical realism without or with structural abnormalities on the elasticity can be considered. A second type of deformable VF approximations focuses on the elasticity regulating function within a human VF rather than on its anatomy. In this case, each VF is mimicked as a pressurised latex tube for which the pressure can be varied. As such, the elasticity of each VF can be varied and imposed. Normal and pathological conditions can be considered.

In order to quantify the elasticity of these mechanical VF's replicas and its influence on the fluid-structure interaction underlying sustained auto-oscillation, experimental as well as model results are presented. It is sought to quantify their elasticity and in turn to contribute to its predictability. In this work the linear elasticity is focused on, which is expressed with the effective Young's modulus (EYM) of the composite-like vocal folds structure.

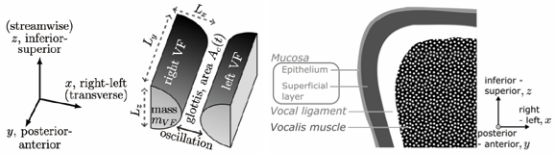


Fig. 1: Illustration of the larynx for VF auto-oscillation and multi-layer structural representation.

II. DEFORMABLE REPLICAS

A. Silicone VF replicas: M5, MRI and EPI

Silicone VF replicas mimic the multi-layer (ML) (micro-)anatomical VF structure to some extent as an overlap of silicone molding layers with constant elasticity following the methodology outlined in [1] and the references therein. Three molded ML silicone VF replicas (M5, MRI and EPI) are shown in Fig. 2 for which layer thickness l_t and overall dimensions L_x and L_z are indicated. The M5 replica is a two-layer (2L) reference model following the body-cover theory of phonation representing thus the vocalis muscle and superficial layer. The MRI replica has a three-layer (3L) structure by adding a thin epithelium-like three-

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

layer (3L) structure by adding a thin epithelium-like to the 2L-M5 structure. The EPI replica is a four-layer (4L) structure obtained by inserting a soft ligament-like deep layer between the muscle and superficial layer of the 3L-MRI replica. Each VF is mounted on a stiff backing layer and replicas have constant elasticity.

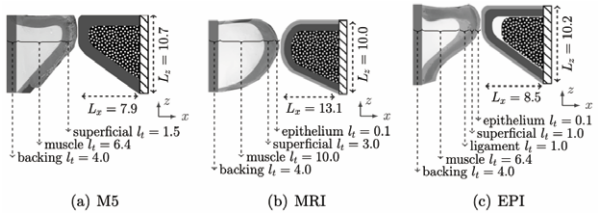


Fig. 2: Coronal section (mm) of a molded silicone multi-layer VF replicas (right VF) and its schematic representation (left VF): a) M5, b) MRI, c) EPI.

B. PLT VF replica

A VF replica with variable elasticity is obtained by representing each VF as a pressurised latex tube (PLT) [2]. Each VF consists of a latex tube enveloping a hollow rigid metal support as depicted in Fig. 3. The latex tube is pressurized (PLT) by filling it with distilled water by means of a water column. The elasticity of the PLT replica depends on the imposed P_{PLT} and thus on the height of the water column. In this work, P_{PLT} is varied between 450 Pa and 6500 Pa (with steps of at most 500 Pa), corresponding to a water column range of about 60 cmH_2O. The PLT VF is positioned in a rigid frame, the same way as during fluid-structure interaction experiments [2], allowing simultaneous observation in the sagittal plane (side view) and the transverse plane (top view).

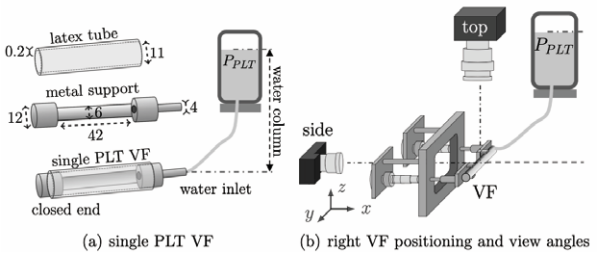


Fig. 3: Overview (mm) pressurized latex tube (PLT): a) single VF, b) spatial VF and camera positioning.

II. EYM ESTIMATION: SILICONE COMPOSITES

A. Six specimens from three silicone VF replicas

Bone-shaped ML silicone specimens [2] with serially stacked layers are designed in order to approximate the ML composition of the three silicone replicas. Each specimen has a test section of length 80 mm in between two clamping ends. The number of layers n, the layer composition, the layer order and layer lengths differ in the same way as between silicone replicas. Specimens are designed as 2L (II_{M5}, $n=2$), 3L (III_{MRI}, $n=3$) and 4L (IV_{EPI}, $n=4$) composites for which the layer lengths l_i match either the layer thickness ratio l_i/L_x or the layer volume ratio V/V_{VF}. As shown in Fig. 4, two different specimens are molded based on the composition of each replica.

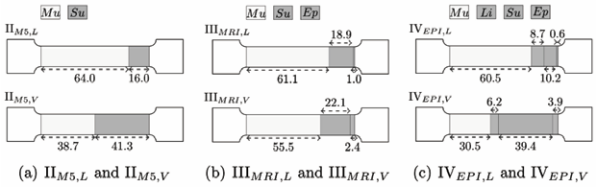


Fig. 4: Serial stacked specimens based on M5, MRI and EPI replicas with layer lengths l_i (mm): muscle-Mu, ligament-Li, superficial-Su, epithelium-Ep.

B. Uni-axial stress tests and EYM estimation

The effective Young's modulus E_{eff} of the molded silicone ML specimen is experimentally estimated from uni-axial stress tests by means of precision loading [1]. Briefly, the force-elongation relationship along the force direction is measured on vertically placed specimens by fixing the upper clamping end and adding a known weight m to the lower clamping end. The weight is gradually incremented. The load force for added mass m is given as its product with the gravitational constant. The specimen's elongation is deduced from geometrical measurements (between 44 and 198 mm) on each layer as a function of weight increment: length and midway area perpendicular to the force direction. The elongation is then the sum of the elongations of each layer, the cross-sectional area of the specimen is obtained as the weighted arithmetic mean of midway areas. Measured force-elongation and area-elongation data are illustrated in Fig. 5a and Fig. 5b for the M5-based specimens. The true stress is then given as the ratio between the force and instantaneous area whereas the true strain is obtained as the natural logarithm of the ratio between the instantaneous length and original specimen length. The EYM corresponds to the slope of a linear fit to the elastic (small strain up to 0.32) region in which the stress is proportional to the strain as plotted in Fig. 5c.

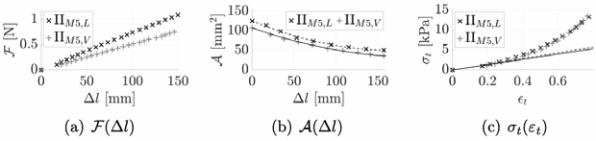


Fig. 5: Uni-axial stress tests for M5-based specimens: a) force-elongation, b) area-elongation, c) stress-strain curves with linear fit (small strain range up to 0.32).

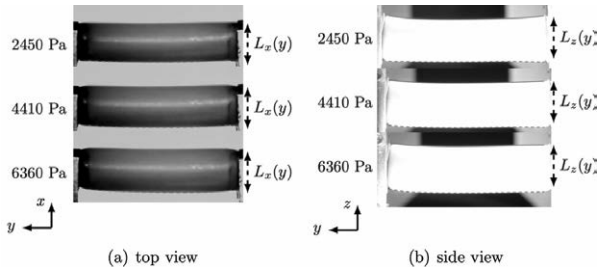


Fig. 5: PLT images and edge detection for several P_{PLT}.

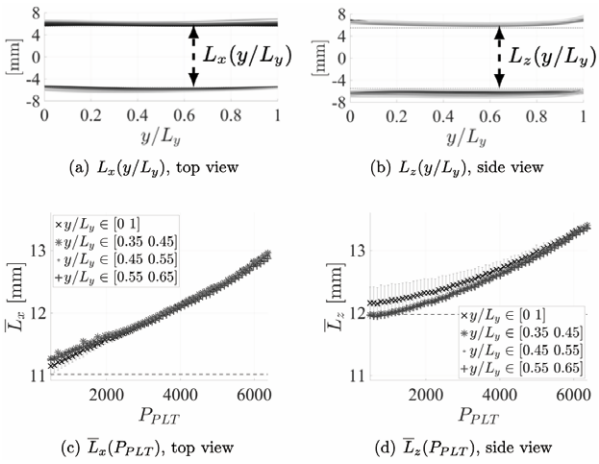


Fig. 6: Characteristic lengths of PLT replica from imaging as a function of P_{PLT} for different y/L_y intervals (symbols): a,b) local and c,d) mean.

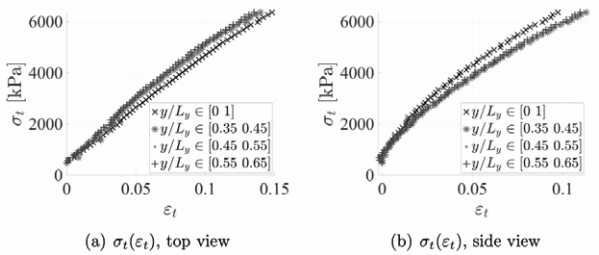


Fig. 7: Image-based stress-strain curves for the PLT replica for different y/L_y intervals (symbols).

B. Modelled EYM for serial stacked specimen

For n serial stacked layers, the stress in the equivalent homogeneous composite and the stress in each layer is constant (Reuss hypothesis) [1]. The effective Young's modulus is then modelled as the harmonic mean of the layers Young's moduli weighted by their lengths. This model was validated in [1] for 2L and 3L silicone specimens for which the layers Young's moduli varies between about 2 kPa up to 65 kPa corresponding to the range of interest for silicone VF replicas. It is noted that model outcome is not affected by the stacking order of the layers within the specimens.

II. EYM ESTIMATION: PLT VF REPLICA

Increasing (or decreasing) the internal water pressure P_{PLT} expands (or shrinks) the PLT replica radially to the posterior-anterior axis. As the PLT can be considered as an inhomogeneous material consisting of both latex and water, the relationship between P_{PLT} and the deformation is governed by a EYM. This EYM is estimated from steady state images taken as a function of P_{PLT}. For each P_{PLT} a top and side view image is taken (see Fig. 3b). Characteristic lengths $L_x(y)$ (top view) and $L_z(y)$ (side view) are obtained as the distances between the replica's edges as illustrated in Fig. 5. Extracted edges and subsequent local characteristic lengths as a function of y/L_y (L_y=*42* mm) are plotted in Fig. 6a (top view) and Fig 6b (side view). Mean characteristic lengths for different y/L_y ranges (overall, short 4 mm intervals) are plotted in Fig. 6c (top view) and Fig. 6d (side view). Mean values agree to within 0.25 mm with respect to the overall mean and to less than 0.1 mm between the 4 mm intervals. The stress-strain curves shown in Fig. 7 are then calculated in the same way as explained for the silicone specimens. It follows that the EYM corresponds again to the slope of a linear fit to the elastic (small strain) region in which the stress is proportional to the strain.

III. RESULTS

A. YEM of replica-based silicone specimen

Measured (filled symbols) EYM and modelled (empty symbols) EYM for the six silicone VF replica-based specimens are plotted in Fig. 8. For MRI and EPI based specimen, measured EYM are close to their overall mean of 5.7 kPa as the standard deviation of 0.4 kPa as well as the maximum difference of 0.8 kPa is small (less than 15%). It follows that imposing either the thickness ratio (subscript L) or the volume ratio (subscript V) does not significantly affect measured EYM for MRI-based or EPI-based specimens. Measured EYM for M5-based specimens exceed values found for MRI or EPI based specimens with 1.0 kPa (volume ratio) or 2.3 kPa (length ratio), so that the imposed ratio affects EYM for M5-based specimens.

The impact (M5-based, EPI-based) or lack thereof (MRI-based) of the imposed ratio (thickness L or volume V) on modelled EYM is understood as the harmonic mean depends on the layer lengths and the layers Young's moduli. For all replicas, the muscle layer has a larger Young's modulus than the superficial layer so that shortening the muscle layer, corresponding to imposing the volume ratio instead of the thickness ratio, results in smaller EYM predictions. The decrease is significant for M5-based (3.4 kPa) and EPI-based (4.7 kPa) replicas. For MRI-based

specimens the decrease is not significant (0.1 kPa) as the muscle layer is shortened with less than 15\% (or less than 5.6 mm) and in addition Young's moduli of the muscle (4.0 kPa) and superficial (2.2 kPa) layer are of the same order of magnitude, which is not the case for the M5-based or EPI-based replicas.

The difference between modelled and measured EYM for all six specimens varies between -2.5 kPa and 2.8 kPa resulting in an overall model accuracy of -0.5 ± 2.1 kPa (mean and standard deviation). Modelled EYM exhibit some of the tendencies described for measured values. Indeed, both measured and modelled EYM for M5-based specimens are greater values associated with MRI-based specimens. Furthermore, the imposed ratio (L or V) affects M5-based specimens more than MRI-based specimens. For the EPI-based replica, measured EYM are in between the range associated with the model. Given the direction of the VF oscillation (Fig. 1), values of specimen for the length ratio is respected are probably more pertinent.

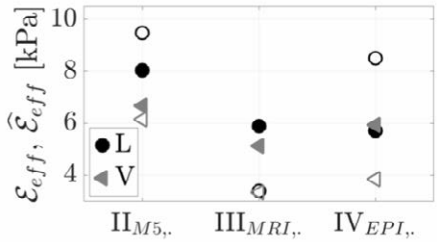


Fig. 8: Measured (filled) and modelled (empty) EYM for specimens respecting either the thickness (L) or volume (V) ratio of silicone VF replicas.

B. YEM of PLT VF replica

The effective Young's moduli estimated from top and side view imaging for the overall and several 4 mm ranges of y/L_y intervals are plotted in Fig. 9. For each viewing angle values obtained for increasing internal pressure P_{PLT} as well as for decreasing internal pressure P_{PLT} are considered. As all P_{PLT} result in a small strain (less than 0.15) deformation, the EYM is obtained from a linear fit to the complete strain range. From Fig. 9 is seen that EYM estimates for different 4 mm y/L_y intervals match for both the EYM obtained for side as well as top viewing. Nevertheless, side EYM exceed values associated with top value with about 6 MPa. Considering the whole y/L_y range stresses this directional difference as it either decreases (to view) or increases (side view) estimated EYM values with about 3 MPa up to 7 MPa. Given the pressure force direction and the movement associated with the oscillation (Fig. 1), values obtained from the top view images are probalby most pertinent. EYM associated with the PLT replica (between 40 MPa up to 57 MPa) are significantly (factor 10^4) greater than values for the silicone based specimens (between 2 kPa up to 10 kPa).

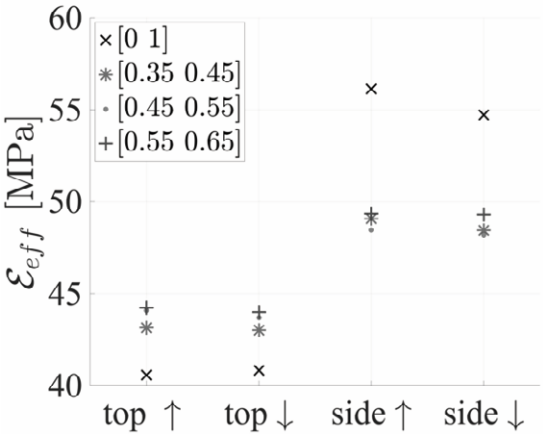


Fig. 9: Measured EYM for the PLT replica from top view and side view images for different y/L_y intervals (symbols) when either increasing (upward arrow) or decreasing (downward arrow) P_{PLT}.

V. Conclusion

Effective Young's moduli for silicone specimens representing silicone VF replicas are measured. For silicone specimens measured and modelled values agree, so that the model can be used as a predictor. Effective Young's moduli of the PLT VF replica are measured as well. In future, the pertinence of found values with respect to mechanical properties and oscillation properties need to be considered.

ACKNOWLEDGEMENTS

Supported by Full3DTalkingHead (ANR-20-CE23-0008-03). Authors thank Cristina Pérez Oms and Dr. Anne Bouvet for their contribution to the experiments.

REFERENCES

[1] M. Ahmad, A. Bouvet, X. Pelorson, and A. Van Hirtum, "Modelling and validation of the elasticity parameters of multi-layer specimens pertinent to silicone vocal fold replicas," *Int J Mech Sciences,* vol. 208, 106685, 2021.

[2] A. Bouvet, X. Pelorson, and A. Van Hirtum, "Influence of water spraying on an oscillating channel," *J Fluid Structures,* vol. 93, 1-20, 2020.

HOW MUCH LOADING DOES COUGH POSE ON THE VOCAL FOLDS? PRELIMINARY HIGH SPEED IMAGE ANALYSIS COMPARING COUGHING AND PHONATION

J. Horáček[1], V. Bula[1], A. Geneid[2], V. Radolf[1], A-M. Laukkanen[2]

[1] Institute of Thermomechanics of the Academy of Sciences of the Czech Republic, Prague, Czech Republic
[2] Phoniatric Clinic, Helsinki University Hospital, Finland
[3] Speech and Voice Research Laboratory, Faculty of Social Sciences, Tampere University, Tampere, Finland
jaromirh@it.cas.cz, bula@it.cas.cz, Ahmed.Geneid@hus.fi, radolf@it.cas.cz, Anne-Maria.Laukkanen@tuni.fi

Abstract:
Coughing offers a risk for voice problems by increasing vocal loading. This study estimates loading by measuring glottal area variation from high speed images and subglottic pressure from oral pressure P_{oral}. Phonation on [o:] and coughing at the same P_{oral} (6 Pa) and SPL (93 dB_{6cm}) were compared from one healthy male. In coughing, the glottal width (GW) at the middle of vocal folds (VFs) was 25% larger. GW measured at VF processes was almost unchanged. Maximum glottal opening velocity dGW/dt was nearly 40% higher, maximum glottal declination rate (MWDR) was up to 3 times higher, and MWDR at vocal processes was 13% higher. The acceleration and deceleration values for VFs were 40% and 47% higher, respectively. *Fo* in the last part of coughing decreased from f_0=222 Hz to 77 Hz at phonation offset. In [o:] f_0 was 116 Hz. Closed quotient CQ≅0.50 in coughing was close to CQ=0.47 in vowel. Vibration frequency of the false vocal folds (FVFs) registered in the first, rough part of coughing, was 293 Hz. Peak-to-peak value of P_{oral} increased 5.4 times in coughing. During vibration of FVFs in coughing, mean P_{oral} increased from 6 Pa to 70 Pa and $P_{oral\ p\text{-}t\text{-}p}$ increased 2.45 times.

Keywords: **Glottal area, EGG, oral air pressure, laryngeal movement, coughing therapy**

I. INTRODUCTION

Throat clearing and coughing are known to be related to voice problems, and recent studies also support this [1]. Coughing involves a tight glottal closure, high subglottic pressure (P_{sub}), abrupt glottal opening and high transglottic airflow [2].

Ross et. al. [3] in 1955 studied the changes of intrapleural air pressure and of airflow at the mouth during coughing and found the pressures up to 18.7 kPa, and flow rates of expired air up to 6.5 L/s. Because he also found the lumen contraction of the trachea during coughs, he estimated maximal airflow velocity in trachea up to unbelievable 280 m/s. Later, in 1975 Evans & Jaeger [4] measured airflow rates at the mouth during coughing and forced expirations in 10 subjects and found out mean values 8.8 L/s in both cases.

These pressure and flow values found in coughing are about one order higher than the maximal values found in human voicing. According to [5], the mean P_{sub} for normal vowel phonation is in the range of 400-2600 Pa (or up to max. 5 kPa), and the mean volume flow rate is in the range 0.07-0.3 L/s in normal vowel phonation in speech mode and up to 0.845 L/s in pathological cases [6]. Therefore, in coughing there is an explosive increase of flow between the vocal folds when the glottis opens.

In addition to such extremely high VF loading during the abrupt opening of the glottis in coughing, an increase of impact stress and acceleration and deceleration related strain on the tissue is also expected during phonation part of coughing. Additionally, in video material it is possible to see vigorous movements in all laryngeal structures during throat clearing and coughing. This may cause shear stress in soft tissues and trauma also on the cartilages, e.g. leading to the development of contact granulomas at the arytenoid region [7].

This pilot study compares coughing and phonation in terms of glottal width variation and movements of the laryngeal structures. The results may shed further light on the loading mechanisms in coughing.

II. METHODS

High-speed laryngoscopic data were obtained from one normophonic male participant with a healthy larynx. He phonated on vowel [o:] and produced coughs. KayPentax Color High-Speed Video System (model 9710¸ KayPentax, NJ) with spatial resolution of 512x512 pixels was used. The sampling frequency was set to 2,000 fps. A rigid scope was inserted through a hole in a T- shaped (2 cm in diameter) mouthpiece into the pharynx. Oral air pressure (P_{oral}) was registered in

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

the mouthpiece, through which the endoscope was inserted into the pharynx. A Glottal Enterprises manometer and a PT75 transducer were used (Glottal Enterprises, Syracuse, NY). The acoustic signal was recorded using an AKG (Type C477; AKG Acoustics, Vienna, Austria) head-mounted microphone at 6cm from the corner of the participant's mouth. The mouthpiece both enabled air pressure registration and helped to fix the endoscope position. Simultaneous recordings of the electroglottograph (EGG) and acoustic and oral pressure signals were made with Computerized Speech Laboratory (CSL; KayPentax, NJ).

For the present study, the glottal width variation was derived from the images at the membranous and cartilaginous parts of the glottis. Maximum amplitude of both glottal widths' variation (GW) was measured. Maximum glottal opening and closing declination rates (MWDR) were obtained from the first time derivative of GW, and acceleration and deceleration values from the 2nd derivative. Similarly, the strong vibrations of the false vocal folds (FVFs) were also quantified.

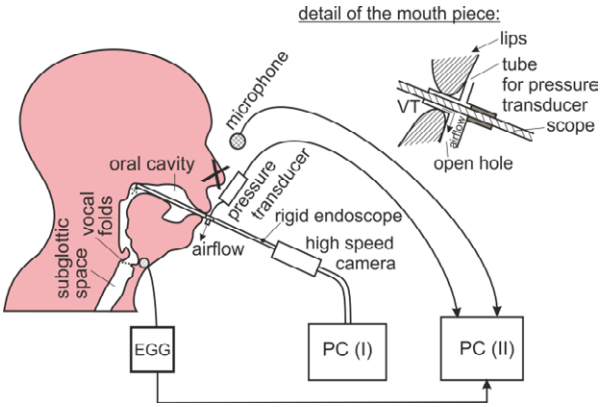


Fig. 1 Measurement set-up.

III. RESULTS

A coughing sample of 0.755 s total duration was analyzed, see Fig. 2. The first part of expiration, 0.263 s of duration, was characterized by a slow squeezing of FVFs and VFs processes during inspiration followed by fast and rough changes of all the laryngeal structures. During sudden expulsion of air only vibrations of FVFs were possible to evaluate, because the VFs were partially hidden, and their vibrations were too fast related to the sampling frequency of the HS camera. In the 2nd part of expiration of the length 0.132 s, a slow variation of laryngeal opening and closing was seen up to the time 0.395 s, where the 3rd part of coughing started by a second rough expiration phase characterized by a transient-like phonation up to phonation offset and final glottal opening.

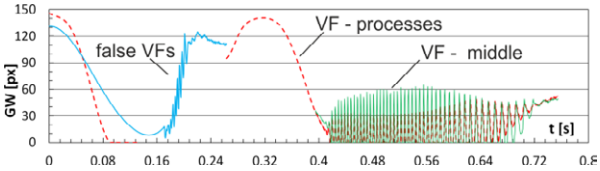


Fig. 2 Variation of distances between FVFs, VFs processes and the glottal width (GW) measured at the middle of the vocal folds, obtained from HS images the during analyzed coughing sample.

Fig. 3 shows in detail the vibration (GW(t) and the time derivative dGW/dt) of the FVFs in the first part of the coughing sample, together with the synchronized audio (Mic), P_{oral} and EGG signals. Similarly, Fig. 4 shows GW(t) and dGW/dt of the VFs measured during the 2nd and 3rd parts of the coughing sample. For comparison, Fig. 5 shows the results measured for 'ordinary' phonation on vowel [o:].

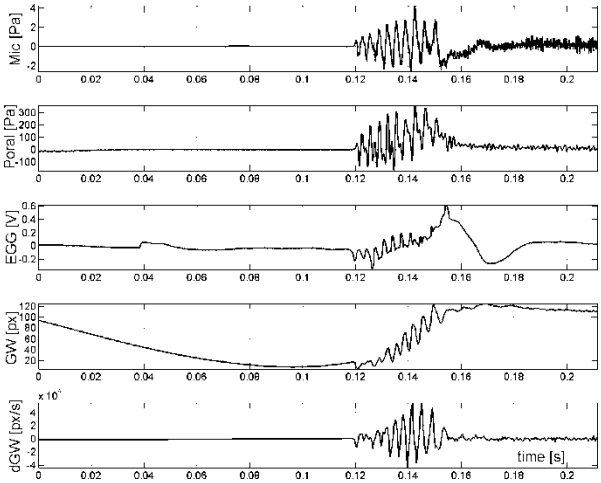


Fig. 3 First part of the analyzed coughing sample, where GW and dGW are shown for false VFs. EGG may reflect vibration of FVFs, or synchronous vibration of FVFs and VFs.

Table 1 compares data on maximal amplitudes of VFs vibrations obtained from the middle of the VFs and demonstrated in Fig. 2 for coughing and in Fig. 5 for the vowel phonation. All values are substantially higher for coughing. The closed quotient CQ was 0.47 for phonation, and in coughing it first increased from 0.43 to 0.50 and then back down to 0.32 in the end of the last part of the sample. Fundamental phonation frequency for vowel was f_0=115.6 Hz, and during coughing it decreased from f_0=222 Hz at the beginning of VFs vibration to only f_0=77 Hz at the phonation offset. Normalized amplitude quotient

$$NAQ = f_0 \text{ (maximum GW/MWDR)}$$

was 0.260 for phonation and during coughing it increased from 0.158 to 0.267.

The FVFs vibrated at the frequency 292.7 Hz, averaged from six periods of the GW(t) waveform, while the VFs vibration was not possible to identify in the HS video. Also from the EGG signal in the first part of the coughing process (see Fig. 3) it was possible to calculate the average value 292.0 Hz for the frequency of VFs vibration that started with the frequency ca 333 Hz, and decreased to 233 Hz after 8 periods.

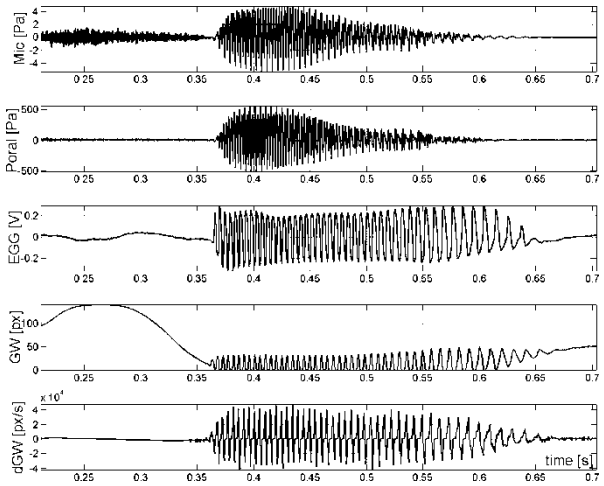


Fig. 4 Second and third part of the coughing sample, where GW and dGW are shown for VF processes.

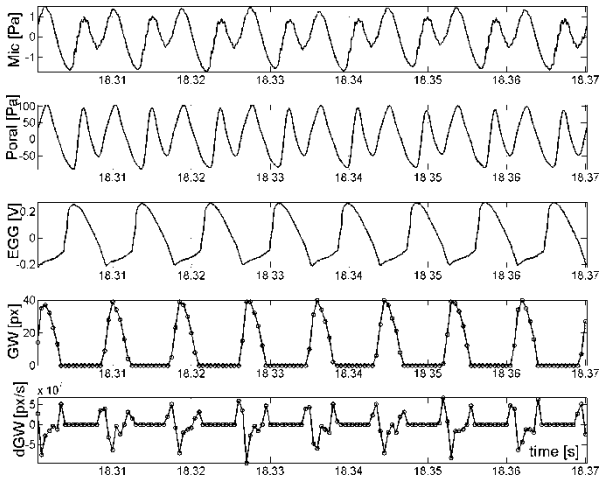


Fig. 5 – Samples of analyzed signals for vowel phonation, where GW and dGW are shown for VFs.

Table 2 compares data on maximal amplitudes of glottis oscillations measured at the VFs processes. The differences between cough and vowel phonation are much smaller at the VFs processes compared to the values found in the VFs middle. Some values measured in coughing were even smaller than in phonation, especially the maximal speed of the glottis opening dGW and its maximal acceleration ACC. This observed phenomenon probably results from a larger total moving mass because laryngeal structures joint to the VFs processes moved simultaneously with them.

Table 1. Maximal values of the normalized glottal width GW, first derivatives of GW specifying maximal speeds of glottis opening dGW=dGW(t)/dt and closing (MWDR), acceleration ACC=$d^2GW(t)/dt^2$ and deceleration DCC measured at the middle of the VFs; Δ is the difference between cough and vowel.

	GW [1]	dGW [1/s]	MWDR [1/s]	ACC [1/s^2]	DCC [1/s^2]
vowel	0.300	346.8	133.4	6.40E5	7.2E5
cough	0.374	483.6	403.0	8.95E5	10.6E5
Δ [%]	+24.7	+39.5	+302.1	+39.8	+47.2

Table 2. Maximal values of the normalized glottal width GW, first derivatives of GW specifying maximal speeds of glottis opening dGW=dGW(t)/dt and closing (MWDR), acceleration ACC=$d^2GW(t)/dt^2$ and deceleration DCC measured at the VFs processes; Δ is the difference between cough and vowel.

	GW [1]	dGW [1/s]	MWDR [1/s]	ACC [1/s^2]	DCC [1/s^2]
vowel	0.263	320.1	213.4	6.67E5	6.40E5
cough	0.282	276.3	241.8	4.84E5	6.91E5
Δ [%]	+7.2	-13.7	+13.3	-0.27	+8.0

Table 3 compares audio and pressure data obtained from the waveforms shown in Figs. 2-4 for coughing and in Fig. 5 for vowel phonation. The highest values of SPL, obtained with external microphone (100 dB) and from P_{oral} (140 dB), and the highest peak-to-peak values of $P_{p\text{-}t\text{-}p}$ (1065 Pa) were measured in the third part of the coughing sample. The highest P_{oral} (70 Pa) was found in the time interval where the false VFs were vibrating. All these values are much higher than the data measured for vowel phonation.

Table 3. Results of audio (acoustic peak-to-peak pressure $P_{mic,ptp}$ and SPL_{mic}) and oral pressure (mean P_{oral} , maxima of peak-to-peak $P_{oral,ptp}$ and SPL_{Poral}) signal measurements in three parts of the coughing sample.

	$P_{mic,ptp}$ [Pa]	SPL_{mic} [dB]	P_{oral} [Pa]	$P_{oral,ptp}$ [Pa]	SPL_{Poral} [dB]
vowel	3	93	6	196	129
cough 1	7	96	70	481	136
cough 2	3	86	11	60	114
cough 3	10	100	6	1065	140

* Data evaluated only in the time interval of false VFs vibration.

IV. Discussion

Figs. 2 - 4 show that the cough example studied here fully corresponds to the typical coughing process published for the voluntary coughing sounds in healthy

subjects, see Yanagihara et all. [8] and Korpáš et all. [9]. The typical sound record consists of three parts. The expulsive phase of cough starts in the moment of glottal opening when the first burst of sound emerges. This is followed by a noisy interval which corresponds to steady-state flow with the glottis wide open. The glottis narrows at the end of the expulsive phase which generates the second burst of the sound.

The results show clearly that even in this type of relatively soft coughing – or rather throat clearing - (with similar P_{oral} and SPL as in ordinary phonation), the estimated vocal loading must be much higher, as both glottal vibration amplitude, and opening and closing rates increased substantially compared to ordinary phonation, particularly the glottal closing rate. This increases impact stress and acceleration and deceleration related stresses [10, 11]. Fast (3 times higher than f_o) vibrations of the FVFs were also observed, and the vocal processes and other laryngeal structures vibrated as well. The closing rate of the vocal processes increased over 10% which fits with the finding that chronic coughing increases the risk of contact granulomas at the arytenoid region [7].

It would be tempting to compare the results with those obtained for loud and strained phonation [e.g. 12] but differences in the research methodology make the comparison difficult. However, according to [12] the glottal area declination rate increased in average 69.6% from typical voice loudness to loud, while in the present study the change from habitual vowel phonation to throat clearing was 302.1%.

On the other hand, the fact that in the present study a mouthpiece was used gives opportunity for an interesting speculation. According to the participant's comments (and those of other participants not studied here), it was difficult to produce forceful coughing with the mouthpiece. It is thus plausible that by offering some flow resistance the mouthpiece raised P_{oral} which is prone to reduce transglottic pressure and glottal closing speed from what it would be without the flow resistance [13]. It remains to be studied, whether this kind of a method could be exploited to reduce vocal fold loading in patients suffering from chronic coughing.

V. Conclusion

This preliminary data show measurable characteristics that enable estimation of coughing related vocal loading. The substantial increase of maximum glottal width declination rate during coughing compared to ordinary phonation implicates much higher vocal folds loading, although in this study a mouthpiece was used and this damped coughing somewhat. Our next study will concern coughing without the mouthpiece to investigate the usability of a mouthpiece as a potential device to reduce vocal loading in chronic coughing. A higher image rate is also warranted for a more detailed image analysis.

VI. Acknowledgement

The study was supported by the Czech Science Foundation, Grant No. 19-04477S: “Modelling and measurements of fluid-structure-acoustic interactions in biomechanics of human voice production.”

References

[1] J.F. Martinez-Paredes, R. Alfakir, CC. Thompson et al. “Effect of chronic cough on voice measures in patients with dysphonia,” *Journal of Voice* 2020 in press. https://doi.org/10.1016/j.jvoice.2020.12.025

[2] J.H. Comroe, *Physiology of Respiration*, 2nd ed., Chicago, Year Book Medical Publishers, Inc. 1974.

[3] B.B. Ross, R. Gromiak, H. Rahn, “Physical dynamics of the cough mechanism,” *J. App. Physiol.* 8, 264-268, 1955.

[4] J.N. Evans and M.J. Jaeger, “Mechanical aspects of coughing,” *Pneumonologie* 152, 253-257, 1975.

[5] M. Hirano, “Clinical examination of voice.” in *Disorders of Human Communication 5*. G.E. Arnold, F. Winckel and B.D. Wyke Eds. Springer-Verlag, Wien, 1981.

[6] R.J. Baken and R. Orlikoff, “*Clinical Measurement of Speech and Voice.*” 2nd edition, Singular, 2000.

[7] M.K. Wani, G.E. Woodson, “Laryngeal contact granuloma.” *The Laryngoscope* 109: 1589-1593, 1999.

[8] N. Yanagihara, H. Leden, E. Werner-Kukuk, „The physical parameters of cough. The larynx in a normal single cough.“ *Acta Oto-laryng* 61, 495–509, 1966.

[9] J. Korpáš, J. Sadloňová, M. Vrabec, „Analysis of the Cough Sound: an Overview.“ *Pulmonary Pharmacology* 9**,** 261–268, 1996.

[10] J. Jiang and I. Titze, "Measurement of vocal fold intraglottal stress and impact stress." *J. Voice* 8, 132–144, 1994.

[11] J. Horáček, A.M. Laukkanen, P. Šidlof, et al. "Comparison of acceleration and impact stress as possible loading factors in phonation. A computer modeling study." *Folia Phon Logop*. 61, 137-145, 2009.

[12] R.R. Patel, J. Sundberg, B. Gill et al. "Glottal airflow and glottal area waveform characteristics of flow phonation in untrained vocally healthy adults." *J Voice* in press, 2020.

[13] I.R. Titze, "Voice training and therapy with a semi-occluded vocal tract: rationale and scientific underpinnings.” *J Speech Lang Hear Res* 49, 448-459, 2006.

SESSION V
SINGING

HUMMING BEATBOXING: THE VOCAL ORCHESTRA WITHIN

Annalisa Paroni[1], Hélène Lœvenbruck[2], Pierre Baraduc[1], Christophe Savariaux[1], Pascale Calabrèse[3], and Nathalie Henrich Bernardoni[1]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, F-38000 Grenoble
[2]Univ. Grenoble Alpes, Univ. Savoie Mont-Blanc, CNRS, LPNC, F-38000 Grenoble
[3]CNRS, Grenoble INP, TIMC-IMAG, F-38000 Grenoble

annalisa.paroni@gipsa-lab.fr, Helene.Loevenbruck@univ-grenoble-alpes.fr, pierre.baraduc@gipsa-lab.fr, christophe.savariaux@gipsa-lab.fr, Pascale.Calabrese@univ-grenoble-alpes.fr, nathalie.henrich@gipsa-lab.fr

Abstract: **Humming beatboxing is a technique used by beatboxers to give the impression of producing multiple sounds synchronously. How this is achieved and what the differences are with regular beatboxing remains mostly unexplored from a scientific standpoint. Four beatboxers were recorded. Electromagnetic articulography was combined with acoustic, electroglottographic, and breathing measurements. The articulatory and breathing behaviors of three boxemes (kick, hi-hat, rimshot) were compared between a regular and a humming realization. When produced as regular beatboxing, the trajectories of the tongue were consistent with a glottalic initiation mechanism and breathing behavior was related to the acoustic outcome. In contrast, for humming sounds the breathing behavior was dissociated from articulation and acoustic outcome, suggesting that the initiation mechanism took place within the oral cavity and the more posterior portion of the vocal tract did not participate in the production of those sounds. Articulatory trajectories were consistent with a closure held at the posterior region of the oral cavity. This suggests that in the humming technique the vocal tract is divided into two sections: the oral cavity functions on its own to produce the rhythmic line, while the melodic line is produced in the laryngeal or pharyngeal spaces and propagates through the nasal cavities.**
Keywords: **Human beatboxing, humming beatboxing, humming**

I. INTRODUCTION

Human beatboxing (HBB) is an emerging and rapidly evolving vocal art that relies on the human vocal instrument to produce all kinds of sounds for the purpose of music making. The core of HBB is instrumental mimicry: typically, the first sounds a beatboxer learns are those that reproduce the drum set sounds, i.e. kick, hi-hat, snare/rimshot, cymbal. More experienced beatboxers, however, can be considered as multivocalists, as they exploit a wide variety of vocal techniques such as rapping, singing, overtone singing, scratching, etc. depending on the style of music they want to produce. In particular, the humming technique can be used to give the impression of multiple sound sources within the same beatboxer: a rhythmic line and a melodic line can be produced simultaneously. This technique is well known by beatboxers and is generally resumed as the technique that allows a beatboxer to produce multiple sounds at the same time using the air present in the mouth to produce the rhythm, and the voice to produce the melody. However, how this is achieved remains mostly unexplored from a scientific standpoint. This study focuses on three categories of drum sounds (kick, hi-hat, rimshot) produced as regular HBB sounds or in the humming technique. Some studies have shown that regular kick, hi-hat, and rimshot are generally produced via a piston-like action of the closed glottis [3, 6, 1], i.e. using a glottalic initiation mechanism [4]. The only published study so far that directly investigates humming boxemes (i.e. HBB sounds) has shown that the humming versions of these three boxemes are produced via a pushing or pulling action of the tongue [5], i.e. using a velaric initiation mechanism [4]. The present study aims at elucidating the similarities and differences in terms of breathing strategy and articulatory mechanism between regular and humming kick, hi-hat, and rimshot as well as giving some insights on how the vocal tract is configured when producing different sounds simultaneously.

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

II. Methods

Four male French speaking beatboxers (S02-S05) were recorded, three professionals and one amateur, aged 20 to 38 years. The recordings took place in the semi-anechoic room of GIPSA-lab in Grenoble, a place of biomedical research authorized by the ARS Auvergne-Rhône-Alpes. Electromagnetic articulography (EMA WAVE, NDI, Canada) and respiratory inductance plethysmography (RIP, ETISENSE, France) were combined with electroglottographic, acoustic and video recordings. Audio and EGG signals were sampled at 20 kHz, RIP and EMA signals at 200 Hz.

The three boxemes kick (**P**), hi-hat (**t**), rimshot (**K**) were produced 12 times each in a row. Each repetition was preceded by [sasɛlə] (English translation: "this is the"). Each boxeme sequence was produced in regular HBB, then in the humming technique. Further, the *beat*, i.e. a musical phrase, **PtKtPtKt** was repeated 10 times as regular HBB, and 10 times as humming HBB. The data recorded by the different systems were all synchronized together. The audio recordings were manually segmented and annotated using Praat software [2]. Spatial trajectories of 9 coils placed on five flesh points of the tongue (apex/blade, middle, right, left and dorsum) and four flesh points of the lips (upper and lower, median and lateral) were extracted from the EMA recordings. Mean trajectories and variance were computed using the commercial software package MATLAB.

III. Results

The three professional beatboxers (S03, S04, S05) produced two versions of the humming boxemes: one was only a sequence of drum sounds, i.e. the rhythmic line (**RL**), the other was a superposition of drum sounds (**RL**) and a hummed melodic line (**ML**). The presence of vocal-fold vibration is attested by the EGG signal in Fig. 2. However, one beatboxer alternated vocal-fold vibration and glottal stops when producing his **ML**. The amateur beatboxer (S02) gave only one version of humming boxemes as post-voiced boxemes: no vocal-fold vibration occurred synchronously to the drum-sound production, but was present right after. No vocal-fold vibration was detected during regular HBB production.

Breathing strategies varied among beatboxers and stimuli. Shorter tasks such as boxeme repetition held the most variability. However, a typical pattern emerged during humming **RL** tasks: an increase in thoracic and abdominal circumferences before the initiation of the *beat* was followed by an alternation of decrease and increase during the *beat*. Fig. 2 shows that this alternation can be similar to breathing behavior at rest, but was not related to the acoustic outcome. When voicing was added during humming **RL+ML** executions, the evolution of thoracic and abdominal circumferences was similar to speech: an increase before vocal-fold vibration initiation attested of air intake, followed by a regular decrease during voicing.

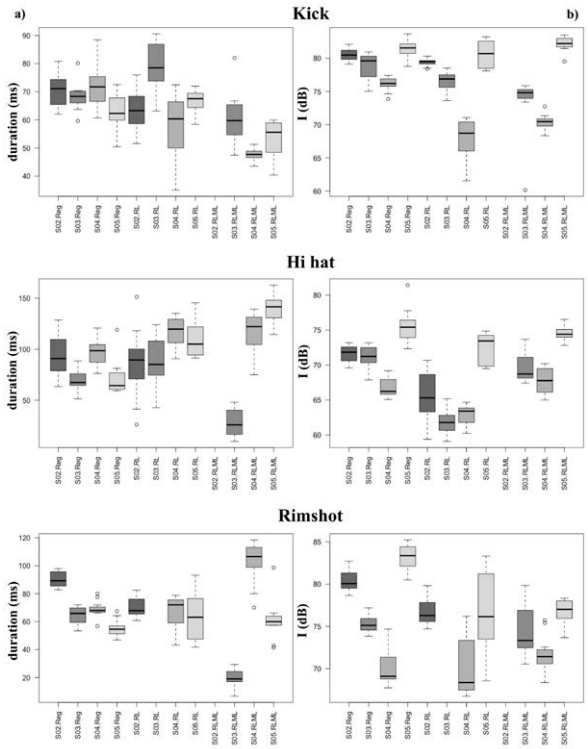


Figure 1: Distribution of a) sound duration (in ms) and b) sound intensity (in dB) for each boxeme produced by each beatboxer (S02-S05) as regular HBB (Reg), humming (RL), and voiced humming (RLML).

Regular HBB production showed the most varied breathing strategies among beatboxers, especially for shorter tasks (boxemes repetition). Longer tasks, more similar to real-life HBB, attested of a typical behavior: a tendency towards stabilization of thoracic (and possibly abdominal) circumference during the *beat* execution, with small local variations in correspondence with each boxeme acoustic production. In the case illustrated in Fig. 2, the more prominent local variations occurred in correspondence with **K** and indicated a rapid increase in thoracic circumference suggesting a small inhalation during the boxeme production.

Articulatory behavior was quite consistent among the four beatboxers for the realization of **P** and **t**, whereas **K** showed more variability. Fig. 3 illustrates mean trajectories and acoustic outcomes for S04. **P** was always produced as a bilabial occlusive. However, the tongue was particularly active in the realization of the three variants (Regular, humming **RL**, and humming **RL+ML**). As for the two humming versions, the articulatory data show

that the superposition of the **ML**, in this case the vocal-fold vibration, to the **RL** did not impact the lingual or labial movements. The tongue was raised high against the palate in the back of the oral cavity and was pushed forward right before and during the occlusion release and the acoustic realization. Breathing data showed no relation between breathing and acoustic production of **P**. In the case of the regular **P**, the tongue assumed a lower position in the oral cavity and underwent an upward displacement that began before the occlusion release and ended after the cessation of the sound. Breathing data showed local decrease in thoracic circumference around sound production, suggesting the use of an egressive airstream. This resulted in slightly shorter and softer humming sounds compared to their regular equivalents (Fig. 1). **t** was always produced as an alveolar occlusive. Two main articulatory strategies were observed for the humming versions. One strategy consisted in holding the tongue against the palate and then suddenly producing a rapid downward movement, especially in the middle region, during which the occlusion was released and the sound took place. The other (shown in Fig. 3) was via occlusion of the vocal tract in the anterior and posterior region of the oral cavity, creating an air pocket between the middle region of the tongue and the palate and subsequently compressing the trapped air via a pushing action of the middle section of the tongue, releasing the anterior occlusion and producing the sound. In both articulations the tongue assumed a high position, especially in the back region of the oral cavity. The breathing data showed no relation with sound production. Regular **t** was achieved with a lower position of the tongue. Only the more anterior part of the tongue made contact with the palate during the occlusion phase. At occlusion release, the anterior portion of the tongue was lowered and at the same time the posterior portion of the tongue underwent an upward movement. Breathing data showed local decrease in thoracic circumference, suggesting the use of an egressive airstream. The humming versions of **t** generally were longer and softer than the regular **t**. **K** was achieved using the most different articulatory strategies among the beatboxers. For the most part, the humming versions were realized pushing the tongue against the palate and then pulling it down in a rapid motion, while the more anterior region of the tongue was kept in contact with the palate (Fig. 3) and the occlusion was released on one side of the tongue. In the humming **PtKt** task, one beatboxer also produced **K** as a bilabial occlusive, where the pressure buildup was achieved via compression of the cheeks. Again, no relation emerged between breath and acoustic realization. Regular **K** was realized in two different ways. An occlusion was created in the back region of the tongue against the palate, then released via a rapid downward motion of the tongue (Fig. 3). In the regular **PtKt** task, two beatboxers used a different articulatory behavior with a different acoustic outcome: the tongue was kept in contact with the palate during the occlusion phase, then the occlusion was released in the back region of the tongue, while the front portion of the tongue was kept in contact with the palate. Breathing data showed a local increase in thoracic circumference, suggesting the use of an ingressive airstream. Once again, humming boxemes resulted as softer and generally slightly shorter than their regular equivalent.

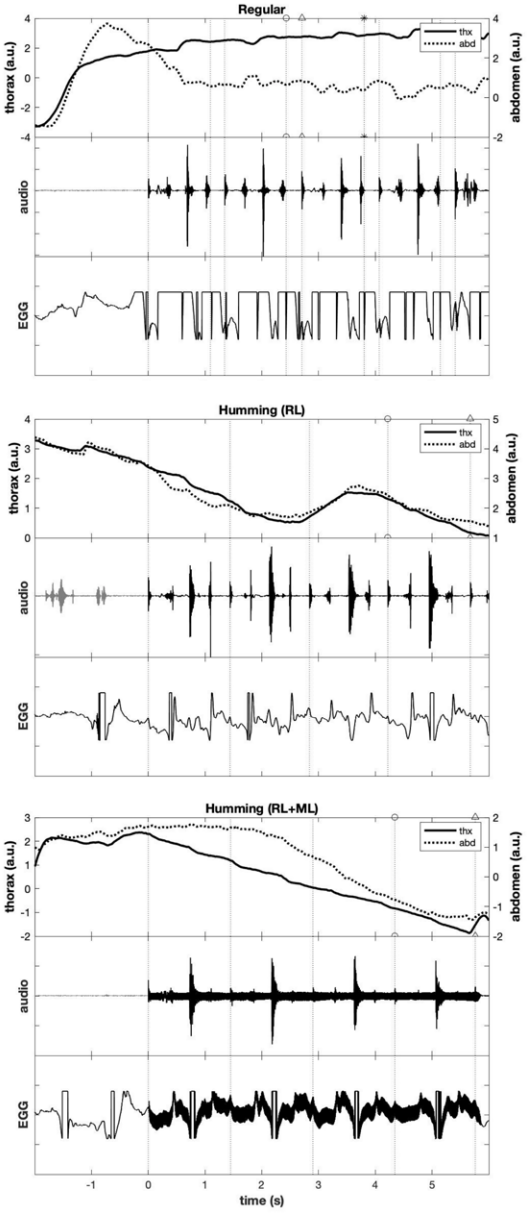


Figure 2: Breathing, audio, and EGG signals of S04 producing the *beat* **PtKt**. y-axes are arbitrary scales.

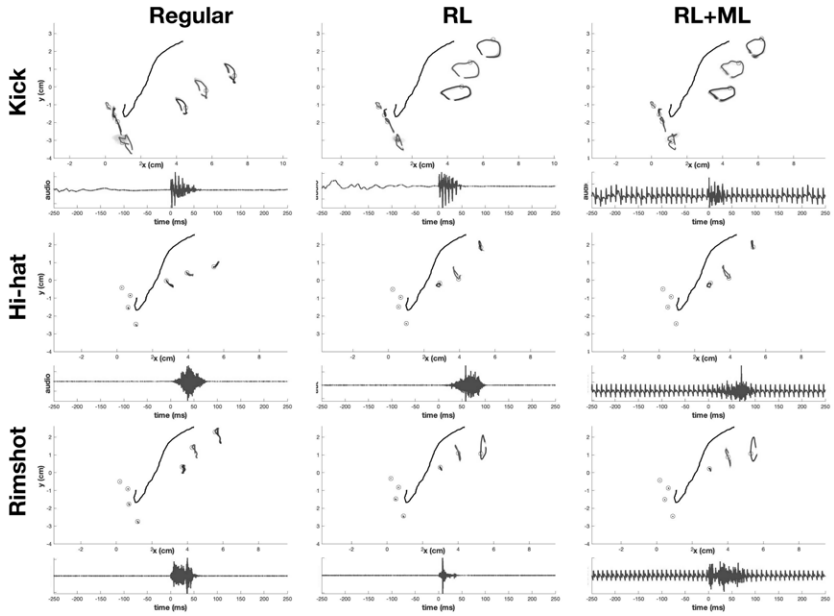


Figure 3: Articulatory trajectories of tongue and lip coils in the mid-sagittal plane during regular beatboxing and humming without (RL) or with (RL+ML) melodic line, for singer S04. For each sequence, audio signal of a representative token is plotted. Solid line: palate contour

IV. Discussion

Beatboxers naturally produced two humming versions of **P**, **t**, and **K**: one was the **RL** without **ML**, the other both **RL** and **ML**. We found that the term "humming HBB" does not imply the presence of a **ML**, but rather the choice of a particular articulatory strategy for the **RL** that is restrained to the oral cavity. This study showed that, while for regular **P**, **t**, and **K** breathing and articulatory behavior are related, with a likely glottalic or pulmonic initiation mechanism in most cases, the humming equivalents systematically switch to a velaric (mostly lingual) egressive or ingressive initiation mechanism. This bares two main consequences: on the one hand, humming boxemes are generally less intense than regular HBB boxemes; on the other hand, the use of an oral airstream to produce the **RL** allows for the dissociation of breathing and articulation. The high position of the back of the tongue divides the vocal tract into two functional sections that can produce two different sounds at the same time.

V. Conclusion

In humming HBB, the synchronous production of a rhythmic line and a melodic line is achieved by isolating the oral cavity from the rest of the vocal tract. The oral cavity functions on its own to produce the rhythmic line. Humming kick **P**, hi-hat **t**, and rimshot **K** are produced via velaric initiation mechanisms.

This leaves the upstream part of the vocal tract (laryngeal and pharyngeal spaces) available for breathing or producing the melodic line. In the latter case, the humming sound source generated by vocal-fold vibration is propagated into the nasal cavities. This is a skilful and original use of the vocal tract, regularly performed by beatboxers.

V. References

[1] R. Blaylock, N. Patil, T. Greer, and S. S. Narayanan. Sounds of the human vocal tract. In *Proceedings of Interspeech*, pages 2287–2291, 2017.

[2] P. Boersma. Praat: doing phonetics by computer. *http://www. praat. org/*, 2006.

[3] T. De Torcy, A. Clouet, C. Pillot-Loiseau, J. Vaissiere, D. Brasnu, and L. Crevier-Buchman. A video-fiberscopic study of laryngopharyngeal behaviour in the human beatbox. *Logopedics Phoniatrics Vocology*, 39(1):38–48, 2014.

[4] P. Helgason. Sound initiation and source types in human imitations of sounds. In *Proceedings of FONETIK*, pages 83–88, 2014.

[5] A. Paroni, N. Henrich Bernardoni, C. Savariaux, H. Lœvenbruck, P. Calabrese, T. Pellegrini, S. Mouysset, and S. Gerber. Vocal drum sounds in human beatboxing: An acoustic and articulatory exploration using electromagnetic articulography. *The Journal of the Acoustical Society of America*, 149(1):191–206, 2021.

[6] M. Proctor, E. Bresch, D. Byrd, K. Nayak, and S. Narayanan. Paralinguistic mechanisms of production in human "beatboxing": A real-time magnetic resonance imaging study. *The Journal of the Acoustical Society of America*, 133(2):1043–1054, 2013.

EVALUATING THE NASALISATION OF THE SINGING VOICE

N. Kotsani[1], E. Angelakis[1], A. Georgaki[1]

[1] Laboratory of Music Acoustics and Technology (LabMAT), Music Studies Department,
National and Kapodistrian University of Athens, Athens, Greece
nathalie@music.uoa.gr, angelakisv@music.uoa.gr, georgaki@music.uoa.gr

Abstract: **This study presents a vocal analysis prototype tool for the quantification of the nasality characteristic in singing. The tool is part of a larger, under development, software suite for the singing voice, which is aimed towards the analysis of both qualitative and quantitative vocal characteristics from specific audio samples. The tool examines the nasality characteristic by assessing formant central frequencies and bandwidths. In order to determine the most relevant acoustic parameters to be extracted and analysed by the tool, a case study experiment was conducted on two professional singers, singing in various degrees of willingly nasal voice. The audio samples recorded through this process were submitted to perceptual evaluation and acoustic analysis. Statistic analysis of the results demonstrate higher correlation coefficients between the nasality rating and the second formant frequency, followed by the first formant bandwidth. This prototype version relies on a regression model based on the above statistics.**

Keywords: **Singing voice; Nasality; Data processing; Formants; Voice Quality**

I. INTRODUCTION

Research on pitch accuracy and timbral analysis of the human voice has led to the development of a plethora of digital vocal analyzers [1, 2]. However, the variety of educational tools pertaining to the qualitative characteristics of the singing voice seems to be limited [3]. Our current bibliographic research revealed that this shortage seems to extend to software for singing voice qualitative analysis as a whole.

The present work introduces a vocal analysis tool targeted towards quantifying the nasality of the singing voice by assessing data from input audio samples of the user's voice. This could be an important contribution, as nasality is an important vocal factor in the evaluation of the overall voice quality, but also a determining factor for the stylistic authenticity of distinct vocal genres, sub-genres, schools of singing, vocal techniques, interpretational approaches and trends. This tool is part of a toolset under development that evaluates various characteristics of the singing voice, aiming to assist in a more spherical singing voice quality appraisal, including quantitative features (such as tonal accuracy, rhythmic consistency, and range) and qualitative characteristics.

Nasality is a perceptual characteristic of the voice, pertinent to the nasal tract [4], as a part of the vocal tract vocal 'filter'. The nasal tract is a complex resonator and has acoustic properties that have been extensively studied in the literature [5, 4]. Nasal tract resonance is associated with a reduction of the amplitude and a widening of the first formant's bandwidth [5]. There are also reports [4, 5, 6, 7] linking nasality with the first and second formant values and amplitudes, spectral tilt and cepstral peak prominence.

Dickson noted as early as 1962 [6] the variability in nasality acoustic features between experiment participants. This report was corroborated by more recent studies that resulted in similar conclusions [8]. Analogous findings have also been published on the nasality characteristic psychoacoustic domain, as Rusko [9] studied the perception of nasality in the human voice and musical instruments to conclude that, in parallel to spectral properties, nasality depends on the "similarity of dynamic changes to their aesthetic templates in speech, and of the formant structure to that of the vowels" [9].

Regarding the nasality in speech, Chen [10] attempted an objective quantification of vowel nasalization degree, through measurements on the acoustic signal, normalizing the parameters by adjusting for the influence of the vowel formant frequencies. Krol et al. [11] presented the preliminary results of a study analyzing the spatial distribution of energy in the acoustic field, using a multi-channel recorder, while Styler [7] reported findings pointing to both a spectral tilt, and a tilt of the first formant bandwidth, as nasality features for the English and French language. The latter [7] also proposed the A1-P0 acoustic parameter (i.e. "the amplitude of the first formant and the amplitude P0 of a nasal peak at low frequencies" [10]) as the most reliable indicator for use in nasality quantification and highlights the significant measurement values variability across speakers. Attuluri and Pushpavathi [12] found the "one third octave spectra analysis" to be an effective

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

hypernasality detection method for patients with a congenital disorder.

In 2007, Sundberg et al. [5] performed a study involving a CAT scan imaging of a singer's vocal tract and nasal cavity system, to determine the sound transfer characteristics of this model by means of sine-tone sweep measurements. Although this work focused on a single subject, the results were in agreement with pertinent, more recent studies. Specifically, for vowels employing the velopharyngeal opening (VPO) [8], there have been reports of: a) a widening of the first formant bandwidth [7], b) a decrease in the first formant's amplitude [13, 14, 15, 10, 16], and c) a strength enhancement of the spectrum partials near 3 kHz [16].

Gill et al. [8] concluded that a wide VPO is associated with a weaker fundamental and a lower level of the highest long-term average spectra (LTAS) peak below 1 kHz and a boost of otherwise low levels in the 24 kHz range. Vampola et al. [16] created a three-dimensional finite element model of the vocal tract for one female subject, for two vowels, proposing frequency bands for the four lowest nasal resonances and concluding in "more dominant acoustic energy" [16] in the region of formants F3–F5 due to nasality. Santoni et al. [17] investigated the effect of altered auditory feedback on the control of oral-nasal balance in song, showing lower nasalance scores in response to both increased and decreased nasal signal level feedback for the participants. Havel et al. [4] studied the sine-sweep response of 3-D models and concluded to a dip in the transfer function at the main resonance of the nasal tract with the VPO.

II. METHODS

For the detection of the nasality characteristic in audio samples the authors extended their previously developed Formant Range Profile (FRP) tool that analyzed the first two formants of the singing voice [18]. The prototype presented herein uses acoustic parameters pertinent solely to the voice formants. These parameters were selected through a case study experiment as the ones most correlated to vocal nasality, and a linear regression model was developed to utilize them for the estimation of a quantitative voice nasality factor.

Audio samples were recorded in a studio booth, by two professional singers with many years of operatic training, following a specified protocol. The protocol consisted of vocal trials on the non-nasal vowel "a" in order to control for the inherit nasality of certain consonants and vowels. Trials involved singing vowels on a) a constant frequency, and b) on two-octave ascending and descending glissandos and arpeggios. Singers performed these trials in three distinct experimental conditions [5], i.e. opting to use vocal sounds which they perceived as (i) nasal, (ii) non-nasal and (iii) progressively changing from nasal to non-nasal and back.

Participants 1 and 2 (P1, P2) performed the trials in 5min 42s, and 6min 32sec time, respectively. Recording was exported into 44.1KHz/16bit mono .wav files using the open software Audacity. File preparation was conducted by removing silent intervals and oral trial descriptions. These files were subsequently segmented into 228 2-second consecutively numbered .wav files.

The above 228 samples were evaluated perceptually by the same two experiment participants, utilizing their expertise as singing teachers. Evaluation was performed aurally on a 9-point scale, from 0 (non-nasal) to 4 (extremely nasal), including 0.5 step ratings. Files were presented to the two judges in distinctively randomized orders and with the use of closed type headphones. Evaluation means for each sample had thus discreet values ranging from 0 to 5 with a 0.25 step.

Aiming towards a more homogeneous class separation of the dataset, and due to the perceptual ratings of two judges resulting to a 0.25 step nasality ratings, samples were classified into five classes with an experiment-specific approach. The first and the fifth class included 3 rating values each (0, 0.25, 05 and 3.5, 3.75, 4 respectively), leaving the rest with 4 rating values each (< 0.75, <= 1.6, <= 2.4, <= 3.2 <= 4).

For each audio sample, the average of the first four formant frequency values (F1, F2, F3, F4) and bandwidth values ($F1_{BW}$, $F2_{BW}$, $F3_{BW}$, $F4_{BW}$) were extracted. The values were computed using python programming language and the python library parselmouth, which uses the Praat's software functions. In addition, the Pearson's correlation coefficients and the P-factors of those features, between the perceptual nasality ratings, were extracted.

III. RESULTS

The regression lines between nasality perceptual ratings (NR) and the features $F1_{BW}$, $F2_{BW}$, $F3_{BW}$, $F4_{BW}$, F0, F1, F2, F3, F4, are shown in Figure 1 (i-ix), respectively. Table 1 shows the average values of the acoustic parameters in correspondence with the five-class nasality perceptual ratings. Table 2 shows the Pearson's correlation coefficients along with the P-factors. We observe that, as the perceptual nasality rating increases, the $F1_{BW}$, F1 and F2 are increasing as well, while the F3 is decreasing.

The above results demonstrate higher correlation coefficients between the nasality rating and the second formant frequency (r = 0.5769, P-factor = 0.0000), followed by the first formant bandwidth (r = 0.4767, P-factor = 0.0000). The experimental protocol of this study included singing phonation in a variety of distinct pitches and on vocal glissandos. Therefore, in order to control for possible fundamental frequency (F0)

effect on nasality assessment, the Pearson's correlation coefficient between the perceptual nasality ratings and F0 was extracted. F0 proved to be a non-statistically significant factor for vocal nasality, with a correlation coefficient equal to 0.1081 and a P-factor of 0.1035.

Based on the above observations, the statistically significant factors F1, F2, F3, $F1_{BW}$ and $F3_{BW}$ were selected, yielding, using linear regression, a correlation coefficient of 0.8112 with P-factor equal to 0.0000.

IV. Discussion

Our results confirm nasalization of the singing voice to be a feature that appears identifiable through acoustic analysis of audio samples. The studied features included the frequency and bandwidth of the first four formants and the results are in accordance with the literature. More specifically, the analysis conducted confirms the increase of the first formant bandwidth in nasal phonation, demonstrating the highest correlation coefficient among the examined formant bandwidths.

The FRP tool was extended using stepwise regression, to extract the aforementioned features and enriched to output a diagram visualizing an overall nasality evaluation of the audio input samples. The proposed tool can be used either (a) independently, assisting singing students identify the nasalization of their voice, become familiar with the vocal tract modifications, and gain control of the articulatory mechanism and VPO use in singing, or (b) as part of the under development comprehensive voice quality assessment toolset. This toolset is a part of the "Assistance for students in Singing and Music Aesthetics" (ASMA) project, involving the authors of the present work, which aims towards the amelioration of the vocal and music education process in Greek primary education, by providing comprehensive information, training and teaching guidelines, as well as suitable tools, to the elementary school music teachers.

Future work on this project includes an extension of our tool using more acoustic parameters, enhancement of the existing model using a large scale dataset with participant groups of distinct vocal proficiency levels, as well as the development of a multivariate model predicting the nasality rate of the singing voice, using neural networks.

V. Acknowledgments

The research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant" (Project Number: HFRI_FM17_TΔ E_3832).

References

[1] Y.L. Shue, P. Keating, C. Vicenik, and Y. Kristine. Voice-Sause: A Program for Voice Analysis. Proceedings of the 17th International Congress of Phonetic Sciences, volume 3, pages 1846–1849, Hong Kong, 2011.

[2] M.S. Morelli, S. Orlandi, and C. Manfredi. BioVoice: A multipurpose tool for voice analysis. Biomedical Signal Processing and Control, 64, 2 2021.

[3] E. Angelakis, G. Kosteletos, A. Andreopoulou, and A. Georgaki. Development and Evaluation of an Audio Signal Processing Educational Tool to Support Somatosensory Singing Control. In Audio Engineering Society Convention 145. Audio Engineering Society, 2018.

[4] M. Havel, J. Sundberg, L. Traser, M. Burdumy, and M. Echternach. Effects of Nasalization on Vocal Tract Response Curve. Journal of Voice, 2021.

[5] J. Sundberg, P. Birch, B. Gümoes, H. Stavad, S. Prytz, and A. Karle. Experimental Findings on the Nasal Tract Resonator in Singing. Journal of Voice, 21(2):127–137, 3 2007.

[6] D. R. Dickson. An acoustic study of nasality. Journal of Speech and Hearing Research, 5(2):103–111, 1962.

[7] W. Styler. On the acoustical features of vowel nasality in English and French. The Journal of the Acoustical Society of America, 142(4):2469– 2482, 10 2017.

[8] B. P. Gill, J. Lee, F.M.B. Lã, and J. Sundberg. Spectrum Effects of a Velopharyngeal Opening in Singing. Journal of Voice, 2018.

[9] M. Rusko. Towards a More General Understanding of the Nasality Phenomenon. Technical report, 1996.

[10] M.Y. Chen. Acoustic correlates of English and French nasalized vowels. The Journal of the Acoustical Society of America, 102(4):2360–2370, 1997.

[11] D. Krol, A. Lorenc, and R. Swiecinski. Detecting laterality and nasality in speech with the use of a multi-channel recorder. In ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing-Proceedings, volume 2015-August, pages 5147–5151. Institute of Electrical and Electronics Engineers Inc., 8 2015.

[12] N. Attuluri. Correlation of Perceived Nasality with the Acoustic Measures (One Third Octave Spectral Analysis & Voice Low Tone to High Tone Ratio). Global Journal of Otolaryngology, 8(3), 6 2017.

[13] P. Delattre. Les attributs acoustiques de la nasalité vocalique et consonantique. Studies in French and Comparative Phonetics, pages 257–263, 1966.

[14] A.S. House and K. N. Stevents. Analog studies of the nasalization of vowels. The Journal of speech and hearing disorders, 21(2):218–232, 1956.

[15] G. Fant. Acoustic Theory of Speech Production. DE GRUYTER, 12 1971.

[16] T. Vampola, J. Horácˇek, V. Radolf, J.G. Švec, and A.M. Laukkanen. Influence of nasal cavities on voice quality: Computer simulations and experiments. The Journal of the Acoustical Society of America, 148(5):3218–3231, 11 2020.

[17] C. Santoni, G. de Boer, M. Thaut, and T. Bressmann. Influence of Altered Auditory Feedback on Oral-Nasal Balance in Song. Journal of Voice, 34(1):9–157, 2020.

[18] A. Andreopoulou, N. Kotsani, G. Dedousis, and A. Georgaki. Evaluating the vocal characteristics of elementary school students: Basic assessment tools and methodology. In Proceedings of Interaction Design and Children, IDC 2021, 216–223. Association for Computing Machinery, Inc, 6 2021.

Table 1, Average values of the formant bandwidths ($F1_{BW}$ - $F4_{BW}$) and formant central frequencies (F1-F4) in correspondence with the five-class nasality perceptual ratings (NR).

NR	$F1_{BW}$	$F2_{BW}$	$F3_{BW}$	$F4_{BW}$	F0	F1	F2	F3	F4
0	191.93	333.71	495.15	381.13	233.49	794.89	1334.58	2835.94	3468.94
1	200.53	242.74	766.86	356.32	257.09	814.53	1357.61	2680.6	3459.73
2	281.15	208.18	544.11	490.42	197.61	858.81	1366.79	2748.16	3736.86
3	355.85	263.34	721.57	535.03	231.03	880.78	1490.02	2678.92	3521.92
4	382.65	340.69	706.58	422.82	271.12	933.81	1688.95	2531.32	3435.98

Table 2, The Pearson correlation coefficients and the P-factors between the NR and the studied acoustic parameters.

	r	P-factor		r	P-factor
			F0	0.1081	0.1035
$F1_{BW}$	0.4767	0.0000	F1	0.4522	0.0000
$F2_{BW}$	0.0115	0.8628	F2	0.5769	0.0000
$F3_{BW}$	0.1639	0.0132	F3	-0.3645	0.0000
$F4_{BW}$	0.0994	0.1345	F4	-0.0413	0.5351

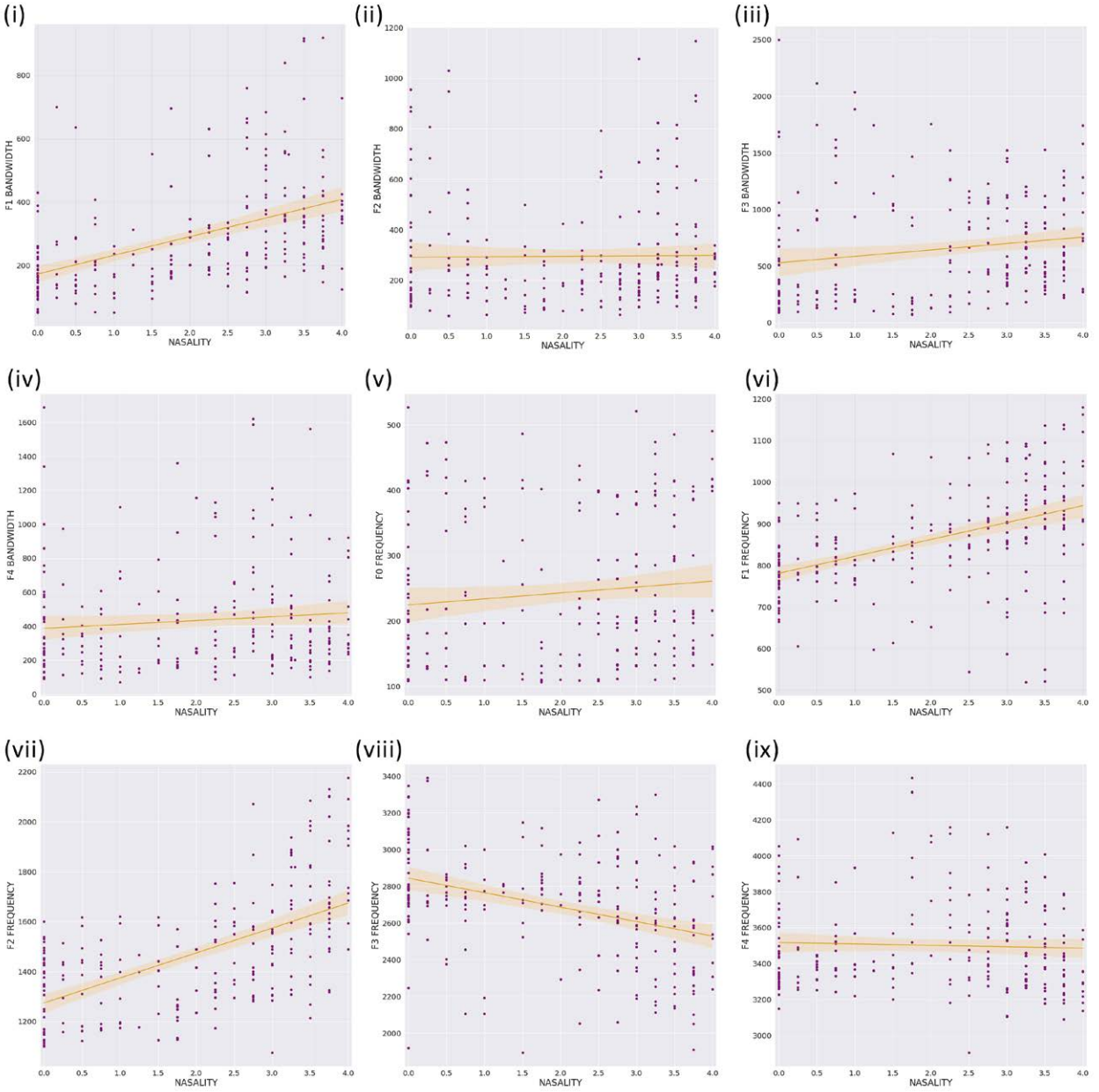


Fig. 1, Regression lines (i-ix) between the nasality perceptual ratings and the features $F1_{BW}$, $F2_{BW}$, $F3_{BW}$, $F4_{BW}$, F0, F1, F2, F3, F4, respectively.

DOES THE CROSSING OF H2 AND F1 AS PITCH CHANGES AFFECT THE PERCEPTION OF HOW OPEN/COVERED THE VOICE TIMBRE APPEARS?

A. Vurma[1]

[1] Estonian Academy of Music and Theatre, Tallinn, Estonia
allan.vurma@eamt.ee

Abstract: **According to the aesthetics of classical singing style, vocalists should "cover" the vowels at pitches in the so-called *passaggio* region. Some authors [1][2] have recently claimed that "covering" is primarily an acoustic illusion when the vocalist tries to avoid the reflexive wider opening of the mouth and raising of the larynx as the pitch ascends to the passaggio. If the singer keeps the vocal tract resonance frequencies or formants (R) invariant, the voice timbre seems "open" if at least two harmonics locate lower than the R1, and "covered" if this concerns only the fundamental. The purpose of this study was to check these assumptions by the use of perception tests. In the case of all the vowels investigated, except /i/, there was a statistically significant tendency to rate the timbre of sounds as more "covered" when the pitch was higher, and more "open" when the pitch was lower, without the expected abrupt changes at those pitches where the H2 passed the R1. Rather than depending on the H2 crossing the R1, the change in perceived timbre with pitch may be related to the expectations the f_0 creates on the apparent characteristics of the singer, including the values of formant frequencies.**

Keywords: **Voice covering, formants, harmonics, timbre, pitch**

I. INTRODUCTION

A "covered" (Italian *coperto*, German *gedeckt*) voice is a quality that is typically pursued in classical singing style (especially in the male voice in its *passaggio* region) in order to avoid a strident, shrill, screeching sound similar to yelling that is sometimes also called "open" or "white" timbre [3]. The aesthetics of "covering" the voice when singing in the *passaggio* region probably emerged in the 1840s. There are various opinions as to how the voice "covering" should be accomplished. Manuel Garcia, a world-famous voice teacher of that period, claimed that to "cover" the voice, the singer has to keep the larynx in a low position [4]. This suggestion contradicted the common practice of opera singers of those times who typically let the larynx rise with pitch [5]. Also, in untrained singers, the larynx usually rises as the pitch ascends [6].

Miller [3] and Appleman [4] relate "covering" to vowel modification, which should typically be in the direction of a back vowel and a wider mouth opening with ascending pitch. Sundberg [7] claims that to "cover" the voice the first resonance of the vocal tract (R1) of the vowels with high R1 such as /a/ and /ae/, and the second resonance of the vocal tract (R2) of the vowels with low R2 such as /i/ should be lowered. This is possible by reducing the mouth opening (which however, contradicts the aforementioned suggestions of Miller and Appleman) and widening the pharynx (ibid.).

Miller and Schutte [8] state that the essence of voice "covering" is the position of R1 in relation to the second harmonics of the voice spectrum (H2): we perceive that the voice is "covered" if the R1 is located below the H2. Their statement was based on the results of a case study where the first author of the paper was trying to sing vowels with "open", "standard" and "exaggeratedly covered" timbres.

"Open"/"white" or "yell" timbre, which is the opposite of "covered"/"closed" timbre, may be related to the articulatory formant tracking where the singer tunes R1 to the H2—a strategy which people often instinctively tend to apply to increase the loudness of the voice [1][9]. Although classically trained singers tend to avoid "yell" or "open" timbre, for many commercial styles such as pop, jazz, belting or folk it is a legitimate quality, as in such styles naturalness is typically sought, and "yell" is a natural, speech-like way similar to how people use their voice in emotional situations.

Miller [1] and Bozeman [2] claim that the "covered" timbre is perhaps, at least partly, an acoustic illusion which emerges when the singer intentionally resists the temptation to track the H2 by the R1 at the *passaggio*. In order to achieve this, when reaching the pitches where H2 rises above the typical R1 for the corresponding vowel, the singer has to let the H2 pass (turn over) the R1 (by avoiding doing anything with the articulatory organs).

In the case of another strategy often used spontaneously—"hoot" [1] or "whoop" [9]—the R1 is tuned to the H1. As H2 = 2H1, the "hoot"/"whoop" is

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

possible at about one octave higher than “yell”, and its addressing is not the focus of this paper.

To summarise: several authors such as Bozeman [2][9] and Miller [1] claim that when R1 is located below the H2, the timbre of the sung vowel seems “covered” or “closed” (according to Miller, this mainly concerns male voice singing in the *passaggio* region) and timbre feels “open” when the R1 is located at or above at least two spectral harmonics. However, the systematic investigation using statistical tools and perception tests with a wider range of participants to prove such claims would appear to be lacking. This study is an attempt to fill this gap.

II. Methods

We used perception tests in which we played singing-voice-like synthesized stimuli to a group of experts and asked them to rate the timbre of each stimulus on the scale “open” – “covered”/“closed”.

Stimuli: Using the program *Madde 3.0.0.2* we synthesized nine series of short (about 2 sec) auditory stimuli. The 5 vowels (/a/, /e/, /i/, /o/, /u/) × 2 voice categories (bass, soprano) paradigm was used to create the series. Each series consisted of 11 to 20 pitches of the chromatic scale depending on the vowel and the voice category used in each specific series. The pitch ranges were chosen to include the region where we expected to meet an abrupt change in the experts' ratings from “open” to “covered” because of the H2 turning over the R1 (i.e., the stimuli where H2 = R1 and thereabouts). For all the stimuli included in the same series, we kept the frequencies and bandwidths of the VT resonances and other parameters that it is possible to specify in the *Madde* program (with the exception of the fundamental frequency, f_0) invariant.

The values of the VT resonance frequencies used as the input parameters for the synthesis by *Madde* were estimates of the VT resonance frequencies of the voice spectrum of two real singers (a soprano and a bass) singing the same vowels at the lower end of their comfortable voice range. The measurements were made using the software *Praat 6.1.08*.

Experts: Altogether we had 44 experts, whom we contacted either in person or by distributing the internet links to the tests online. For sharing the links, we used the personal Facebook contacts and the postings to the Acoustic Vocal Pedagogy Group. In the online tests we asked the participants to deliver some information about their background. The reported ages were between 21 and 67 years (45 years on average). All the participants were or had been professional singers or were still voice students at the tertiary level of music education (10 people); 18 were also active as voice teachers (including nine professors at academic institutions). Most of the experts had a background in the classical style. The reported countries of origin included Estonia, the USA, the Netherlands, Finland, Latvia and Argentina.

Tasks: In the tests, the experts had to give their responses on a 5-point Lickert scale where the rating “1” corresponded to the most “open” and “5” to the most “covered/closed” timbre, with the midpoint at the rating “3”. The tests were administered via the online platform *PsyToolkit* [10][11].

The participants were also asked to give comments on the thoughts that running the experiments may have created. As not all the participants had enough time to complete all nine tests, the number of participants who completed each test was different, and remained between 28 and 44.

The participant could run the test on any convenient device such as desktop, laptop, tablet or mobile (they were asked to give information about their equipment in a special box on the screen). They were asked to use earphones if they did the tests on a mobile so as to ensure the sound quality. The experts chose the order and the time they did the tests themselves.

III. Results

In the panels of Figure 1, we see that (1) in all tests, except the female /i/, the linear trendline is rising with pitch indicating that the experts tended to rate the timbre as more “covered” at higher pitches, and as more “open” at lower pitches; (2) no one of the graphs showed the hypothesized abrupt changes around the region where the H2 passes over the R1 (H2 = R1).

Inter-individual differences: While 42% of the trendlines rose by at least one grade-point with one-octave pitch ascent, in the case of 13% of the trendlines, the tendency was opposite to the general trend—at higher pitches, the timbre was rated as more open compared to the low-pitch notes. Moreover, 45% of the trendlines lacked a clear trend, or it was quite weak.

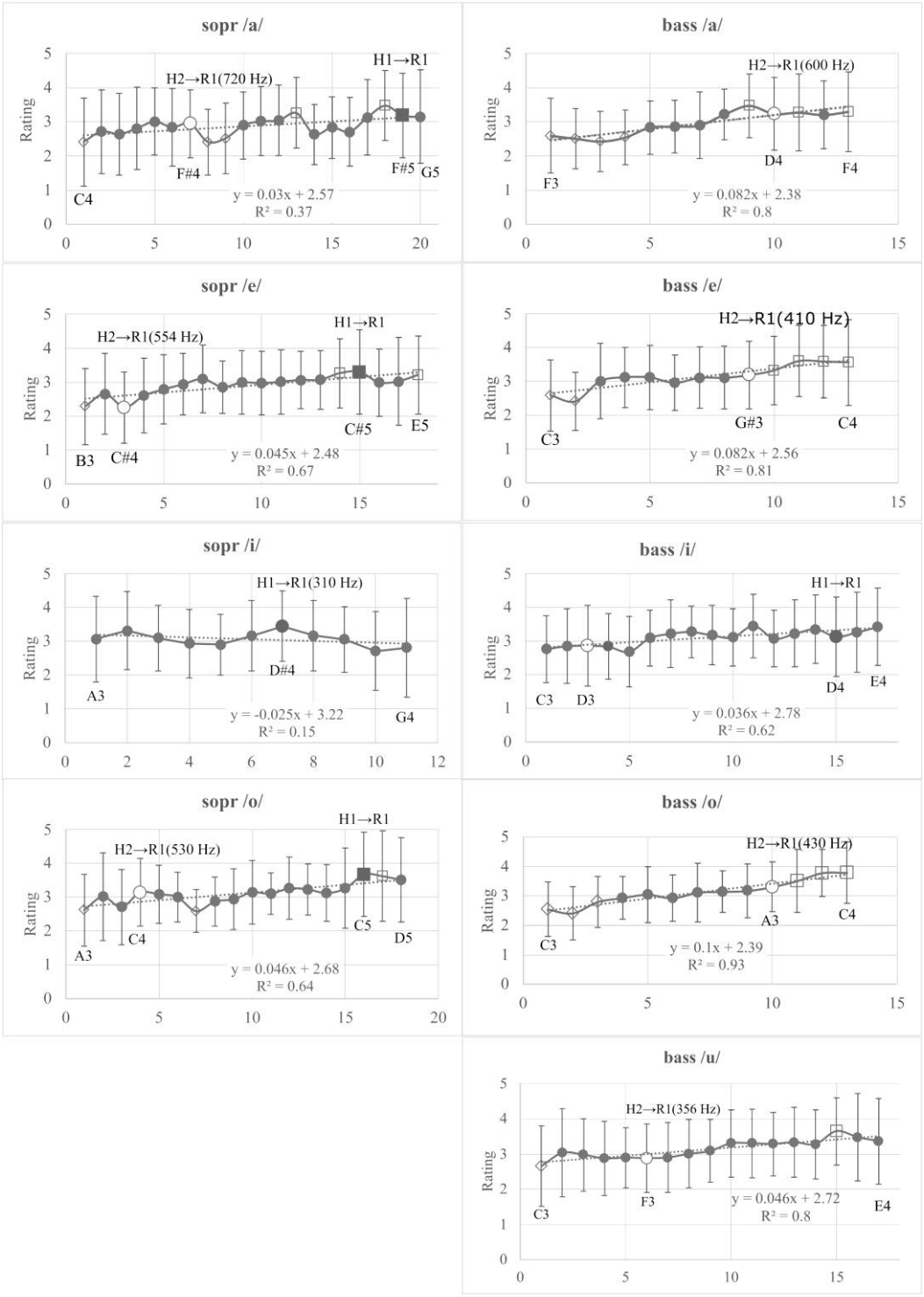


Figure 1. Average ratings with standard deviation whiskers to the timbre of stimuli on the scale "open" – "covered". Each panel corresponds to a separate test. The horizontal axes show the chromatic scale steps. Empty ring markers indicate pitches at which H2 crosses over the R1, dark enlarged markers indicate pitches at which H1 crosses the R1. The difference in ratings to the stimuli at pitches marked with diamonds at the beginning and with squares at the end of the corresponding pitch scale is statistically significant according to the *PostHoc* tests of ANOVA. Also, the linear trendlines of the curves with the corresponding formulas and correlations are presented.

IV. Discussion

Although our perception tests showed a statistically significant tendency of voice specialists to perceive the timbre of sung vowels with invariant VT resonance frequencies as more "open" at low pitches and as more "covered"/"closed" at high pitches, this tendency was often small or absent; moreover, contrary to our hypothesis, such timbral change seems not to be related to the specific pitch at which the H2 turns over the R1 (i.e., where H2 = R1) as the change from the "open" to the "covered/closed" timbre with rising pitch was expressed by a smooth statistical trend and not by an abrupt jump or change. Such finding compelled us to seek alternatives to the explanation that was used to describe the phenomenon by Miller and Schutte [8] and by Bozeman [2][9]. The results of our study concur with the findings of Traunmüller [12], who found that the decisive factor which defines how "open" the timbre of the vowel seems to the listeners is the tonal distance between the R1 and the fundamental component of the spectrum (f_o = H1).

Another possible hypothesis suggests that the f_o's impact on the perceived "openness" of the vowel could be related to the expectations the f_o may create in listeners with regard to the frequencies of the VT resonances (vowel formants) [13]. In general, shorter people typically have somewhat higher vocal tract resonances than taller people since their vocal tracts tend to be physically smaller, and they produce a higher f_o with their voice in common situations like speaking as their vocal folds tend to be shorter.

Therefore, when we hear a human voice with a high pitch, we subconsciously expect that the timbre of such a voice also has formants at somewhat higher frequencies than a voice with a lower f_o. We may speculate that in the case of our tests the subconscious expectations of the experts on the formant frequencies matched quite well the actual acoustical parameters of our stimuli at the low end of the pitch scales used in our tests (the values of the VT resonance frequencies used to synthesize our stimuli corresponded to the values produced by a real bass and soprano). This was probably less so at high pitches, which might enforce the impression of voice "covering".

V. Conclusions

Pitch is able affect the perceived timbre of the voice on the scale "open" – "covered"/"closed". However, the content of the corresponding terminology may be understood even in opposite ways by different users, for quite a number of whom the terms do not appear to have a consistent meaning. The specific acoustic condition H2 = R1 seems not to play a substantial role in changing the perception from one timbral category to the other.

References

[1] D.G. Miller, *Resonance in Singing*. Princeton, NJ: Inside View Press, 2008.

[2] K.W. Bozeman, *Kinesthetic Voice Pedagogy*. Gahanna, OH: Inside View Press, 2017.

[3] R. Miller, *On the Art of Singing*. Oxford: Oxford University Press, 1996.

[4] D.R. Appleman, *The Science of Vocal Pedagogy*. Bloomington, IN: Indiana University Press, 1986.

[5] G.W. Bloch, "The pathological voice of Gilbert-Louis Duprez," *Cambridge Opera Journal*, 19(1), pp. 11–31, 2007.

[6] J. Stark, *Bel Canto*: *A History of Vocal Pedagogy*. Toronto: University of Toronto Press, 1999.

[7] J. Sundberg, "Perception of singing," in. *The Psychology of Music*. D. Deutsch Ed. San Diego, CA: Academic Press, 2013, pp. 69–100.

[8] D.G. Miller, and H.K. Schutte, "Toward a definition of male 'head' register, passaggio, and cover in western operatic singing," *Folia Phoniatr Logop*, 46, pp. 157–170, 1994.

[9] K.W. Bozeman, *Practical Vocal Acoustics: Pedagogic Applications for Teachers and Singers.* Hillsdale, NY: Pendrago Press, 2013.

[10] G. Stoet, "PsyToolkit - A software package for programming psychological experiments using Linux," *Behavior Research Methods, 42(4)*, pp. 1096–1104, 2010.

[11] G. Stoet, "PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments," *Teaching of Psychology, 44(1)*, pp. 24–31, 2017.

[12] H. Traunmüller, "Perceived dimension of openness in vowels," *J Acoust Soc Am,* 69(5), pp. 1465–75, 1981.

[13] T.M. Nearey, and P.F. Assmann, "Probabilistic "sliding template" models for indirect vowel normalization," in *Experimental Approaches to Phonology*, M.-J. Solé, P.S. Beddor and M. Ohala Eds. New York, NY: Oxford University Press. 2007, pp. 246–269.

FORMANT TUNING IN CRETAN RIZITIKO SINGING

S. Kalozakis[1], A. Georgaki[2] , G. Kouroupetroglou[3]

[1] National and Kapodistrian University of Athens/Music Department, Panepistimioupolis-Zografou, 157 84, Athens, Greece
Hellenic Mediterranean University/Department of Music Technology and Acoustics, 74133, Rethymnon, Greece

[2] National and Kapodistrian University of Athens/Music Department, Panepistimioupolis-Zografou, 157 84, Athens, Greece

[3] National and Kapodistrian University of Athens/Department of Informatics and Telecommunications, Panepistimioupolis- Ilisia, 157 84, Athens, Greece

Email addresses: kalozakis@hmu.gr[1], georgaki@music.uoa.gr[2], koupe@di.uoa.gr [3]

Abstract: **Recent research into different folk vocal styles has been conducted by examining the acoustic parameters of the singing voice [1] [2] [3] [4]. On the Greek island of Crete, the acoustical parameters of a song are "strongly" depended upon the origin of the singer, due to the peculiar pronunciation which each Cretan region adopts. The most known non-dance folk songs in Crete are "Rizitika" songs. (plural of "Rizitiko" song). Although Rizitiko has a distant chronological root (root in Greek means Riza which is etymologically related to Rizitiko) is a living culture, a dynamic legacy and heritage that is spread all over Crete and mainly at the western and central regions of the island.**

In this paper, we research thoroughly and present the formant characteristics of the Cretan Rizitiko singing style sung by sixteen (16) men. Specifically, we demonstrate (via illustrative panel) the formant tuning of two (2) singers whose origin belongs to different Cretan region. Also, we compare the vocal acoustical differences of formant frequencies between all participating singers for one singing diphone ("ki") and for one singing vowel ("a")

Keywords: **Rizitiko, Formant Tuning, Formant Frequency**

I. INTRODUCTION

Folk music is synonymous with traditional music. The island of Crete is a geographical part of Greece that still supports and embellishes its traditional identity [5]. Tradition, from the Latin verb tradire (to deliver) can be delivered (among other elements) through song, music and dance. These three are "strong" elements of the Cretan tradition [6]. The traditional music of a region can often be divided into dance and non- dance music. The most known non-dance folk songs in Crete are Rizitika (solemn slow songs, possibly of Byzantine origin) [7]. Rizitika songs are strong and a dynamic symbol of Cretan identity. There is a significant difference in pronunciation between the Prefectures of Crete, as each Cretan region adopts a characteristic pronunciation. This characteristic pronunciation becomes more noticeable with the use of velar consonants (such as k/x/g) which is followed by anterior vowel (i/e/ou) [8]. Particularly, the case of Diphones (consonant with a following vowel) meets the most characteristic element of the Cretan pronunciation that is not eliminated [9]. The question that arises is whether this characteristic pronunciation has an impact on the singer's voice. By this we mean, how this feature (pronunciation) is captured and imprinted during the performance of the modern Cretan singing voice, the existence (or not) of Formant Tuning along with the classification and distribution of formant frequencies by region under consideration.

II. METHODS

Sixteen (16) singers were recorded from four different regions. Specifically, these were three (3) Cretan regions Chania (or as pronounced Xania) Rethymnon and Heraklion. In these Cretan Counties the Rizitiko is found especially in the province of Xania [10]. The fourth region was Athens which is non-Cretan. The reason of recording two non-Cretan singers was to find how their singing (the acoustical and musical parameters) differentiates compared to Cretan singers and what is the "role" of the origin and how it functions to the "interpretation" of Rizitiko.
From Xania four (4) singers recorded as well as from the region of Heraklion. From Rethymnon six (6) singers recorded and from Athens two (2). All studio recordings were made in the same acoustic environment, in order

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

to achieve the correct comparison of the acoustic measurements that would follow. The equipment used was state of the art and it was identical to all recordings.

The equipment consisted of a condenser microphone with omnidirectional polar pattern (Earthworks Audio M30) with direct response across the frequency spectrum, keeping the same "working distance" (the distance between the singer and the microphone) which was thirteen centimeters (13 cm) for each session.
The microphone was connected to the low-noise preamplifier (Avalon M5 Pure Class A). Similar to previous studies [11] electrodes were placed to all singers during recordings. These were placed externally on the neck of the singers at the height of the thyroid cartilage to detect impedance changes of the vocal cords and were connected to the Electroglottograph Console (Kay Pentax Model 6103). The signal of the electroglottography (EGG) was stored as a "mono" signal to the digital recorder (Tascam X-48MKII hard-disk recorder). The digital recorder (using 44.1Khz sampling rate/16-bit resolution) was in connection with the mixing console (Audient ASP 8024) in order to have during playback the two signals (microphone and electroglottography). All recordings accomplished at the Studio of the Department of Music Technology and Acoustics, Rethymnon- Crete.

The whole recording process was identical for all participating singers. All singers were exclusively male. This was due to the fact that Rizitika songs always performed by male voices, in contrast to other vocal styles dominated by female voices [12]. The method included the following procedure:

Firstly, all singers (after they understood the procedure that would follow) filled in a questionnaire regarding their origin, where they grew up, age, musical studies, years of experience as performers and their discography, if they are smokers etc. After filling the questionnaire:

- It was found with the use of piano the singer's voice extend (registro) in order to be categorized the voice of all singers (tenor)
- Note selection of the Rizitiko by the singers
- All singers performed the same Rizitiko song "Se Psilo Vouno"
- The singers performed a major scale for all Greek vowels (a/e/i/o/ou) for both ascending and descending form (musical scale)
- After completing the recordings, the pre-selected Cretan characteristic singing diphones, all performed vowels, as well as the recitation of the lyrics of the performed Rizitiko song, isolated

in order to proceed data mining (using Praat software program).

Praat software program developed by Paul Boersma and David Weenick from the Institute of Phonetic Sciences of the University of Amsterdam.
Most of our measurements used in our analysis were acquired using the PRAAT software, as it is a valuable and flexible software tool in the field of phonetics and voice analysis [13]. This program can handle large audio files and extract measurements of the vocal parameters using its built- in function.

Mostly intensity, pitch and formant analysis were used in our measurements, as mentioned. More specifically, for the functions of pitch and formant analysis, PRAAT uses an algorithm that performs an acoustic periodicity detection trough a precise autocorrelation method and also, it can capture a value every 6.25 millisecond, giving the average value for the formant frequency we aim to find, respectively.

III. RESULTS

Formant frequencies initially differ depending on age and gender, as the anatomical features of the vocal tract (length) depend on the above two factors. That is why all singers had the same gender (male) and similar age range (32 to 43 years old).

Each of the preferred resonating frequencies of the vocal tract is known as a formant. In the vocal tract the five (5) lowest formant frequencies (usually referred to as F1 for the first, F2 for the second etc.) play a role in shaping the spectrum of the voice and the timbre of the voice (sound color).

Formant frequencies differ from the vowel that is pronounced each time, since the position and the shape of the tongue, the lips, the soft palate, the jaw, is on the substance, the articulatory movements of the face. The formant frequencies depend on the articulatory movements.

The position of the articulators affects only the first two formants (henceforth F1, F2) so the quality or recognizability of the vowel depends on the first two formants [14]. F1 is more susceptible to the changes of the jaw [14] [15]. Specifically, as long as the jaw opens, it increases the frequency of the first Formant and vice versa. F2 is more susceptible to the changes of the tongue [14]. When the tongue compresses the upper part of the vocal tract, it occurs a frequency increment of the second formant, or if we simplify it, F2 is mostly determined by the frontness/backness of the tongue body.

"Fig. 1, 2" show the average value of F1 and F2 for the vowel "A" and the diphone "KI" respectively. It appears

that singers whose origin is from Rethymnon, compared to other regions, use a smaller opened jaw position resulting to a low F1 value (below 500 hz at the diphone and well below 600 hz at the examined vowel). It can be easily ascertained that Rethymnon singers, have the lower F1 value for both vowel and diphone. Singers from Athens have obviously significantly higher F1 value for the vowel "A" something that is interpreted in a larger jaw opening compared to all Cretan singers. Vowel "A" is an "open" vowel and emphatically characterizes the opening of the jaw.

At "Fig. 1, 2" we can see as well that singers from Heraklion, Chania (or Xania) and Athens show interesting similarity to F1 value at diphone "KI". Comparing Cretan regions only, we see that Heraklion provided a higher F2 value at vowel "A" whilst, Chania at diphone "KI". The latter, is due to the fact perhaps that at the region of Chania this characteristic diphone [8] [9] ("ki") pronounced more "strongly" and has greater intensity.

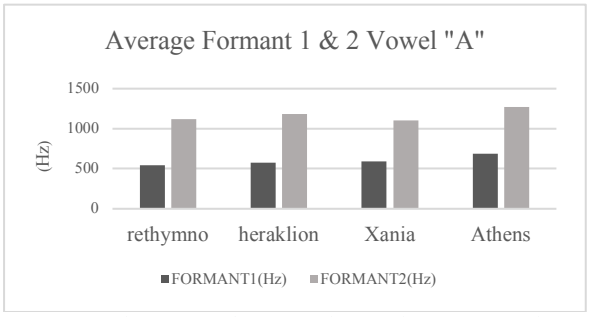


Fig. 1 Vowel "A". On the vertical axis: the average values of F1, F2 sorted by region with intense and pale color respectively. On the horizontal axis: the frequencies (hz)

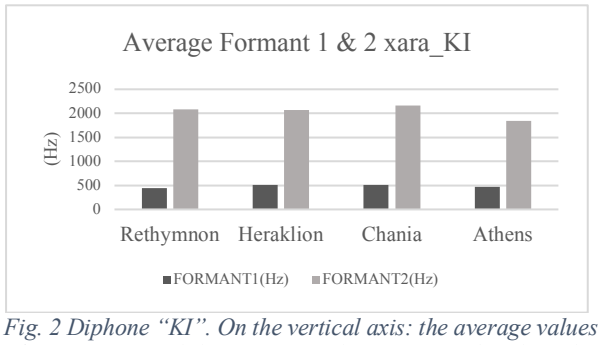


Fig. 2 Diphone "KI". On the vertical axis: the average values of F1, F2 sorted by region with intense and pale color respectively. On the horizontal axis: the frequencies (hz)

Having mined the values of the first two formants for all singers, our main objective was to investigate whether Rizitiko singers apply formant tuning. Formant tuning suggests that a singer increases Sound Pressure Level (SPL) without expense of vocal effort by adjusting his lower formant frequencies to coincide with partials (harmonic frequencies) in order to gain SPL. By doing this, a singer exposes his voice and can be heard with less vocal effort in large auditoria [15] [16] [17]. In many cases, singers can adjust the articulation of the vocal tract (formant tuning) in order to enhance and gain acoustic output [17]. Sometimes, singers tune their two lowest formant frequencies (F1, F2) to coincide harmonic partials in order to increase the audibility of the voice [18].

Earlier literature in formant tuning, considered that in order to be occurred formant tuning F1 and F2 must be tuned to a partial, either F1 is tuned to the fundamental frequency (f0) or F1 is tuned to the vicinity of a partial. In the latter, previous studies considered that vicinity between F1 or F2 is either over a semitone (100 cents) of a partial, or under (below) a semitone of a partial. [19] [13].

In the present study we consider formant tuning occurs if the F1 and F2 has maximum one semitone distance (above or below) a partial, or F1/F2 is tuned exactly at a partial. "Fig. 3, 4" represent a typical formant tuning phenomenon for two Cretan singers (from Rethymnon and Chania respectively) performing a major scale (ascending/descending form) singing vowel "A".

Rethymno singer at "Fig. 3" produces formant tuning since F1 (lower curve) is in most cases aligned with the third harmonic (H3). More specifically, at dominant, submediant and supertonic F1 is tuned exactly on H3. F2 (upper curve) formant tuning extends from H4 to H8.

F1 (lower curve) of the singer from Chania at "Fig. 4" is in almost complete alignment with the partials (harmonics). In fact, his F1 "follows" the performing note along the third harmonic (H3). At his F2 (upper curve) we observe evidence of formant tuning as well.

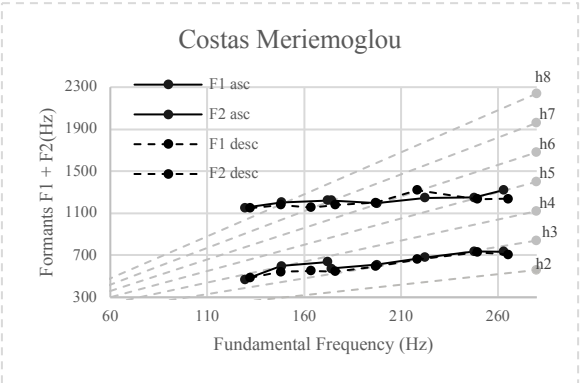


Fig. 3 Singer from Rethymno performing a major scale singing vowel "A". The continuous and dashed lines represent the ascending and descending form respectively. The oblique dashed lines are the partials (harmonics). Lower and Upper curve represent F1 and F2 respectively.

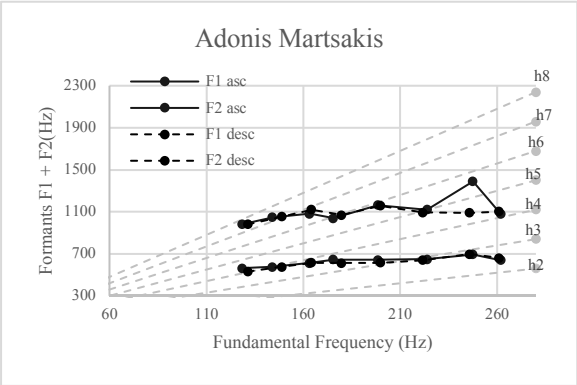


Fig. 4 Singer from Chania performing a major scale singing vowel "A". The continuous and dashed lines represent the ascending and descending form respectively. The oblique dashed lines are the partials (harmonics). Lower and Upper curve represent F1 and F2 respectively

IV. Discussion

In order to find evidences of formant tuning in modern Cretan singing and the classification and distribution of formant frequencies by Cretan region under consideration, we presented primarily our formant analysis and the results of our measurements. Despite the fact that this technique (formant tuning) more frequently appears among opera singers, none of the recorded Rizitiko singer has undergone operatic training. More vowels and diphones will be analyzed soon, to draw "solid" conclusions.

V. Conclusion

Our main goal was to find evidences of formant tuning in Cretan Rizitiko singing. The results revealed that formant tuning occurs in the modern Cretan singing voice, as in other non-operatic vocal styles [13] [20] [21]. All participating fourteen (14) Cretan singers, performed at vowel "A" the ascending/descending scale, with strong elements of formant tuning. In many cases of Cretan singers, F1 was tuned to H4, H3 and even at H2. At these harmonics, formant tuning becomes more noticeable as an increase of SPL is observed. Moreover, it has become clear from our measurements so far, that Rethymno singers use a smaller opened jaw position. The latter is reinforced by the fact, that Vowel "A" is an "open" vowel and emphatically characterizes the opening of the jaw.

REFERENCES

[1] [12] Nathalie Henrich, Mara Kiek, John Smith, Joe Wolfe. Resonance strategies used in Bulgarian women's singing style: a pilot study.
[2] H. Domitrovic. P. Boersma, C. Kovacic "Long – Term average spectra in professional folk music voices: a comparison of the KLAPA and DOZIVACKI styles." Institute of Phonetic Sciences, Proceedings 25 pp.53-64, 2003
[3] Ana P. Mendes, Aira F. Rodrigeus, and David M. Guerreiro. "Acoustic and Phonatory Characterization of the Fado Voice". Journal of Voice, Vol. 27, pp.655.e9-655e15, 2012
[4] J.J. Cabrera, J.M. Diaz-Banez, F.J. Escobar-Borrego, E. Gomez, F. Gomes, J. Mora. "Comparative Melodic Analysis of A Cappella Flamengo Cantes", Thessaloniki,Greece, Proceedings on Interdisciplinary Musicology, 2008
[5] [6] Hnaraki. M. "Cretan Music-Unravelling Ariadne's Thread", Athens, Kerkyra Publications, pp111-151, 2007
[7] D. Conklin, Chr. Anagnostopoulou "Comparative Pattern Analysis of Cretan Folk Songs" Journal of New Music Research, 40:2, 119-125, 2011
[8] [9] Ioanna Kappa "Phonological Features of the Cretan Dialect", University of Crete press, pp 53, 2014
[10] Hnaraki. M. "Cretan Music-Unravelling Ariadne's Thread", Athens, Kerkyra Publications, pp 97, 2007
[11] G. Kouroupetroglou, D. Delviniotis, G. Chrisoxoidis. "Standard Tagged Collection of Chanting Voices". Athens, Proceedings of International Musicological and Chanting Conference, 2006
[13] G. Chrisoxoidis, G. Kouroupetroglou."Formant Frequency Tuning in Professional Byzantine Chanters". Recent Advances in Electrical and Computer Engineering, 2014
[14] Sundberg Johan, "The Science of the singing Voice", University Press, Northern Illinois, 1987
[15] Joliveau. E, Smith. J, Wolfe. J "Vocal Tract Resonances in Singing: The Soprano Voice", Acoustical Society of America, 2004
[16] Summerfield. Q, Foster. J, Tyler. R "Influences of Formant Bandwidth and Auditory Frequency Selectivity on Identification of Place of Articulation in Stop Consonants", University of York, UK, 1984
[17] H. K. Shutte, D. G. Miller and J. G. Svec. "Measurements of Formant Frequencies and Bandwidths in Singing". Philadelphia, Journal of Voice, Vol. 9. No. 3. pp. 290-296, 1995
[18] G. Carlsson and J Sundberg. "Formant Frequency Tuning in Singing". New York, Journal of Voice, Vol 6. Pp. 256-260, 1992
[19] J. Sundberg, F. M. B. La and B. P. Gill." Formant Tuning Strategies in Professional Male Opera Singers". KTH Stockholm, Journal of Voice, Vol. 27. pp. 278-288, 2013
[20] H. Smith, K. Stevens, R. Tomlinson. "On an Unusual Mode of Chanting by Certain Tibetan Lamas". Journal of Acoustic Society of America, 41:1262-1264,1967
[21] J. Sundberg. "Vocal Tract Resonance in Singing ". New York: Raven Press, The NATS Journal, 1967.

LARYNGEAL AND VIBROACOUSTIC FACTORS IN ESTILL VOICE MODEL FIGURES - CASE STUDY

M. Frič, P. Amarante Andrade, A. Dobrovolná

Musical Acoustics Research Centre, Academy of Performing Arts in Prague, Czechia
marek.fric@hamu.cz, pedro.andrade@hamu.cz, alena.dobrovolna@gmail.com

Abstract: **Laryngeal dimensions and glottal and acoustic parameters were measured in a single female subject who performed 7 out of 13 voice figures and 6 qualities from Estill Voice Training (EVT). High speed videolaryngoscopy and acoustic analysis showed that the influence of voice figures and qualities on vocal fold oscillations was significant. Factor analysis identified the glottal behavior in open and closing quotients and glottal stiffness as significant factors, which mainly describe body cover control figures. The second type of interpretable factor of vocal fold oscillation was related to the speed quotient; therefore changes in the degree of adduction as well as acoustic interaction between the vocal folds and the supraglottal vocal tract could be assumed. With the exception of the second formant, which was related to the vertical position of the larynx and tongue, changes in acoustic features were mainly manifested in the total SPL. This was primarily influenced by the level of the first harmonic component and secondly by the energy in the 2 - 4 kHz bandwidth.**

Keywords: **Singing, Estill voice model, Estill voice training, glottal vibrational parametrization, acoustic parameters, Glottis Analysis Toolbox**

I. INTRODUCTION

Estill Voice Training (EVT) is a program for developing special vocal skills [1]. Experimental studies suggested that EVT is potentially an effective educational system for developing and controlling distinct voice qualities in contemporary commercial singing [2]. EVT teaches six vocal qualities that differ in the level of aryepiglottic narrowing and the occurrence of the singer's formant [3]. An emphasis on body-cover figures helps students to develop the ability to discriminate among slack, thick, thin and stiff vocal conditions. These conditions can be objectively identified using SPL, subglottal pressure, glottal airflow and perturbation measures; however, contact quotient (CQ) values from electroglottography (EGG) have not been shown to be able to distinguish among voice conditions [4]. The aim of this work is to describe how acoustic, laryngoscopic, and vibrational parameters of the voice change across 7 different figures and 6 voice qualities (i.e. conditions) within the EVT system.

II. METHODS

A single female subject (45 y.o.) with a Certificate of Figure Proficiency in EVT participated in the study. Synchronized acoustic and EGG signals (Laryngograph D200) were measured together with high-speed videolaryngoscopy (HSV, Phantom V611 VisionResearch) using a 90° rigid laryngoscope (Olympus) at 6000 fps. The subject performed 7 out of 13 Estill figures (the list is given in the caption of Fig. 1) in two pitches, C4 and A4. Prior to HSV analysis, the antero-posterior axis of the glottis was aligned in the vertical direction and normalized to the width of the epiglottis. Subsequent measurements of the antero-posterior glottal length (AP) and false vocal fold (FVF) width were measured from the middle parts of 166ms length excerpts. Subsets of the HSV files were analyzed using the Glottis Analysis Toolbox (GAT, University Hospital Erlangen, Erlangen, Germany). The acoustical parameters (frequencies and amplitudes of the first three formants, SPL @30 cm, the level of the first harmonic, the level difference between the first and the second harmonics, and the singing power ratio) were calculated from a synchronously recorded sound signal.

A factor analysis (varimax normalized, STATISTICA 6.0) was used to find the main components of variability in the glottal area waveform and acoustical parameters separately. The factor analysis was done with all utterances where the glottal area and other glottal parameters could be calculated. Slack body cover utterances in both pitches produced unreliable glottal parameters and therefore were excluded from further analysis. In addition to this, utterances where the visibility of the full vocal fold length or width (e.g. antero-posterior compression and FVF constriction) were impaired were also excluded, as they affected the glottal analysis.

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

The goal of the factor analysis was to identify if glottal and acoustical parameterization had similar factorial trends in both pitches and if they were characterized by similar singing conditions.

III. RESULTS

The first 2 graphs in Fig. 1 (left) show the ratio of the length of the glottis to the width of the gap between the false vocal folds relative to the width of the epiglottis for C4 and A4. In both pitches, the longest ***glottal length*** was found for **stiff** body cover and **low tongue** position, while **belt**, **speech** and **twang** showed the shortest glottal length. ***Wide FVF*** was measured for **opera**, **low tongue** and **thyroid tilt**, while ***narrow FVF*** were found for **belting**, **thick** body cover and **cricoid tilt**.

The correlation coefficients between laryngeal dimensions and measured glottal and acoustic parameters are depicted in Tab. 1 and Tab. 2, respectively. Most glottal parameters correlated with the length of glottis in C4, while in A4 they correlated mainly with FVF width and the ratio between glottal length and FVF width.

The results of the factor analysis of glottal parameters are depicted in Tab. 1 (Factor and Corr. columns). At both pitches, glottal parameters could be divided into 5 factors that described 82.1% (for C4) and 87.4% (for A4) of the variability.

The main components found for both pitches included: *open*, *closing*, *rate* and *amplitude quotients*, *glottal area index* and *stiffness*. In addition to this, for C4, the *Maximum-Area-Declination-Rate* was included in the main component, while *shimmer* and *plateau quotient* were part of the main factor for A4. The second components in both pitches were related to the *time periodicity*, *jitter* and *HNR*, and for C4, there was also the influence of the *plateau quotient*. The third factor in C4 and the fifth in A4 were related to the *speed* and the *asymmetry quotients*, respectively. In C4 the third factor was additionally influenced by *waveform* and *amplitude symmetry indexes*. The third factor in A4 was related to *waveform* and *phase asymmetry indices* and *contour angles symmetry [CP]*.

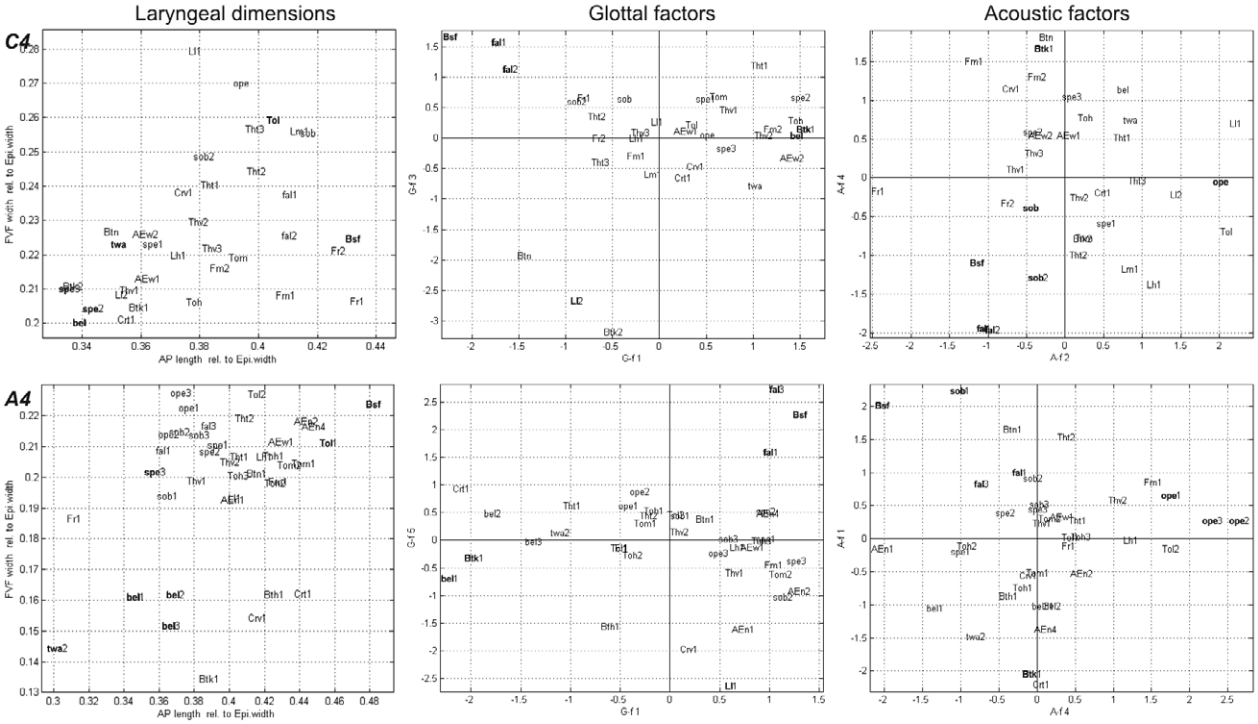


Fig. 1 Distribution of utterances according to laryngeal dimensions (left column) and most relevant glottal (middle column) and acoustic factors (right column) for C4 (upper row) (left column) and A4 (bottom row). Abbreviation of vocal figures: False voc. folds (F): constrict [Fc], mid [Fm], retracted [Fr]; Body cover contr. (B): slack [Bsl], thick [Btk], thin [Btn], stiff [Bsf]; Thyroid cart. contr. (Th): vertical [Thv], tilt [Tht]; Cricoid cart. contr. (Cr): vertical [Crv], tilt [Crt]; Ary-epiglotic sphincter (AE): wide [AEw], narrow [AEn]; Laryngeal vert. pos. (L): low [Ll], mid [Lm], high [Lh]; Tongue pos. (To): low [Tol], mid [Tom], high [Toh]) and six voice qualities: speech [spe], falsetto [fal], sob [sob], opera [ope], oral twang [twa], belt [bel]

Tab. 1 Factor loadings of Glottis analysis toolbox parameters and correlation with vocal fold dimensions for C4 and A4 pitch respectively

Pitch	***C4***					***A4***				
Parameter / Factor	***Factor***	**Corr.**	AP length	FVF width	FVF/ AP	***Factor***	**Corr**	AP length	FVF width	FVF/ AP
Open-Quotient(OQ)	**1**	-0.8	0.53*			**1**	0.93		0.61•	0.41*
Closing-Quotient(ClQ)	**1**	-0.79	0.71•			**1**	0.86		0.68•	0.5•
Amplitude-Quotient	**1**	-0.8	0.63•			**1**	0.78		0.75•	0.58•
Rate-Quotient(RQ)	**1**	0.74	-0.65•			**1**	-0.89		-0.73•	-0.52•
Stiffness	**1**	0.79	-0.45*			**1**	-0.92		-0.8•	-0.54•
Glottal-Area-Index(AC/OQ)	**1**	0.84	-0.55•			**1**	-0.95		-0.66•	-0.44*
Maximum-Area-Declination-Rate	**1**	0.81	-0.43*			**6**				
Shim(%)	**NaN**					**1**	-0.87		-0.67•	-0.58•
Time-Periodicity	**2**	0.96				**2**	0.94			
HNR(dB)	**2**	0.83				**2**	0.71		0.43*	
Jitt(%)	**2**	-0.96				**2**	-0.94			
Plateau-Quotient(PQ)	**2**	0.82	0.6•			**1**	-0.87			
Speed-Quotient(SQ)	**3**	-0.87	-0.67•			**4**	-0.92		-0.5*	-0.46*
Asymmetrie-Quotient	**3**	-0.84	-0.67•			**4**	-0.93		-0.46*	-0.44*
Waveform-Symmetry-Index	**3**	0.75				**3**	-0.86			
Amplitude-Symmetry-Index	**3**	0.86	0.59•			**NaN**			0.69•	0.49*
Phase-Asymmetry-Index	**NaN**					**3**	0.78			
ContourAngles-Symmetry*[CP]						**3**	0.71			-0.42*
Peak-Closing-Velocity	**4**	0.9		0.76•	0.45*	**NaN**			0.77•	0.52•
Peak-Acceleration	**4**	0.89		0.78•	0.44*	**NaN**			0.78•	0.53•
Amplitude-Length-Ratio	**4**	0.93		0.61•	0.57•	**NaN**			0.76•	0.58•
Amplitude-Periodicity	**5**	-0.7				**1**	0.87		0.66•	0.58•
mean-WMC	**NaN**					**5**	-0.85			

Tab. 2 Factor loadings of acoustic parameters and correlation with vocal fold dimensions for C4 and A4 pitch respectively

Pitch	***C4***					***A4***				
Par	**Fact.**	**Corr.**	AP length	FVF width	FVF/ AP	**Fact**	**Corr.**	AP length	FVF width	FVF/ AP
F1	1	0.9	-0.48*	-0.53*		5	-0.84			
F2	5	0.92		-0.65•	-0.47*	2	-0.87			
F3	NaN			-0.73•	-0.65•	2	-0.72			
A1	1	-0.82	0.51*	0.55•		1	0.84		0.67•	0.6•
A2	3	0.97				3	-0.95			
A3	4	0.94	-0.56•			NaN				
SPL	2	0.84	-0.55•		0.46*	4	0.96			0.44*
SPR	4	0.81	-0.62•	-0.55•		1	-0.94		-0.57•	-0.52•
L(f0)	2	0.96			0.67•	4	0.93		0.49*	0.54•
L(f0)-L(2f0)	1	-0.93	0.51*	0.68•		NaN			0.56•	0.63•

The fourth factor in C4 included the *peak closing velocity*, the *peak acceleration* and the *amplitude-length ratio*, while in A4, significant value was achieved with the *mean-WMC* parameter alone.

Score coefficients of interpretable factors according to glottal parametrization are depicted in Fig. 1 (middle column). In the first component in both pitches, **stiff** and **falsetto** were in opposition to **thick** and **belt**

conditions. The second factor revealed no clear relations among conditions. The third components in C4 and the fifth in A4 revealed **low larynx** to be in opposition to the **falsetto** condition.

The acoustic parameters could be divided into 5 factors for both pitches (C4 and A4, see Tab. 2), these factors describe 92.3% and 88.2% of the variability for C4 and A4 pitches, respectively, but their distribution varied according to pitch. The second factor in C4 and the fourth factor in A4 included *SPL* and *level of the first harmonic* (Lf0). The distribution of utterances along those factors (see Fig. 1, right column) clearly differentiated the **stiff** from the **opera** configurations. The fourth factor in C4 and the first in A4 included the *Singing power ratio* parameter which separated **falsetto** and **sob** qualities from the **thick** condition. Factors containing the *second formant position* (F2) (the fifth in C4 and the second in A4) were found to be related to the **vertical position** of the **tongue** and **larynx**.

IV. Discussion and conclusion

In this study, false vocal folds constriction was found to affect the type of vocal fold vibration (similar to [5]) and to produce apperiodic glottal behavior. Laryngeal dimensions showed that the length of the glottis was systematically affected by most glottal parameters in the lower pitch (C4), while in the higher pitch (A4) they were affected mainly by FVF width. This finding supports the assumption that the length-to-width ratio of the glottis depends on the vocal register [6].

Glottal and acoustic parameterization showed similar significant and interpretable factors in both pitches. The most important GAT factor consistently included parameters associated with open and closing quotients and stiffness. These can be linked to their respective laryngeal mechanisms [7] and level of adduction [8], differentiating falsetto from thick conditions. From the EVT [1] point of view, this factor describes the best body cover control figure. In the acoustical parametrization, SPR was found to be an important factor, probably due to an enhancement in the 2 - 4 kHz spectral band [9]. Stiff vs. opera conditions were found to be at the opposite extremes of the total acoustical energy and therefore, can be assumed to be related to glottal adduction and the presence of glottal insufficiency.

A factor containing speed quotient consistently differentiated falsetto (low speed quotient) from low laryngeal position (high speed quotient) in both pitches, which suggests higher levels of adduction with a lower larynx position. The difference in speed quotient found between falsetto and low laryngeal position could have also been caused by supraglottal acoustic interaction [10].

Finally, the vertical position of the tongue and larynx had a systematic effect on F2, however no systematic effects in glottal parameters were found for both pitches.

The results from this case study confirm that similar fundamental acoustic and glottal factors occur for the two different pitches tested (C4 and A4). Therefore, it can be assumed that the glottal factors open and speed quotients as well as the acoustic factors of total energy (SPL) and singing power ratio are some of the underlying elements that form the EVT figures and voice qualities. Further studies with multiple expert subjects are needed in order to confirm the preliminary findings from this study.

Acknowledgement: This publication was written at the Academy of Performing Arts in Prague as part of the project “Subjective and objective aspects of musical sound quality.”

REFERENCES

[1] Steinhauer KM, McDonald Klimek M, Estill J. *The Estill Voice Model: Theory & Translation*. 2nd ed. Estill Voice International; 2017.

[2] Fantini M, Fussi F, Crosetti E, Succo G. Estill Voice Training and voice quality control in contemporary commercial singing: an exploratory study. *Logop Phoniatr Vocology*. 2017;42(4):146-152.

[3] Yanagisawa E, Estill J, Kmucha ST, Leder SB. The Contribution of Aryepiglottic Constriction to "Ringing" Voice Quality A Videolaryngoscopic Study with Acoustic Analysis. *J Voice*. 1989;3(4):342-350.

[4] Barone NA, Ludlow CL, Tellis CM. Acoustic and Aerodynamic Comparisons of Voice Qualities Produced After Voice Training. *J Voice*. 2021;35(1):P157.E11-157.E21.

[5]Madill, C., Nguyen, D. D. Impact of Instructed Laryngeal Manipulation on Acoustic Measures of Voice–Preliminary Results. *J Voice* (In Press).

[6] Larsson, H., Hertegård, S. Vocal Fold Dimensions in Professional Opera Singers as Measured by Means of Laser Triangulation. *J Voice*, 2008;22(6), 734–739

[7] Henrich N. Mirroring the voice from Garcia to the present day: some insights into singing voice registers. *Logoped Phoniatr Vocol*. 2006;31(1):3-14.

[8] Herbst CT, Qiu Q, Schutte HK, Švec JG. Membranous and cartilaginous vocal fold adduction in singing. *J Acoust Soc Am*. 2011;129(4):2253-2262.

[9] Omori K, Kacker A, Carroll LM, Riley WD, Blaugrund SM. Singing power ratio: Quantitative evaluation of singing voice quality. *J Voice*. 1996;10(3 2):228-235.

[10] Rothenberg M. Acoustic interaction between the glottal source and the vocal tract. In: Stevens KN, Hirano M, eds. *Vocal Fold Physiology.* Tokio, Japan: University of Tokyo Press; 1981:305–328.

PRESSURE, FLOW AND GLOTTAL AREA WAVEFORM PROFILE CHANGES DURING PHONATION USING THE ACAPELLA CHOICE® DEVICE

P. Andrade[1], M. Frič[1], B. Saccente-Kennedy[2], V. Hruška[1]

[1]Musical Acoustics Research Centre, Academy of Performing Arts in Prague, Praha, Czechia
[2]University College London Hospitals NHS Foundation Trust, London UK
pedro.andrade@hamu.cz, marek.fric@hamu.cz, brian.saccente-kennedy@nhs.net, hruska.viktor@hamu.cz

***Abstract*: Vibratory positive expiratory pressure devices (PEP) are now considered a suitable resource for voice therapy. PEP devices produce large low frequency intraoral pressure modulations in the vocal tract that influences glottal behaviour. In this study, the impact of phonation into an Acapella Choice device (a type of PEP) on glottal behaviour was assessed. Phonation was produced by 2 males and 1 female participant whilst audio, EGG, pressure, flow and high-speed videoendoscopic data were collected. The results showed a systematic effect on glottal behaviour with changes in pressure caused by the Acapella device. When Acapella pressure was maximum, vocal fold vibration was hindered (lower: EGG amplitude, airflow, contact quotient (CQ), fundamental frequency (fo) and glottal area (GA)) as Acapella pressure reduced the opposite trend was observed. This systematic change in the supraglottic pressure modulates the behaviour of the vocal folds between what seems to be hindered and aided vibration. This behaviour confirms a mechanistic impact of the Acapella device on the phonatory apparatus that can be used for specific voice therapy purposes.**

I. INTRODUCTION

Voice therapy using semi-occluded vocal tract exercises (SOVTE) has received increasing adoption worldwide. More recently, the use of vibratory positive expiratory pressure (PEP) devices has been incorporated into the array of tools used as SOVTEs [1-4]. PEP devices are composed of a mouthpiece connected to a tube with an oscillatory valve at its distal end. Some devices such as the Flutter or Shaker use a plastic cone containing a metal sphere that is displaced by the airflow whilst the Acapella Choice device (henceforth Acapella) is composed of a tube with a distal oscillatory arm that closes and opens with airflow (Fig. 1). Although originally designed to mobilise secretion from the lungs in conditions such as cystic fibrosis and neurogenic diseases, the shaking mechanism of PEP devices can be used to produce a massage-like effect in the laryngeal muscles, consequently counteracting harmful effects of tension in the phonatory apparatus. PEP devices share some of their properties with tube phonation (another category of SOVTE), as both cause an artificial lengthening of the vocal tract increasing its impedance in relation to that of the vocal folds whilst allowing continuous airflow [5]. In specific, they increase positive (inertive) reactance close to the fundamental frequency (fo), consequently increasing intraoral acoustic pressure leading to greater vocal economy [6].

In addition to changing the overall configuration of the vocal tract, PEP devices also affect the pressure-flow profiles in the vocal tract by the addition of a second vibratory valve at its distal end (e.g., rocker arm in the Acapella - Fig. 1). Changes in pressure and flow are modulated by the amplitude and frequency of the rocker arm opening and shutting mechanism. In this regard, PEP devices also behave like water resistance therapy (WRT) as both techniques involve the modulation of pressure and flow in the vocal tract by changes in the configuration of the distal end of the tube (rocker arm and water bubbling for Acapella and WRT respectively). In the case of WRT, pressure and flow modulation is controlled by the frequency, size, shape, and vibration regime of the water bubbles [7]. For tubes, as well as for PEP devices, the pressure values are largely determined by flow rate [3,8], this contrasts with WRT where pressure values are predominantly determined by the height of the water column above the submerged distal end of the tube [8].

In addition to the changes in the pressure-flow profiles caused by the PEP device, the vibration pattern of the vocal folds can also be affected using SOVTEs. In specific, previous studies have shown changes in contact quotient (CQ) values during SOVTEs. Most studies found CQ to increase with WRT [9,10,11] with fewer studies showing no clear trends in CQ [12,13]. Furthermore, it has been shown that exercises with two sources of vibration tend to produce larger ranges of CQ values than exercises with a single source of vibration in the vocal tract [14].

Even though PEP devices share common characteristics with tube phonation and WRT, their impact on the vibratory pattern of the vocal folds during exercises has not been described. PEP devices, specifically the Acapella, has been shown to produce large and systematic changes in pressure values [3] and therefore can be expected to produce observable changes in the vocal fold vibratory pattern. The aims of this study are to describe

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

the influence of the Acapella on the glottis and its effect on the vibration pattern of the vocal folds.

II. METHODS

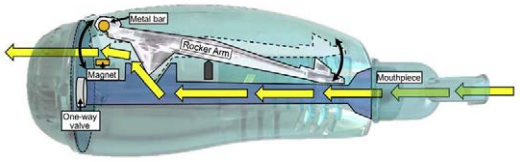


Fig 1. Acapella Choice (Taken from Saccente et al, 2020[3])

Three subjects with no known laryngeal pathologies participated in the study, two males and one female. The Acapella Choice (Acapella Choice, Smiths Medical ASD, Inc, Rockland, Massachusetts) was used in this study as it was shown to produce large mechanistic changes in intraoral pressure [3]. In addition to this, due to the Acapella's mechanism, its pressure-flow values are less likely to be affected by changes in the angle between the device and the floor which can occur with other PEP devices [4].

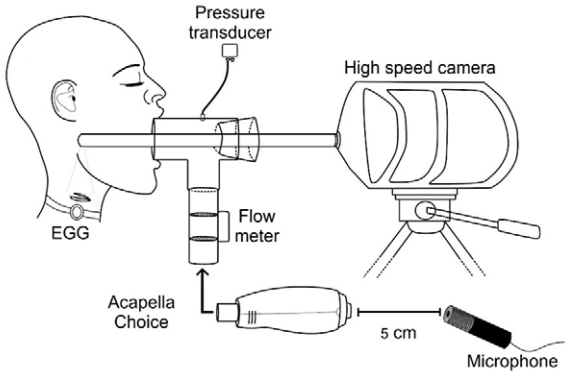


Fig 2. Data collection setup. For clarity, the T-shaped mouthpiece is shown vertically, however during the experiment, the flow meter and tested devices were kept horizontally.

The high-speed videoendoscopy (HSV) was recorded at 6000 fps using a VisionResearch Phantom V611 (VisionResearch Phantom, New Jersey, USA) with an Olympus CLV-S45 (300W) light source (Olympus Corporation, Tokyo, Japan). Half a second segments were recorded for each utterance at a stable part of phonation. Pressure was measured using a digital manometer Honeywell ASDXAVX001PD2A3 (Honeywell International Inc, North Carolina, USA). A 10 cm long measuring probe was placed at the T-shaped mouthpiece junction (Fig. 2). Airflow was measured using a Sensirion SFM3000 digital flow meter (Sensirion AG, Staga, Switzerland) placed between the mouthpiece and the Acapella. Pressure and airflow measurements were recorded at 2 kHz and resample to 48 kHz using Octave ([GNU Octave] version 6.1.0, www.gnu.org/software/octave/index). The rigid endoscope was placed across the straight portion of the T-shaped mouthpiece and sealed at the distal end to avoid air leakage (fig 2). The pressure probe was placed through a small hole at the outside part of the perpendicular joint in the T-shaped tube. Vaseline was applied to all joints to avoid air leakage. Glottal area waveform (GAW) signals were obtained using the Glottal Analysis Tools 2020 software (GAT) (University Hospital Erlangen, Erlangen, Germany). The GAW was resampled to 48kHz using Octave. Electroglottography (EGG) signal was also recorded using a Laryngograph A-100 device (Laryngograph, Wallington, UK). Acoustic data was obtained using a Sennheiser ME 62 microphone (Sennheiser, Wedemark, Germany) placed at 5 centimetres from the distal end of the Acapella. EGG and Audio signals were recorded synchronously (48 kHz, 24-bit). The audio signal was also used for annotation during the experiment. Data was collected in a sound treated room.

Each subject was asked to align the centre of the endoscopic viewing field with the larynx to provide a view of the entire extent of the glottis. The /i/ vowel was used as a target, however distortions in its sound qualities were allowed due to the presence of the endoscope. The subjects were asked to sustain phonation at E3 for males and E4 for the female at habitual loudness for at least 4 seconds.

III. RESULTS AND DISCUSSION

Fig. 3 and 4 shows synchronous EGG, VKG, glottal area, pressure, and airflow data for the Acapella for one male subject (M1) and the female (F1) subject respectively. Three components of pressure and flow data are shown in the graphs and related to a) glottal = the high frequency modulation by the glottal cycle, b) Acapella = low frequency modulation by the Acapella, and c) DC = static elements used to pressurise the Acapella and vocal tract prior to oscillation. As the data for all three subjects showed similar patterns, M2 date is not available in this study.

From the data, it can be observed that when the pressure produced by the Acapella ($pressure_{Acapella}$) increases, CQ and amplitude of the EGG signal ($EGG_{envelope}$) decrease. These reduced values for CQ and $EGG_{envelope}$ are likely caused by a larger supraglottic pressure which increases the intraglottic pressure (assuming that the subglottic pressure remains somewhat constant) consequently hindering the contact between the vocal folds. In addition to this, the peak-to-peak amplitude of pressure modulation by the glottal cycle ($pressure_{glottal}$) also reduces. As the $pressure_{Acapella}$ drops towards minimum values, the opposite trend is observed for CQ, EGG amplitude and peak-to-peak amplitude of the $pressure_{glottal}$. This suggests that the ability of the vocal folds to modulate pressure values is affected by changes in the $pressure_{Acapella}$. When the $pressure_{Acapella}$ is maximum, lower extreme values in $pressure_{glottal}$ modulation are seen, when $pressure_{Acapella}$ drops the peak-to-peak amplitude of the $pressure_{glottal}$ increases.

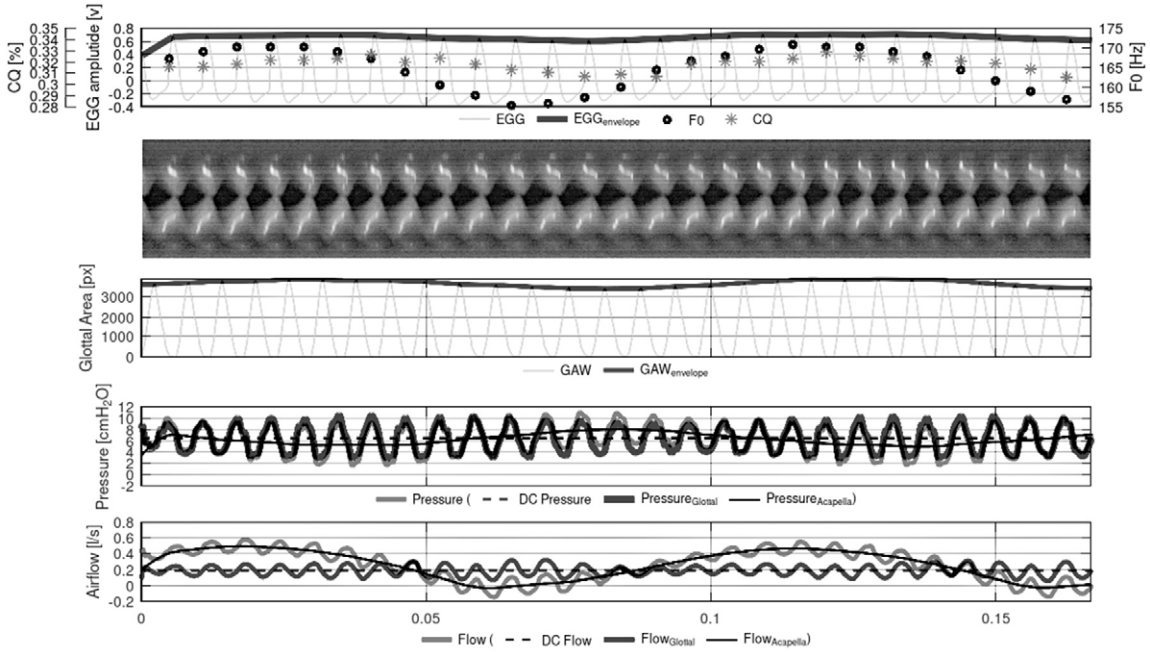


Fig 3. Data for a male participant (M1) producing an E3 at a habitual level. From top to bottom are EGG and CQ, VKG, glottal area, pressure, and airflow. All data presented is shown for 0.15 seconds duration.

The impact of $pressure_{Acapella}$ oscillation on the vocal fold vibratory behaviour is also seen in the videokymogram (VKG) (clearer for M1) and glottal area signals. During high $pressure_{Acapella}$ values the contact between vocal folds (clearer seen in the CQ values) and maximum GA (peak values) are reduced. The opposite trend is found when $pressure_{Acapella}$ reduces towards minimum. Limitations in achieving the same glottal opening (max GA values) during maximum and minimum $pressure_{Acapella}$ conditions, shows an inability of the vocal folds to achieve maximum lateral displacement when $pressure_{Acapella}$ is large. It is worth noting that even when $pressure_{Acapella}$ reaches minimum values, the vocal tract is still pressurized by the static pressure (DC element). Unsurprisingly, when $pressure_{Acapella}$ increases, the fo of the glottal cycle also reduces because of the stronger opposition to vocal fold vibration. Possibly this is likely caused by the increased intraglottal pressure. However, the lower fo values may also be caused by the positive pressure above the level of the glottis which slows down the upstream flow from the lungs. This effect can be expected at instances when $pressure_{Acapella}$ is increasing and the Acapella flow signal ($flow_{Acapella}$) is decreasing causing the flow modulation by the glottis ($flow_{Glottal}$) to reach close to zero values.

Although consistent across all 3 subjects, the described glottal behaviour associated with changes in $pressure_{Acapella}$ seem to be affected by pitch and/or gender as data from male subjects showed clearer trends (as described in this study) than for the female subject.

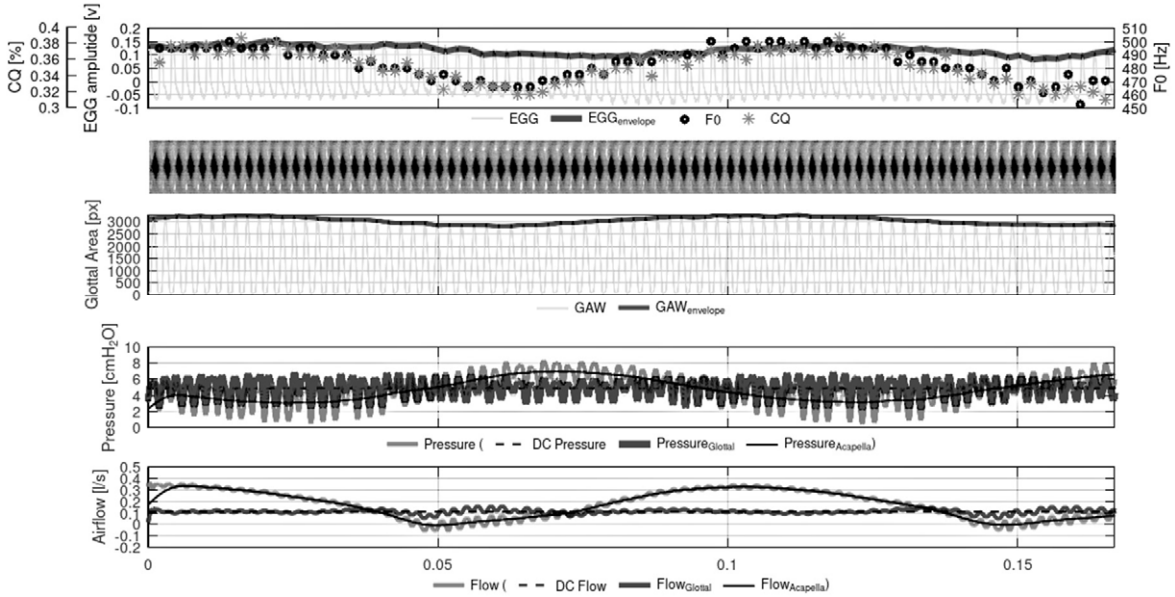


Fig 4. Data for a female participant (F1) producing an E4 at a habitual level. From top to bottom are EGG and CQ, VKG, glottal area, pressure, and airflow. All data presented is shown for 0.15 seconds duration.

IV. CONCLUSION

In this study, the impact of Acapella on glottal behaviour was investigated. As $pressure_{Acapella}$ increases, the vibration of the vocal fold's decreases, showing lower fo, less contact and amplitude of vibration of the vocal folds. When $pressure_{Acapella}$ reduces, the opposite trend is observed where vocal fold vibration seems to be more effective in modulating $pressure_{glottal}$ and $flow_{glottal}$ values. This systematic change in $pressure_{Acapella}$ seems to alternate the behaviour of the vocal folds between hindered and unhindered vibration. This information can be used to clarify the understanding of vocal fold vibration patterns under changeable loading conditions during phonation into PEP devices. It also confirms a mechanistic impact of the Acapella device on the phonatory apparatus that can be used when considering specific voice therapy outcomes.

Acknowledgement

This study was written at the Academy of Performing Arts in Prague as part of the project "Application of Semi-occluded vocal tract and neuromuscular electrical stimulation for professional voice user" with the support of the Institutional Endowment for the Long-Term Conceptual Development of Research Institutes, as provided by the Ministry of Education, Youth and Sports of the Czech Republic.

REFERENCES

[1] Saters TL et al. The Voiced Oral High-frequency Oscillation Technique's Immediate Effect on Individuals With Dysphonic and Normal Voices. *J Voice.* 2018;32(4):449-458

[2] da Silva Antonetti AE, et al. Voiced High-frequency Oscillation and LaxVox: Analysis of Their Immediate Effects in Subjects With Healthy Voice. *J Voice*. Published online May 31, 2018.

[3] Saccente-Kennedy B, et al. A Pilot Study Assessing the Therapeutic Potential of a Vibratory Positive Expiratory Pressure Device (Acapella Choice) in the Treatment of Voice Disorders. *J Voice*. 2020;34(3): 487.e21-487.e30.

[4] Laukkanen AM, et al. Buzzer versus water resistance phonation used in voice therapy. Results obtained with physical modeling. *Biomed Signal Process Control*. 2021; 66:102417

[5] Story BH. Acoustic impedance of an artificially lengthened and constricted vocal tract. *J voice*. 2000;14(4):455-469

[6] Titze IR, Laukkanen AM. Can vocal economy in phonation be increased with an artificially lengthened vocal tract? A computer modeling study. *Logop Phoniatr Vocology*. 2007;32(4):147-156.

[7] Wistbacka G, et al. Resonance Tube Phonation in Water—the Effect of Tube Diameter and Water Depth on Back Pressure and Bubble Characteristics at Different Airflows. *J Voice*. 2018;32(1): 126.e11-126.e22.

[8] Amarante Andrade P, et al. The Flow and Pressure Relationships in Different Tubes Commonly Used for Semi-occluded Vocal Tract Exercises. J Voice. 2016;30(1):36-41.

[9] Guzman MA, et al. Do Different Semi-Occluded Voice Exercises Affect Vocal Fold Adduction Differently in Subjects Diagnosed with Hyperfunctional Dysphonia? Folia Phoniatr Logop. 2015;67(2):68-75.

[10] Laukkanen A-M, et al. High-speed registration of phonation-related glottal area variation during artificial lengthening of the vocal tract. Logop Phonatrics Vocology. 2007;32(4):157-164.

[11] 1. Tyrmi J, Laukkanen AM. How Stressful Is "Deep Bubbling"? J Voice. 2016;31(2): 262.e1-262.e6.

[12] Gaskill CS. The Effect of an Artificially Lengthened Vocal Tract on Estimated Glottal Contact Quotient in Untrained Male Voices. J Voice. 2008;24(1):57-71.

[13] Quinney DM, Gaskill CS. The effect of resonance tubes on glottal contact quotient with and without task instruction: a comparison of trained and untrained voices. J Voice. 2012;26(3):79-93.

[14] Amarante Andrade P, et al. 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. 2014;28(5):1-7.

SESSION VI
NEWBORNS AND CHILDREN

EVALUATING THE ACCURACY OF DECODING IN CHILDREN WHO READ ALOUD

E. Bruno[1], S. Giulivi[2], C. Cappa[3], M. Marini[4], M. Ferro[1]

[1]Institute for Computational Linguistics ILC-CNR Pisa, Italy

[2]University of Applied Sciences and Arts of Southern Switzerland SUPSI Locarno, Switzerland

[3]Institute for Clinical Physiology IFC-CNR Pisa, Italy

[4]Department of Information Engineering, University of Pisa, Pisa, Italy

ester.bruno@ilc.cnr.it, sara.giulivi@supsi.ch, cludia.cappa@cnr.it,
marco.marini@phd.unipi.it, marcello.ferro@ilc.cnr.it

Abstract: **Digital tools based on automatic speech recognition (ASR) could be a useful support for teachers in assessing the reading skills of the students. We focus on the evaluation of the decoding accuracy of children with grade level ranging from the 3rd to the 6th performing a reading aloud task on a narrative text displayed on an ordinary tablet using the ReadLet platform. On the basis of previously collected data, we built a gold dataset with sentences characterised by the audio data, the original text to be read, and the text actually spoken by the child. By using the open-source Kaldi toolkit an ASR system based on the GMM-HMM model was trained on the training portion of the gold dataset. The accuracy of the ASR system was calculated as the ability to correctly decode the test audio data with respect to the annotated text, and the decoding accuracy of the children was estimated by measuring the gap between the results obtained with the annotated text and the original text. A consistent trend with increasing grade level was found in terms of word correctness, substitutions and insertions, while the trained model appears to be significantly able to evaluate the children decoding accuracy.**

Keywords: **speech recognition, decoding accuracy, reading aloud, voice parameters, Kaldi, GMM-HMM acoustic model**

I. Introduction

Reading and understanding a written text are among the most relevant skills in everyone's life [1]. Whether it is to study, to read for personal pleasure, to obtain information, to use instructions, to find communications or updates, we are faced with the need to access the content of a written text. The results of the OECD-PISA 2018 international survey is the most recent in which reading skills were the main area of investigation, and return an uncomfortable international picture, from which Italy does not differ [2]. The assessment of reading skills is achievable by the educational institutions, and the combination of NLP and ICT technology can substantially help the teachers in this task [3].

The process of decoding and understanding during reading were considered by the American Psychiatric Association 2013 as two independent processes, however able to influence each other [4]. The assessment of such processes in ecological conditions on primary school children is the objective of the AEREST protocol [5], which is implemented into the ReadLet platform [6] so that, by using an ordinary tablet, the reading efficiency is automatically evaluated as the integration of the ability to decode and understand a text.

II. Materials and methods

The AEREST protocol provides for the administration of narrative-descriptive texts in three decoding modalities: silent reading, reading aloud, and listening. The decoding step is followed by a questionnaire to evaluate the comprehension of the text just read. By using an ordinary tablet, ReadLet takes care of recording the speech produced by the child, keeps track of child's finger movement on the screen and, finally, stores the answers given to the comprehension questionnaire. All acquired data are aligned over time. Three contributions are calculated to evaluate the reading efficiency of the child:

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

i) the decoding speed, ii) the correctness of the reading and iii) the understanding of the text. Points i) and iii) are already fully automated within the ReadLet platform and in this article we focus on point ii), with the aim of creating a tool that is able to automatically draw the decoding accuracy in terms of correct words, deletions, substitutions and self-corrections.

As part of the AEREST project in 2019, we created a gold dataset starting from the data acquired using from 419 children with a grade level between the third and the sixth. The overall database includes 419 reading-aloud trials and a total of 13118 sentences. To create the gold dataset, a first step involved the selection of the trials in which the child marked the text with the finger for at least 70% of the text length. Since the speech and the finger tracking data were simultaneously recorded during the trial and subsequently aligned over time, we relied on the finger tracking data to automatically split the audio data into sentences. The audio segmentation was then refined manually by means of an ad-hoc audio editing tool and, additionally, the annotation was augmented by taking into account the text actually spoken by the child compared to the original sentence.

From ReadLet we obtained a gold dataset composed by 873 sentences characterized as i) the audio data (i.e. the speech of the child), ii) the original sentence (i.e. the text that should have been pronounced by the child), and iii) the annotated sentence (i.e. the transcription of the actual speech of the child).

The ReadLet dataset was integrated with the CLIPS dataset, 16120 recordings about 8 hours and 30 minute from 250 adult subjects [7]. Once the total dataset was obtained, training and testing of an ASR system based on the GMM-HMM model [8] was performed using the open-source Kaldi toolkit [9]. The GMM-HMM model is composed by 15019 gaussians and it has been trained with the Speaker Adaptive Training (SAT) algorithm [10]. The feature vector was projected by Linear Discriminant Analysis criterion and transformed by Maximum Likelihood Linear Transformation [11] (LDA + MLLT + SAT). The final vector consisted of 40 features. MFCC features were extracted from the audio data and the decoding was performed on the fully expanded decoding graph (HCLG) that represents the language-model, pronunciation dictionary (lexicon), context-dependency, and HMM structure. Both mono-phone and tri-phone model were run and, since the latter outperformed the mono-phone model, we will focus on the tri-phone model only.

Finally the training set was obtained by all CLIPS recordings plus the 60% of the gold dataset, while the test set was built with the remaining 40% of the gold dataset. The random selection of the training and testing datasets was repeated 5 times and the results were averaged accordingly.

We trained the ASR system by feeding the model with the audio data and the annotated sentences belonging to the training dataset. During testing, we fed the model with the testing audio data and we compared the ASR transcriptions with two kind of references: i) the annotated sentences and ii) the original sentences.

III. Results

The predictions of the model run on the test audio data were compared to the target text. The accuracy of the ASR was first measured by Word Error Rate (WER) which is computed as the overall number of predicted words not matching the target text, divided by the number of total words. The preliminary results of the model show a mean WER equal to 10.95% (std=2.00%). Going more in deep, for each grade level the accuracy was evaluated as i) the average number of words per sentence correctly recognised by the model (correctness), ii) the average number of words per sentence substituted into the target text (substitutions), iii) the average number of words per sentence removed from the target text (deletions), and iv) the average number of words per sentence added to the target text (insertions). By using the annotated sentences as the target text we obtained the results shown in Fig. 1.

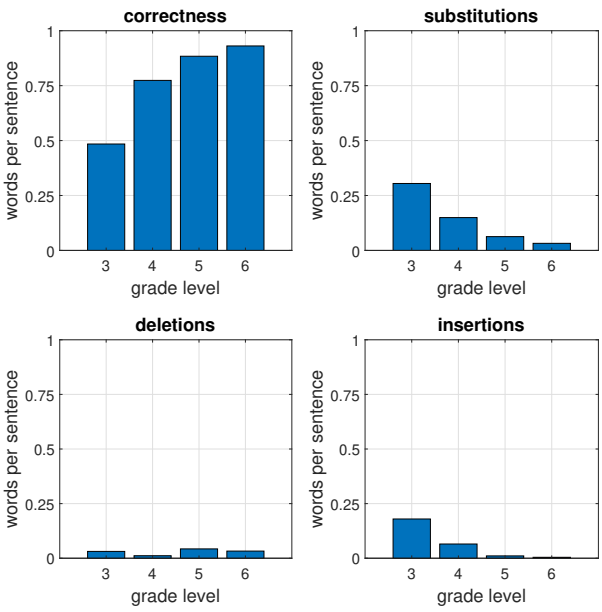


Figure 1: Accuracy of ASR system fed with the test audio data and using the annotated sentences as the target text. For each grade level, the average number of correct/substituted/deleted/inserted words per sentence is shown.

The model accuracy was also calculated on the same test audio data using the original sentences as the target text. The difference of the correctness obtained using the two target texts (i.e. the correctness on the annotated text minus the correctness on the original text) is shown in Fig. 2. While the correctness of the model on the annotated text should tell us about the accuracy of the ASR system itself, the difference of such correctness with the one obtained on the original sentences should tell us about the performance of the children.

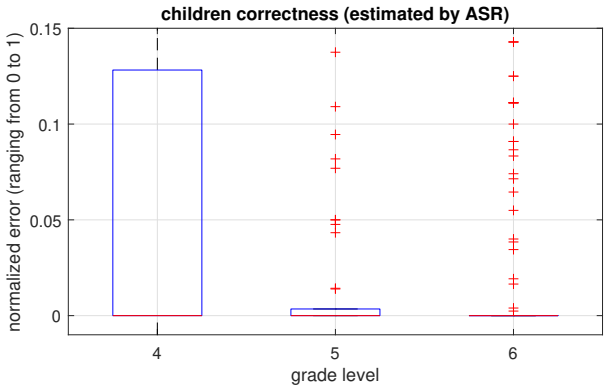


Figure 2: Children decoding accuracy estimated by the ASR system, expressed as the average number of misspelled words per sentence and calculated as the difference between the ASR correctness on the annotated sentences (Fig. 1 top-left) and the ASR correctness on the original sentences.

Finally, we evaluated the normalised edit distance between the annotated and the original sentences to obtain the reference correctness baseline for comparing the correctness estimated by the ASR system (see Fig. 3).

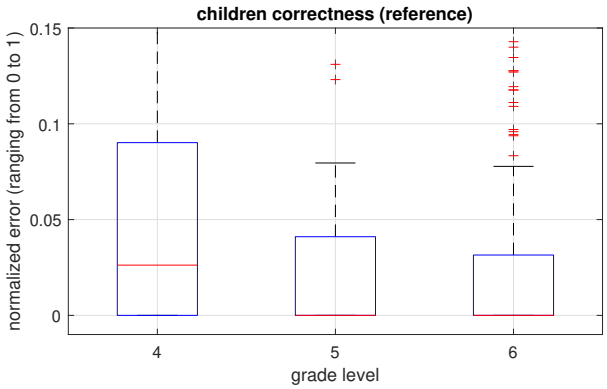


Figure 3: Children decoding accuracy used as reference, calculated as the normalised edit distance between the annotated sentences and the original sentences in the test set.

To validate the results provided by the ASR system about the performance of the children in correctly decoding the text during the reading aloud task, we calculated the Spearman rank correlation between the data shown in Fig. 2 and in Fig. 3. For each grade level, the correlation value along with the statistical significance is shown in Table 1.

grade level	r	p-value
4	0.63	$<10^{-3}$
5	0.50	$<10^{-5}$
6	0.66	$<10^{-10}$

Table 1: Spearman rank correlation between the children accuracy in decoding estimated by the ASR shown in Fig. 2 and the decoding accuracy calculated on the basis on the manually annotated sentences shown in Fig. 3. For each grade level the correlation value is shown together with its statistical significance.

IV. Discussion

As it can noticed in Fig. 1, the accuracy of the ASR model in terms of correctness is around 50% on 3rd graders, while the accuracy grows to 90% on 6th graders. The trend of substitution and insertion statistics goes in the same direction, showing that the more the reader is skilled, the more the model is able to predict the annotated text which, by definition, should reflect the audio data. Anyway a number of factors (e.g. the limited dataset, the poor annotation, the noisy audio, the poor fluency of the reader among all) may prevent the model to gain the 100% accuracy. For grade levels where the accuracy of the model is above 75% (i.e.. grade level ranging from 4 to 6) we show in Fig. 2 the evaluation of the accuracy of the children by by measuring the gap between the correctness obtained on the annotated sentences (i.e. the upper limit the ASR system can reach) and the correctness on the original sentences. Such gap, which decreases along with the grade level, appears to be highly and significantly correlated (see Table 1) with the reference error shown in Fig. 3, being the latter calculated independently on the basis of the edit distance between the annotated sentences and the original sentences.

V. Conclusion

The preliminary ASR system seems to be able to estimate the decoding accuracy of the children and to approximate the reference accuracy calculated on the gold dataset (see Fig. 2 and 3). Nonetheless, the accuracy of the ASR system itself is still poor, especially for young readers (see correctness on 3rd graders in top-left pane of Fig. 1). The improvement of the quality of the sentence annotation together with the creation of a larger gold dataset will help to fill such gap.

Moreover, the next objective consists in estimating, precisely for the words to which the model associates a high level of uncertainty, the sequence of phonemes actually pronounced by the child. This will allow for the automation of the procedure for evaluating the correctness of the decoding of the reading aloud trials. This procedure, for each of the 419 reading trials, was performed manually and these data will constitute a useful benchmark for the automatic analysis system.

A detailed analysis of decoding errors, with particular attention to those words to which the model associates a high level of uncertainty, will be integrated into the ReadLet platform to support professionals to assess the level of reading skills reached by the child, and decide which intervention programmes and measures are most appropriate.

Acknowledgments

This work was supported by the Swiss grant "AEREST: An Ecological Reading Efficiency Screening Tool" (2017-2020) funded by the Department of Teaching and Learning of the University of Applied Sciences and Arts of Southern Switzerland (SUPSI), by the Italian project "(Bio-)computational models of language usage" (2018-) funded by the Italian National Research Council (DUS.AD016.075.004, ILC-CNR), and by the PRIN grant 2017W8HFRX "ReadLet: reading to understand. An ICT-driven, large-scale investigation of early grade children's reading strategies" (2020-2022), from the Italian Ministry of University and Research.

References

[1] Stephen K. Reed. *Cognition. Theories and Applications*. Wadsworth Cengage Learning, Belmont, California (USA), 2012.

[2] OECD. Assessment and Analytical Framework. Technical report, Organisation for Economic Co-operation and Development (OECD), Paris (France), 2019.

[3] Jorge Proença, Carla Lopes, Michael Tjalve, Andreas Stolcke, Sara Candeias, and Fernando Perdigão. Automatic Evaluation of Reading Aloud Performance in Children. *Speech Communication*, 94, 2017.

[4] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5)*. American Psychiatric Association, Washington DC (USA), 2013.

[5] Marcello Ferro, Sara Giulivi, and Claudia Cappa. The AEREST Reading Database. In Johanna Monti, Felice Dell'Orletta, and Fabio Tamburini, editors, *7th Italian Conference on Computational Linguistics (CLIC-IT'20)*, Torino (Italy), 2020. aAccademia University Press.

[6] Marcello Ferro, Claudia Cappa, Sara Giulivi, Claudia Marzi, Ouaphae Nahli, Franco Alberto Cardillo, and Vito Pirrelli. ReadLet: Reading for Understanding. In *IEEE 5th International Congress on Information Science and Technology (CiSt'18)*, pages 1–6, 2018.

[7] Federico Albano Leoni, Francesco Cutugno, Renata Savy, and Valentina Caniparoli. Corpora e lessici di italiano parlato e scritto (CLIPS). `http://www.clips.unina.it/`, 2004. Last accessed on September 9th, 2021.

[8] Dan Su, Xihong Wu, and Lei Xu. Gmm-hmm acoustic model training by a two level procedure with gaussian components determined by automatic model selection. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4890–4893, 2010.

[9] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December 2011. IEEE Catalog No.: CFP11SRW-USB.

[10] Tasos Anastasakos, John McDonough, Richard Schwartz, and John Makhoul. A compact model for speaker-adaptive training. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 2, pages 1137–1140. IEEE, 1996.

[11] R. Gopinath. Maximum likelihood modeling with gaussian distributions for classification. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, 2:661–664 vol.2, 1998.

QUALITATIVE CHARACTERIZATION AND ANALYSIS OF CRYING WAVES FROM BABIES OF FOUR DIFFERENT ETHNIC GROUPS IN THE SIERRA OF THE STATE OF GUERRERO IN MEXICO

C.A. Reyes-Garcia[1], I. Gallardo-Bernal[2]
[1] FullTime Researcher INAOE, Puebla, Mexico
[2] FullTime Researcher UAGRO, Chilpancingo, Gro., Mexico
kargaxxi@inaoep.mx, igallardo@uagro.mx

Abstract: **At present in Mexico there are still living 68 indigenous peoples, each one speaking their own native language, which are organized into 11 linguistic families and are derived in 364 dialect variants, it is estimated that there are approximately 12 million people still speaking these languages in the Mexican territory[1]. The indigenous population of the Guerrero State is mainly made up of four ethnic groups, Amuzgo (ñomndaa), Mixtec (na savi), Tlapaneco (Me'phaa) and Nahuatl (Mexican from Guerrero). It is of our interest to analyze and study for the first time the crying of babies from some of the original ethnic peoples of the Sierra of the State of Guerrero in order to qualitatively characterize them. Among the qualitative characteristics [3] to be extracted are the melody type, shifts, glides, noise concentration, fundamental frequency, intensity, etc. For this research, recordings were directly made by doctors and nurses in babies less than 6 months of age. The infant crying signal recordings were then processed and analyzed to extract the relevant information that allows us later to identify any potential evidence of neurological diseases in newborns, through the use of relevant features extraction and selection techniques, pattern recognition and classification [4].**

In this article we describe, besides some basic information about the ethnic groups, how the samples were collected, the databases used, the qualitative feature extraction process, the knowledge based inference system used, some obtained results as well as a brief analysis of them.

Keywords: **Cry Analysis, Pattern Recognition, Classification.**

I. INTRODUCTION

The state of Guerrero in Mexico has 7 regions; Acapulco, Centro, Norte, Costa Chica, Costa Grande, Tierra Caliente and La Montaña, where a high percentage of indigenous population is concentrated. The four analyzed languages are spoken in different municipalties where the concentration is as follows: The highest percentage of *Nahuatl* language speakers is in Ahuacuotzingo, Cualác, Chilapa, Olinalá and Zitlala. Chilapa, Ahuacuotzingo and Tlapa where *Nahuatl* speakers are about 75% of the population. Speakers of *Tlapaneco* and *Mixteco* languages predominate in more than 40 percent of the population of the municipalities of Alpoyeca, Atlixtac, Copanatoyac and Xalpatláhuac. On the other hand, most of the Amuzgos are located in the municipality of Xochistlahuaca, also having a presence in Tlacoachistlahuaca and Ometepec, as well as other towns on the Costa Chica of the State.

Mixtec is characterized by a strong nasal tendency, which accounts for the large number of nasal and prenasalized phonemes in its phonological repertoire [6], it is heard fast, loud and clear. While Tlapaneco is a highly accentuated language. the variation between the tones is important for its translation, it is a language that is heard in principle like the Chinese language. Amuzgo consists of 14 consonant segments of the basic forms, divided into 11 consonants, one semi-vowel and two slips. The Amuzgo, Mixtec, and Tlapaneco languages are tonal languages that belong to the Ottomangue group. As for the vulnerability and fragility sides of these peoples, according to data from the National Council for the Evaluation of Social Development policy (CONEVAL) in Mexico:

- About 66% of the population suffers from food poverty
- 72% do not have the resources to access health and education services
- 40% of people over 15 years of age are illiterate and 85% did not complete basic education
- 85% do not have their own equity
- Two of the 10 municipalities with the highest extreme poverty in the country are located there, Cochoapa el Grande, with the first place and Metlatónoc, in the tenth.

All these disadvantageous living conditions, along with a congenital endogenous malnutrition, are directly reflected in newly born babies´ health. Theoretically, this weak health influences the acoustic characteristics of the infant cry signal. Our qualitative infant cry signal analysis is directed to help identify pathological trends in newly born babies from the studied ethnical groups as early as possible, through the use of easy to use affordable high technology tools.

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

With this idea as a target, we consider it important to provide technological tools to medical specialists who provide care to this special population in an ancestral state of vulnerability. These tools are directed to support their clinical diagnosis under language barriers with their patients. In particular, our proposed tool seeks to identify pathologies by performing qualitative analysis over the infant crying wave which essentially consists in the observation of the fundamental frequency (F_0) changes as a function of time. The qualitative features to identify represent the shapes that F_0 takes and are called melodic forms. The melodic form can be usually ascending, descending, ascending-descending, descending-ascending, flat and without melodic form, shifts and glides.

By means of the fundamental frequency values extracted from the crying units, the presence of the qualitative characteristics: shift, glide and noise concentrations could be determined automatically. Shifts are defined as an increase or decrease in the fundamental frequency of at least 100 Hz and less than 600 Hz, with a minimum duration of 0.1 sec [10], there may be more than one in the same crying unit, on the other hand Glides, They are defined by an increase or decrease of at least 600Hz, with a minimum duration of 0.1 sec [10] and in the same way there can be varying units in the same crying unit.

In general, pathological crying is associated with the following characteristics:

• Extreme or unstable values in the Pitch (F_0)

• Poor cry quality

• Melody type usually is descending, descending-ascending or flat.

• Sometimes it is impossible to detect the type of melody, occurring shifts, biphonations and glides.

According to the characteristics and definitions of pathological crying mentioned above, as well as considering the opinion of expert doctors, the following knowledge based rule could be established: *if a cry has a descending, descending-ascending melody type, or without melodic form, in addition to having shifts, and glides in more than 70% of the total crying, is considered as: crying with a tendency to pathological*.

In the present work, a program was implemented for, in the first instance, to extract crying units from the recordings made by doctors and nurses, to identify the qualitative characteristics. These features are grouped as pathological and non-pathological, along with the total and percentages of shifts and glides, as proposed in [7]. Taking this information as an instance of the inference rule allows to determine the type of crying of the baby.

II. METHODOLOGY

The sample capture process was carried out in two ways, the first with the support of specialist doctors and nurses in the areas of nurseries, primary care and intensive care, at the Hospital del Niño y la Madre Indígena Guerrerense located in the city of Chilpancingo, Capital of the State of Guerrero, as well such as the General Hospital of Tlapa, located in the heart of the high mountain of the state. The second way is an non-hospital capture carried out in the homes of some babies with parents Nahuatl speakers. The captures were made from mobile devices, all the cries were stored in WAV format, with mono digital streams. A capture time limit was not determined and they were stored in a database with a consecutive as well as a type of language, as well as a number to know if it was a first or second time capture.

Some capture incidents collected from support personnel, worth to be known are; some babies in cribs spent whole days without crying. They stayed in hospitals if the mothers had problems during childbirth, so in some cases they did not present any ailment. The audios of the recordings were affected by the noises generated in the hospital environment, for example footsteps from other personnel, respirator alerts, and ambient sounds. Additionally, it was agreed that all samples were taken from spontaneous cries. In order to record the samples in a uniform way, training was provided to the collaborators, who had to observe the following protocol:

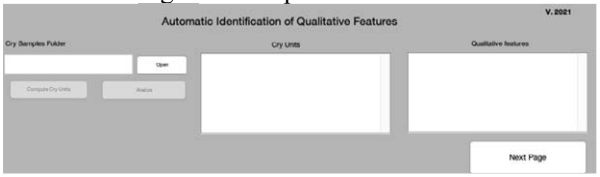
• Babies sampled must be at least 2 days old and not exceed 6 months.

• Place the recording device at least 15 cm away from the baby's mouth.

• Capture in site the following data; date of birth, medical diagnosis, ethnic group, capture number, municipality and locality.

An interface was also developed in Matlab® (fig. 1), which allows the manipulation of samples and facilitates their processing to extract the qualitative characteristics of all the instances in a database and provides a .csv document as an output. In the first stage the folder containing the files with .wav extension, which contain the raw samples of the babies' cries, are processed to form the form the cry units. The software eliminates noise and obtains the fundamental frequency of each unit to proceed to the identification and addition of melodic forms, as well as shifts and glides.

Fig 1. Developed Interface



Clicking NEXT shows the results in a table that can later be downloaded in csv format with the diagnoses automatically predicted.

III. RESULTS

In Table 1, the sampled babies (Px) are listed, as well as the ethnic group to which they belong; Amuzgo (Am), Náhuatl (Nh), Mixteco (Mx), and Tlapaneco (Tl). In the second column the initial diagnosis, given by medical doctors when the sample was taken, is shown. The diagnosed pathologies in the sampled new born babies are; Normal (Nor), Respiratory distress syndrome (RDS), Hyperbilirubinemia (Hip), Pneumonia (Neu), Malnutrition (Des). Next the total of pathological characteristics (CP) and normal characteristics (CN) found in the crying units, are shown. In the S/G column are the total of Shifts and Glides in the crying units. Finally, u30%, u50%, u70% specify the threshold that allows establishing the automatic diagnosis defining whether there is crying with a tendency to pathological (TP), or the crying is normal (Nor).

Table 1. Some results with different thresholds

Px/Et	Dx	C.P	C.N.	S/G	u30%	u50%	u70%
06/Am	Nor	110	16	68%	T.P.	T.P.	Nor
07/Am	Nor	242	32	44%	T.P.	Nor	Nor
17/Am	SDR	134	21	59%	T.P.	T.P.	T.P.
08/Nh	Hip	164	165	42%	T.P.	T.P.	T.P.
09/Nh	SDR	85	42	60%	T.P.	T.P.	T.P.
15/Nh	SDR /Neu	91	99	73%	T.P.	T.P.	T.P.
02/Tl	Nor	222	137	50%	T.P.	T.P.	T.P.
13/Tl	SDR /Des	46	93	42%	Nor	T.P.	T.P.
22/Tl	Nor	203	35	28%	T.P.	Nor	Nor
11/Mx	SDR	107	187	68%	T.P.	T.P.	T.P.
25/Mx	Nor	54	17	8%	T.P.	T.P.	Nor
26/Mx	Nor	20	6	61%	T.P.	T.P.	Nor

Table 1 shows the automatically identified diagnoses. It can be noted that by establishing the threshold contemplated in the 30% rule, 90% of the diagnoses are shown with a tendency to pathological, when the threshold increases to 50%, 80% of the diagnoses are shown with a tendency to pathological. On the other hand, when the threshold goes below 70%, the cries are correctly classified, agreeing with the diagnoses determined by the experts. Table 1 shows the automatically identified diagnoses. It can be noted that by establishing the threshold contemplated in the 30% rule, 90% of the diagnoses are shown with a tendency to pathological, when the threshold increases to 50%, 70% of the diagnoses are shown with a tendency to pathological. On the other hand, when the threshold is 70%, the cries are correctly classified, agreeing with the diagnoses determined by the experts. In order to verify the effectiveness of the proposed method and the validity of the threshold established, a comparison was done with the BabyChillanto Database, of INAOE, with normal babies and babies with hyperbilirubinemia

Table 2. Normal Results u70%

Px	Dx	C.P.	C.N.	S/G	u70%
52a	Normal	16	15	50%	Normal
52b	Normal	24	12	70%	Normal
52c	Normal	10	6	73%	Normal
52d	Normal	6	1	77%	Normal
52e	Normal	23	6	66%	Normal
52f	Normal	10	5	68%	Normal
52g	Normal	9	7	79%	Normal
52h	Normal	9	1	79%	Normal
52i	Normal	16	7	73%	Normal
52j	Normal	18	2	64%	Normal

In Table 2, the prediction (Dx) of the BabyChillanto database is shown, with cries of the Normal class, when performing the test with a threshold of 70% (u70%) in all cases the cries were classified as Normal.

Table 3. Hiperbilirubinemia Results at u70%

Px	Dx	C.P.	C.N.	S/G	u70%
01	Hiperbilirrubinemia	49	23	73%	T.P.
02	Hiperbilirrubinemia	18	12	52%	T.P.
03	Hiperbilirrubinemia	25	16	70%	T.P.
04	Hiperbilirrubinemia	12	1	90%	T.P.
05	Hiperbilirrubinemia	20	12	72%	T.P.
06	Hiperbilirrubinemia	11	0	95%	T.P.
07	Hiperbilirrubinemia	1	0	99%	T.P.
08	Hiperbilirrubinemia	1	0	99%	T.P.

As shown in Table 3, the same test was performed with data from the BabyChillanto Database belonging to the Hyperbilirubinemia class. 100% of the diagnoses generated automatically with the threshold equal to 70% (u70%) determined that the sample has a pathological tendency, which confirms the class established in the database.

IV. DISCUSSION

As shown in the results, the predictions obtained automatically through our system were possible after we were able to set up the right threshold which, consequently, allowed the rule to correctly classify the cries. When obtaining the first results with the support and supervision of the experts, we decided to test

different values of the threshold until setting up 70% as the best value for this database. The normalization of the samples was achieved in the first instance based on the time of each of the patients' cries, homologating their duration to a minimum of 1 minute and a maximum of 3 minutes. In addition to it and in order to have comparable samples, the average obtained for each baby was 184 crying units. Previously, the medical specialists reviewed the diagnoses of the sampled patients, validating the information of some and correcting the information of others in order to confirm the initial diagnoses by the doctors and nurses of the different shifts at the moment the samples were taken. In this review the specialists verified that the case of sample 08/Nh and 09/Nh corresponds to the same neonate at different moments of capture, the diagnosis had changed from Hyperbilirubinemia in the first registration of his record to SDR at a later date. It is worth mentioning that our system found that the crying in both cases showed pathological trend.

A similar case occurred with the neonate 25/Mx and 26/Mx, for this situation the initial diagnosis reported in his file said "to confirm respiratory distress syndrome" (RDS). However, after 3 days of the initial capture, the medical diagnosis was changed to Normal. For both cases our proposed system predicted from the first capture, that they were normal cry.

V. CONCLUSION

This study allowed us to carry out research for the first time over infant cry in Mexican indigenous groups. Literature and overall studies are scarce for these sectors of the original populations. In short, we did not find studies of extraction of qualitative characteristics of crying of indigenous babies, for which we consider that this research could give rise to new gaps and horizons to deepen into issues related to any of the 364 dialectical variants. In addition the system developed and presented in this article may be tested with real patients in any medical unit or general hospital within the health sector, helping to give opportune care and a better diagnosis on babies with still very limited communication skills.

An interesting finding derived from the analysis of the samples was the variability of the sound intensity of the cries. When listening to each one of them it is noticeable that the babies that came from the most vulnerable regions were heard weaker than those from the less vulnerable ones, especially between the Tlapaneco and Mixtec languages, even when using the same capture instrument. We consider that there is an important relationship between intensity of crying and tongue and in turn nutritional status of pregnant mothers, this being a study that could be materialized in future research.

ACKNOWLEDGMENTS

The present work was supported by CONACYT (Posdoctoral Stays in Mexico: Id-154867) The authors thank the support of INAOE for the facilities to elaborate this investigation as well as the expert collaborators from HNMIG and HGT.

REFERENCES

[1]https://www.iwgia.org/en/mexico/3625-iw-2020-mexico.html
[2]https://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG%202015%20rev%20(July%201).pdf
[3]M. Antonia Ruíz Díaz, Carlos A. Reyes, Luis C. Altamirano, Análisis del Llanto Infantil: Identificación Automática de Características Cualitativas del Llanto Infantil para la Ayuda de un Diagnóstico Oportuno, Editorial Académica Española, 104 pages, ISBN-10: 3845486813, ISBN-13: 978-3845486819, August 2011.
[4]C. Manfredi, R. Viellevoye, S. Orlandi,, A. Torres-García, G. Pieraccini,C.A. Reyes-García, “Automated analysis of newborn cry: relationships between melodic shapes and native language”, in Biomedical Signal Processing and Control, vol 53, 2019, pp 1-10, ISBN 1746-8094,
https://doi.org/10.1016/j.bspc.2019.101561, Q2 Health Informatics.
[5] Farris, K. B. D. (2004). Diccionario básico del mixteco. México
[6]Aburto M., Pacual. Mason, David. y Mason, Elena. Tiuelis titlajtlajtos nahuatl. Puede hablar el náhuatl. Instituto Lingüístico de Verano, A.C. México, 2004
[7]Reyes-García, C. A., Torres-García, A. A., & Ruiz-Diaz, M. A. (2018, October). Extracción de Características Cualitativas del Llanto de Bebé y su Clasificación para la Identificación de Patologías Utilizando Modelos Neuro-Difusos. In Memorias del Congreso Nacional de Ingeniería Biomédica (Vol. 5, No. 1, pp. 106-109).
[8]Launey, Michel (1992). Introducción a la lengua y a la literatura náhuatl. México: UNAM
[9]Sarmiento-Silva, Sergio (2001). Tlapanecos de Guerrero. Proyecto Perfiles Indígenas de México, Documento de trabajo.
[10]Wermke, K., Lind, K., Development of the vocal fundamental frequency of spontaneous cries during the first 3 months. Int J Pediatric Otorhinolaryngology, 1999.

AUTOMATING QUASI-STATIONARY SPEECH SIGNAL SEGMENTATION IN SUSTAINED VOWELS: APPLICATION IN THE ACOUSTIC ANALYSIS OF PARKINSON'S DISEASE

A. Tsanas[1], A. Triantafyllidis, S. Arora [3]

[1] Usher Institute, Medical School, University of Edinburgh, Edinburgh, UK
[2] Information Technologies Institute, Centre for Research and Technology Hellas
[3] Mathematical Institute, University of Oxford, Oxford, UK
(A. Tsanas): atsanas@ed.ac.uk; (A. Triantafyllidis): atriand@gmail.com; (S. Arora): arora@maths.ox.ac.uk

Abstract: **Acoustic analysis of sustained vowels is typically used to quantify perturbations in fundamental frequency (F0), amplitude, and deviations from periodicity, and associate these with clinical outcomes of interest. Computational and practical constraints suggest that 2-3 seconds are often sufficient to acoustically characterize a sustained vowel phonation. The question then is how to best determine a short quasi-stationary segment from a typical 20-30 seconds speech recording. We computed the F0 contour in 10 millisecond epochs using SWIPE, a state-of-the-art F0 estimation algorithm, which we had previously demonstrated is very competitive in F0 estimation for sustained /a/ vowels. Subsequently, we determined the two second signal segment that exhibits the smallest mean absolute successive F0 difference. We tested the segmentation algorithm on 100 randomly selected sustained vowel /a/ phonations from the Parkinson's Voice Initiative, where we had hand-labeled the quasi-stationary segments. We found the algorithm correctly identified the quasi-stationary segments in all cases, thus demonstrating it can be deployed at large scale studies automating further processing of sustained vowels. We also demonstrated that this pre-processing step can have a major influence in the acoustic characterization of the phonations.**

Keywords: **acoustic analysis, F0 estimation, speech signal segmentation, sustained vowels**

I. INTRODUCTION

The use of sustained vowels to assess voice disorders is well established in clinical practice [1]. Compared to conversational speech or reading out loud specific abstracts of phonetically rich text, sustained vowels have the advantage that they circumvent linguistic confounds and accent effects [1]. The acoustic analysis of sustained vowels towards the development of robust clinical decision support tools has received considerable research attention. Indicatively, we had previously used sustained vowel /a/ phonations to demonstrate: (i) almost 99% accurate differentiation of people diagnosed with Parkinson's Disease (PD) from Healthy Controls (HC) [2]; (ii) accurate replication of the most widely clinical tool assessing overall PD symptom severity reporting an error that is considerably lower than the inter-rater variability [3]–[6]; (iii) assessing PD voice rehabilitation [7]; and (iv) potential on early PD diagnosis/precursors [8], [9]. Researchers have also developed mechanistic models of speech articulation using sustained vowels, which may provide insights into the underlying vocal production mechanism and voice disorders in a physically interpretable way [10], [11].

In practice, the raw speech signal recordings typically include the prompt by the researcher/clinician, possibly some prior discussion, and one or more prolonged sustained vowel phonations by the study participant. Using the entire sustained vowel phonation (typically 20-30 seconds) is computationally demanding and may be prone to problems (e.g. participant coughing, running out of breath). Computational and practical constraints suggest that processing 2-3 seconds of the sustained vowel phonation are sufficient to acoustically characterize the sustained vowels [1], [12] and to develop mechanistic models [10].

The natural question then arises on how best to choose the short signal segment from the raw recording for further processing. Often, this is done manually by selecting the segment that 'looks best' (low amplitude and low frequency variation) or by selecting a pre-specified signal segment (e.g. the middle of the phonation because that would likely be a stable part of the phonation). For small datasets it may be possible to manually detect segments, however as we move on larger datasets, such as with the Parkinson's Voice Initiative (PVI) study with more than 18,000 sustained vowel phonations [13], [14], the need to develop an automated approach becomes obvious. Previous work in the context of speech signal analysis has focused on removing the non-sustained vowel segment of the recording (e.g. the prompts by investigators and silences). Surprisingly, to the best of our knowledge there is no published work on principled objective detection of short signal sustained vowel segments that

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

would be best applicable towards further acoustic analysis. Moreover, this crucial pre-processing step is rarely reported in the research literature.

If we revisit the underlying principle of using sustained vowels, the aim is to elicit "stable" phonations and assess deviations from signal periodicity [1]. In practice, minor perturbations from maintaining constant amplitude and frequency are common even for people with no vocal pathologies, where larger fluctuations may be hinting towards a vocal pathology (which may be secondary e.g. to PD or other disorders) [1]. Hence, if we want to work on a short signal segment it would be reasonable to identify the most stable part of the phonation. Technically, that would be the most quasi-stationary segment, where stationarity suggests that the central order moments of the signal remain constant [15]. Relaxing the requirement of quantifying non-stationarity, we can instead aim to quantify changes in the fundamental frequency (F0), i.e. the F0 *contour*. The F0 is a key characteristic of speech and its computation is often a pre-requisite for many speech signal processing algorithms [1], [12].

The aim of this study is to develop an algorithmic approach towards automatically detecting the most quasi-stationary short signal segment from a speech recording that comprises a longer sustained vowel phonation which might also exhibit background noise (prompts, silence etc.). We demonstrate the effectiveness of the proposed approach towards the acoustic characterization of PD voices, although in principle the developed method is generalizable across applications focusing on sustained vowels.

II. METHODS

A. Data

We used data from the large PVI study [13], [14], which was set in seven major geographical locations. Participants were invited to call in a dedicated phone number and contribute two sustained vowel /a/ phonations along with basic demographic information (age, gender), and whether they had been clinically diagnosed with PD. The phonations were sampled at 8 kHz and stored on secure cloud servers. For the purposes of this study we have randomly selected phonations from 50 PD participants and from 50 control participants from the US cohort.

B. F0 estimation and signal segmentation

We had previously performed a thorough empirical comparison of multiple F0 estimation algorithms to establish the most accurate for the analysis of sustained vowels [16]. We had found that the Sawtooth Waveform Inspired Pitch Estimator (SWIPE) [17] was very competitive [16] and hence it was used in this study. We used 10 msec epochs to obtain the F0 contour in accordance to standard practice [1], [12], [16].

Following the computation of the F0 contour, we subsequently aimed to determine the short signal segment that exhibited the smallest mean absolute successive F0 differences (without loss of generality we searched for the best short segment of 2 seconds in duration). For convenience, we will simply use the term jitter later on to refer to the mean absolute successive F0 differences. We remark that alternative definitions of jitter variants (F0 perturbations) are possible [1], [4], [12]; here we wanted to explore the simplest approach.

C. Manual hand-labeling of quasi-stationary segments

We have manually hand-labeled the quasi-stationary segments of the 100 speech recordings by aural and visual inspection (e.g. that the quasi-stationary window appears between 4th to the 12th second). We assessed whether the 2-second segment determined by the proposed segmentation algorithm falls completely within the hand-labeled segments.

D. Acoustic analysis of speech segment

We used the Voice Analysis Toolbox which we had previously developed (open source MATLAB code, available at https://www.darth-group.com/software) for the analysis of sustained vowels [5], [12], [18]. We extracted 307 acoustic features which characterize the speech signal: broadly, these features quantify frequency changes (jitter variants), amplitude changes (shimmer variants), signal-to-noise ratio concepts, F0 variability using wavelets, and envelope modulation. For further information on the acoustic features, their algorithmic expression and their tentative interpretation please refer to the Voice Analysis Toolbox and the cited studies above. These features have been previously explored in detail in our PD work [4], [6], [9], [12].

We applied the algorithmic expressions for the computation of the acoustic features using two different segments for comparison: (i) the segment between 1-3 seconds, and (ii) the automatically determined 2-second segment with the algorithm in this study.

III. RESULTS

Fig. 1 presents an indicative sustained vowel recording and the F0 contour to visually illustrate the result of the segmentation algorithm. As a first step, we verified across all phonations used in the study that the automatically detected segment was indeed a short signal where the F0 variability appeared minimal and matched the hand-labeled quasi-stationary segments. Fig. 2 is the zoomed version of Fig. 1 focusing only on the selected signal segment. We can visually observe

from Fig. 1 that if we had pre-fixed a segment at the middle section of the phonation this would have included some large F0 fluctuations. This problem could have occurred at any point in the phonation, which cautions on the use of pre-fixed time segments for further acoustic analysis.

So far, we have demonstrated that the proposed segmentation algorithm correctly identified a short quasi-stationary segment within a speech recording. The next question is whether this makes any practical difference in the subsequent step with the acoustic characterization of the phonation. Table 1 provides summary statistics across some indicative acoustic features (selected to be representative of different acoustic feature families). We remark that some of the acoustic features exhibit considerable differences in the summary values, which indirectly suggests that this pre-processing segmentation step can have a major influence on the reported results.

IV. Discussion

We have developed a robust algorithmic approach towards detecting the quasi-stationary speech signal segment in sustained vowel /a/ phonations that exhibits the lowest F0 fluctuations. This was achieved by first estimating the F0 contour in 10 msec epochs (which is standard in F0 estimation), and subsequently determining the two consecutive seconds segment that exhibited the lowest jitter. We visually verified that in all cases the algorithm had correctly identified a short signal segment where F0 does not fluctuate considerably (see Fig. 2). Finally, we reported that the signal segment that is passed for further processing affects the computed acoustic features (see Table 1).

Although segmentation is a well-researched area in the signal processing and image processing research literature, we are not aware of any similar work that presents a principled approach towards determining a short speech segment within sustained vowels which would be a useful pre-processing step prior to further acoustic analysis. For example, Badawy et al. [19] attempted to correctly estimate the entire duration of the sustained vowel phonation, whereas here we aimed to determine the most quasi-stationary segment within a full recording. Other work has focused on removing silences in recordings [20] so that only the voiced segment could be presented to acoustic analysis algorithms. We remark that our algorithm can intrinsically automatically detect when the F0 fluctuations are above a maximum threshold of F0 fluctuations or unrealistic F0 ranges (e.g. silence recordings, background noise) and hence identify phonations of insufficient quality, prompting further investigation or rejecting those recordings from further processing.

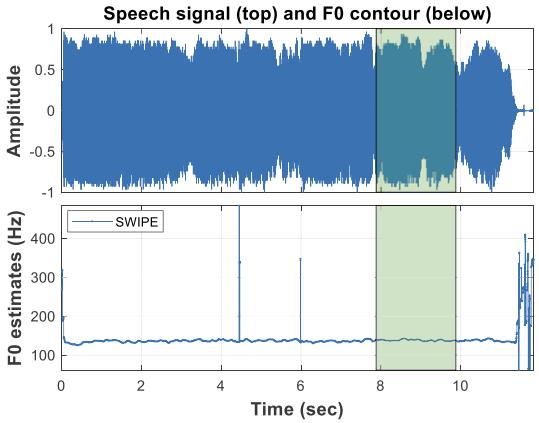


Fig. 1: *Indicative plot visually illustrating the selected signal segment (in transparent green) both in terms of the raw voice signal and the computed F0 contour.*

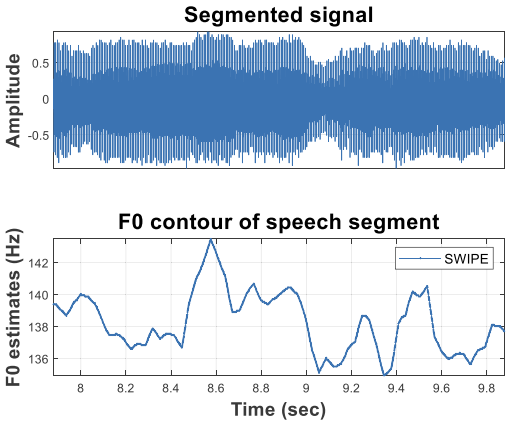


Fig. 2: *Focusing on the segmented signal and the F0 contour (zoomed in version from Fig. 1).*

Table 1: Summary statistics of indicative acoustic features for the phonations used in the study.

Indicative acoustic features	**Benchmark segment (1-3 sec)**	**Automatically determined segment**
Jitter	1.65±2.37	1.33±1.94
Shimmer	0.21±0.07	0.20±0.06
HNR	8.36±10.21	8.49±10.46
GNE	1.47±0.36	1.45±0.40
EMD-ER$_{NSR,TKEO}$	5.87±2.98	7.24±3.70
VFER$_{TKEO}$	0.72±0.59	2.51±1.74

The features are summarized in the form mean±standard deviation. HNR = Harmonics to Noise Ratio, GNE = Glottal to Noise Excitation, EMD-ER = Empirical Mode Decomposition Excitation Ratio, VFER = Vocal Fold Excitation Ratio. For the algorithmic definition of the features in the Table see [12].

This study focused exclusively on sustained vowels /a/ phonations. We remark that in principle these findings should generalize well in other settings with

sustained vowel phonations (e.g. the other two corner vowels /i/ and /u/), but that remains to be tested. So far, we are not aware of any work that has empirically extensively tested F0 estimation algorithms beyond /a/, and future work would likely also need to be done for other vowels or phonetically rich sounds used in clinical practice [1]. A seemingly very different speech signal analysis area to sustained vowels which is, perhaps surprisingly, intrinsically linked is processing of voice fillers. Voice fillers essentially exhibit similar properties to sustained vowels [21] even though they originate in conversational speech, which is a more generic setting where participants are not specifically instructed to produce a specific type of phonation. Previously, we had extracted the corresponding voice fillers for further acoustic analysis manually [21]; in principle, the presented algorithm herein should be generalizable.

We are currently working on extending our early work using the noisy speech data collected as part of the PVI project [13], [14]. This dataset presents considerable challenges because of its large size and the data have not been collected under carefully controlled acoustic conditions. This very challenging setting requires robust methodologies to extract clinically useful information, where automating segmentation and reducing the computations demands on acoustic characterization of phonations is crucial.

Collectively, these results provide a compelling argument that speech segmentation should be carefully considered and reported. This may also have important implications for real-time biomedical signal processing applications (e.g. processing on smartphones), where computational constraints need to be carefully considered. We envisage the proposed algorithm providing a convenient, robust approach to determining a short signal segment from a longer sustained vowel phonation towards standardizing acoustic analysis.

REFERENCES

[1] I. R. Titze, *Principles of voice production*. Iowa City: National Center for Voice and Speech, 2000.

[2] A. Tsanas *et al.*, "Novel speech signal processing algorithms for high-accuracy classification of Parkinsons disease," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1264–1271, 2012

[3] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 884–893, 2010

[4] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity," *J. R. Soc. Interface*, vol. 8, no. 59, pp. 842–855, 2011, doi: 10.1098/rsif.2010.0456.

[5] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "New nonlinear markers and insights into speech signal degradation for effective tracking of Parkinson's disease symptom severity," in *International symposium on nonlinear theory and its applications (NOLTA)*, 2010, pp. 457–460.

[6] A. Tsanas, M. A. Little, and L. O. Ramig, "Remote assessment of Parkinson's disease symptom severity using the simulated cellular mobile telephone network," *IEEE Access*, vol. 9, pp. 11024–11036, 2021

[7] A. Tsanas, M. A. Little, C. Fox, and L. O. Ramig, "Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 1, pp. 181–190, 2014

[8] S. Arora *et al.*, "Investigating voice as a biomarker for leucine-rich repeat kinase 2-associated Parkinson's disease," *J. Parkinsons. Dis.*, vol. 8, no. 4, pp. 503–510, 2018

[9] S. Arora, C. Lo, M. Hu, and A. Tsanas, "Smartphone speech testing for symptom assessment in rapid eye movement sleep behavior disorder and Parkinson's disease," *IEEE Access*, vol. 9, pp. 44813–44824, 2021

[10] P. Gómez-Vilda *et al.*, "Phonation biomechanics in quantifying parkinson's disease symptom severity," in *Recent Advances in Nonlinear Speech Processing*, vol. 48, 2016, pp. 93–102.

[11] A. Gómez *et al.*, "A neuromotor to acoustical jaw-tongue projection model with application in Parkinson's disease hypokinetic dysarthria," *Front. Hum. Neurosci.*, vol. 15, p. 622825, 2021

[12] A. Tsanas, "Accurate telemonitoring of Parkinson's disease using nonlinear speech signal processing and statistical machine learning," University of Oxford, 2012.

[13] 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 voice," *J. Acoust. Soc. Am.*, vol. 145, pp. 2871–2884, 2019.

[14] A. Tsanas and S. Arora, "Biomedical speech signal insights from a large scale cohort across seven countries: The Parkinson's voice initiative study," in *Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, 2019, pp. 45–48.

[15] S. Theodoridis, K. Koutroumbas, *Pattern recognition*, Academic press, 4th ed., 2019

[16] A. Tsanas, M. Zañartu, M. A. Little, C. Fox, L. O. Ramig, and G. D. Clifford, "Robust fundamental frequency estimation in sustained vowels: detailed algorithmic comparisons and information fusion with adaptive Kalman filtering.," *J. Acoust. Soc. Am.*, vol. 135, no. 5, pp. 2885–901, 2014

[17] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music.," *J. Acoust. Soc. Am.*, vol. 124, no. 3, pp. 1638–52, Sep. 2008

[18] A. Tsanas, "Acoustic analysis toolkit for biomedical speech signal processing: concepts and algorithms," in *8th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, 2013, pp. 37–40.

[19] R. Badawy *et al.*, "Automated quality control for sensor based symptom measurement performed outside the lab," *Sensors*, vol. 18, no. 4, pp. 1–22, 2018

[20] T. Giannakopoulos and A. Pikrakis, *Introduction to Audio Analysis: A MATLAB Approach*. Academic Press, 2014.

[21] E. San Segundo, A. Tsanas, and P. Gomez-Vilda, "Euclidean Distances as measures of speaker similarity including identical twin pairs: a forensic investigation using source and filter voice characteristics," *Forensic Sci. Int.*, vol. 270, pp. 25–38, 2017

LONGITUDINAL EFFECT OF REPETITIVE TRANSCRANIAL MAGNETIC STIMULATION ON PHONATION IN A PATIENT WITH PARKINSON'S DISEASE: A CASE STUDY

A. Gómez[1], J. Mekyska[2], L. Brabenec[3], P. Simko[3,4], I. Rektorova[3,5], P. Gómez[6], A. Tsanas[1]

[1]Usher Institute, Faculty of Medicine, University of Edinburgh, Edinburgh, UK; {a.gomezrodellar, athanasios.tsanas @ed.ac.uk}
[2]Department of Telecommunications, Brno University of Technology, Brno, Czech Republic; mekyska@vut.cz
[3]Applied Neuroscience Research Group, Central European Institute of Technology – CEITEC, Masaryk University, Brno, Czech Republic; {lubos.brabenec, patrik.simko, irena.rektorova@ceitec.muni.cz}
[4]Faculty of Medicine, Masaryk University, Brno, Czech Republic.
[5]First Department of Neurology, Faculty of Medicine and St. Anne's University Hospital, Masaryk University, Brno, Czech Republic; irena.rektorova@fnusa.cz
[6]NeuSpeLab, CTB, Universidad Politécnica de Madrid, Madrid, Spain; pedro.gomezv@upm.es

Abstract: **The present paper describes a case study exploring the longitudinal effect of repetitive Transcranial Magnetic Stimulation (rTMS) on hypokinetic dysarthria in a patient with Parkinson's Disease (PD). Several correlates from phonation and articulation such as jitter, shimmer, Harmonic-Noise Ratio (HNR), first two formant amplitudes and quality factors, and the Absolute Kinematic Velocity (AKV) were estimated and compared against equivalent features from a normative dataset using the Jensen-Shannon Divergence. The action of rTMS showed a positive impact on the amplitude and quality factor of the first formant, and consequently, on the AKV. These promising findings will need to be explored in further detail in follow up work.**
Keywords: **Transcranial Magnetic Stimulation, Hypokinetic Dysarthria, Parkinson's Disease.**

I. INTRODUCTION

Hypokinetic Dysarthria (HD) is a common manifestation of PD in speech, difficult to treat. Repetitive Transcraneal Magnetic Stimulation (rTMS) has been used to assess its long-term treatment effects on HD in PD [1]. A case study is analized here to observe functional changes after the rTMS, based on acoustic features quantifying phonation (Energy profile in dB *logEn*, *F0* profile, *HNR*, Zero Crossings (*ZX*), jitter (*Jt*) and shimmer (*Sh*)), and vowel articulation stability (first two formant profiles (vF_1, vF_2), formant bandwidth correlates (mF_1, mF_2) and Absolute Kinematic Velocity (AKV) [2]).

II. MATERIALS AND METHODS

A PD participant (male, 61 years-old, UPDRS-III=9, LED=990 mg) was monitored uttering a sustained vowel [a:] before (baseline, day 0) and after rTMS (four recordings at 15, 26, 78 and 109 days into the study). The stimulation was performed using the instrumentation and methods described in [1].

A. Signal Analysis

The analysis was based on the sustained vowel [a:]. The signal was low-pass filtered and down-sampled to 8 kHz to maintain compliance with telephone-line quality. It was framed on sliding windows of 128 ms (equivalent to 1024 samples) with a stride of 2 ms (equivalent to 16 samples) to capture fast changes in phonation and articulation phenomena. Each signal frame was inverse-filtered using a lattice LPC filter with order k=9 for male voice and k=7 for female voice.

B. Feature extraction

The following features were estimated for each sliding window $W_e(m)$:

- The squared signal envelope in dB (*logEn*)

$$logEn_m = 10\ log_{10}(\mathcal{F}_{LP}\{X_m\});$$
$$X_m = \sum_{n=1}^{W_e(m)} x_n^2 \quad (1)$$

  where x is the speech signal, n is the time index, m is the index of the Hamming sliding window, and $\mathcal{F}_{LP}\{\cdot\}$ is a fourth-order Butterworth low-pass filter with cutoff frequency at 4 Hz.
- The fundamental frequency F0 estimated from the autocorrelation function R_m [3]

$$F_0 = \frac{1}{\tau\ argmax\{R_{m>1}\}} \quad (2)$$

  where τ is the sampling interval.

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

- The zero-crossing function (*ZX*), estimated from the speech signal, counting the number of sign changes on the estimation window $W_e(m)$.
- The harmonic-to-noise ratio (*HNR*) estimated from the first maximum of the autocorrelation function

$$HNR_{dB} = 10\ log_{10}\left\{\frac{R_{max}}{1 - R_{max}}\right\} \tag{3}$$

to be positive where the energy of harmonic contents is larger than that of turbulent contents.
- The first two formants, estimated from the zeros of the prediction-error polynomial $H_K(z)$ estimated for each window frame (vF_1 and vF_2)

$$H_K(z = z_i) = 0;$$
$$z_i = r_i e^{j\varphi_i};\ \varphi_i \geq 0;\ vF_i = \varphi_i \frac{f_s}{2\pi} \tag{4}$$

- The proximity of each zero in the prediction-error polynomial to the unity circle is used as a quality factor (mF_1 and mF_2)

$$mF_1 = r_1;\ mF_2 = r_2 \tag{5}$$

- The Absolute Kinematic Velocity, as a measure of the neuromotor control of the jaw-tongue system is estimated from vF_1 and vF_2 [1]

$$|v_r(t)| \approx \left[H_1\left(\frac{dF_1(t)}{dt}\right)^2 + H_2\left(\frac{dF_2(t)}{dt}\right)^2\right]^{1/2} \tag{6}$$

where H_1 and H_2 are quadratic forms of formant-velocity projection weights w_{ij}. In this way an estimation of the AKV may be produced exclusively in terms of formant dynamics.

- The perturbation features known as jitter and shimmer. Jitter (*Jt*) is estimated as the difference between the F0 values in two neighbor windows relative to their average as

$$Jt_m = 2\frac{F0_m - F0_{m-1}}{F0_m + F0_{m-1}} \tag{7}$$

- Similarly, Shimmer (*Sh*) is estimated as the difference between the energy *En* of the signal from two neighbor windows relative to their average as

$$Sh_m = 2\frac{En_m - En_{m-1}}{En_m + En_{m-1}} \tag{8}$$

where *m* is the index of the sliding window.

[1] Normative male database (50 speakers, age 30.83±10.37 years), ENT services, Hospital Gregorio Marañón of Madrid.

C. Jensen-Shannon Divergence (JSD)

JSD is a generalization of the measure of the mutual information contents between two probability density distributions of feature ξ_m to be compared $p_i(\xi_k)$ and $p_j(\xi_k)$, given by Kulback-Leibler's Divergence [4]

$$D_{JSk}\{p_i(\xi_k), p_j(\xi_k)\} = \frac{1}{2}D_{KL}\{p_i, p_a\} + \frac{1}{2}D_{KL}\{p_j, p_a\}; \quad p_a = \frac{p_i + p_j}{2} \tag{9}$$

with $1 \leq k \leq K$ being the feature index, and *K* the number of features, where

$$D_{KL}\{p_i, p_j\} = D_{KL}\{p_i(\zeta), p_j(\zeta)\} = -\int_{\zeta=0}^{\infty} p_i(\zeta) log\left[\frac{p_i(\zeta)}{p_j(\zeta)}\right] d\zeta \tag{10}$$

is Kulback-Leibler's Divergence defined for a generic feature $\zeta \geq 0$. For this comparison a normalized histogram must be estimated from each feature amplitude. As an example, the probability density for feature the sampled value of the AKV given in (6), estimated as:

- An N-bin histogram of counts by amplitudes is built from each subject's AKV. The interval covered for speeds is $[0, |v_r|_{max}]$, with $|v_r|_{max}$=50 cm.s^{-1}, and K=50, with bin $\Delta b_k = [|v_r|_{max}/K]$=1 cm.s^{-1} wide.
  The count histogram is built for each bin $b_k = k \cdot \Delta b_k$ c_k being the number of counts for bin b_k
  if $b_{k-1} \leq |v_r(t)| < b_k$ *then* $c_k = c_{k+1}$
- Count histograms c_k ($0 \leq k \leq N$) are normalized to their total number of counts $C_t = \Sigma b_k$ ($0 \leq k \leq N$), to be considered estimators of probability density functions $p_k = c_k / C_t$.

Thence $p(|v_{rk}|) = p_k$ will be an estimate of the AKV probability density function. This feature has proven to be quite relevant in differentiating dysarthric from normative speech [5]. The signal analysis was carried on using an application built on MATLAB [6], and optimized for the fast analysis and feature extraction as well as for the statistical comparison of the case study participant's features against the same ones from an age-matched normative subset of a larger database[1]. For such, the distributions of each feature were compared using (9) against distributions from 16 age-matched male speakers selected from the normative database.

III. RESULTS

The following plots in Fig. 1 show the pre-stimulus phonation, *En*, *F0*, *ZX* and *HNR* profiles.

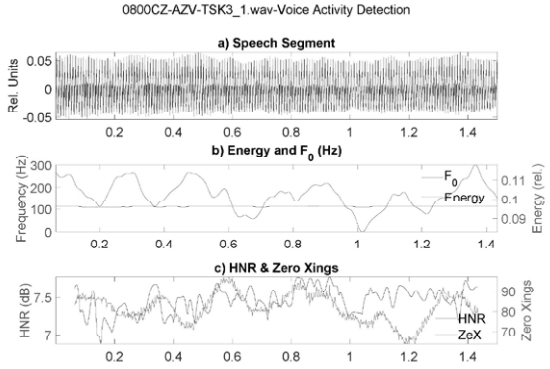


Fig. 1 Example of the signal quality correlates from an emission of vowel [a:] by a 58-year old male participant (pre-stimulus): a) speech; b) *En* and *F0* contours; c) *HNR* and *ZX* contours.

Their respective probability density distributions may be seen in Fig. 2

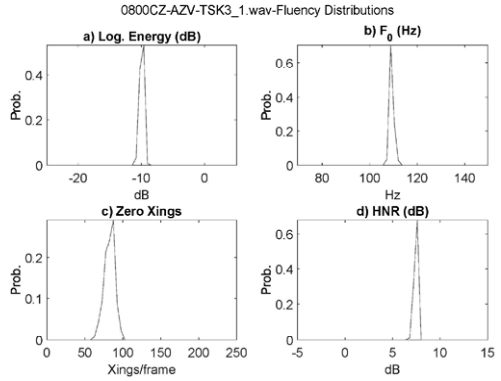


Fig. 2 Distributions of the *logEn*, *F0*, *ZX* and *HNR*.

An example of the two first formant contours and their plot on the vowel triangle are shown in Fig. 3

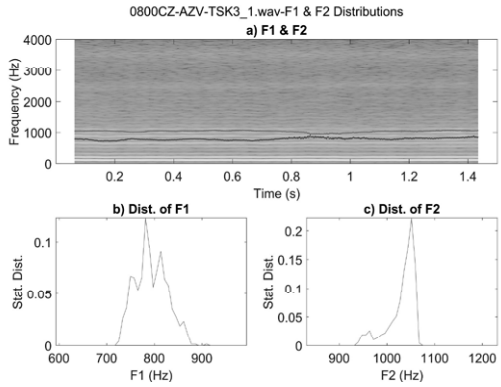


Fig. 3 Formant correlates of the pre-stimulus vocal emission: a) Longitudinal evolution of the first two formants (F_1: blue, F_2: red): b) F_1 density distribution; c) F_2 density distribution.

Important correlates informing on signal quality and estimation robustness are the moduli of the prediction-error polynomial zeros, being shown in Fig. 4.

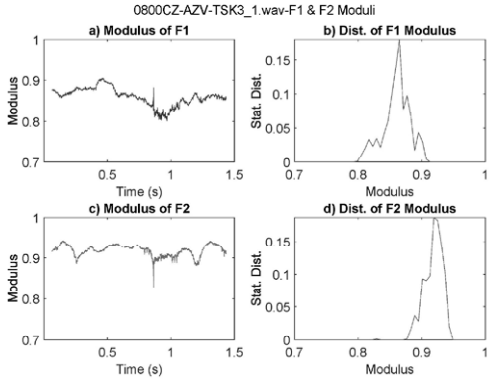


Fig. 4 Formant moduli of the pre-stimulus vocal emission: a) Time evolution of the modulus of F_1: b) density distribution of the modulus of F_1; c) Time evolution of the modulus of F_2: b) density distribution of the modulus of F_2.

The Absolute Kinematic Velocity is a semantic correlate, with a density distribution showing a χ^2-behavior, which can be related to the "emotional temperature" of the speech production [7]. Its temporal evolution and density distribution are shown in Fig. 5.

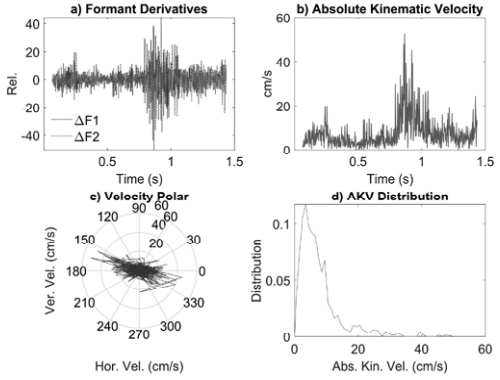


Fig. 5 Formant Kinematics: a) Time derivatives of the first two formants (F_1: blue, F_2: red): b) Absolute Kinematic Velocity from formant derivatives; c) Polar plot of the Absolute Kinematic Velocity; d) AKV density distribution.

The results from the evaluation of the JSD from pre-(baseline) and post-stimulus (15 days, 26 days, 78 days and 109 days, respectively) using the density distributions of the *logEn*, *F0*, *Jt*, *Sh*, *ZX*, *HNR*, vF_1, vF_2, mF_1, mF_2 and AKV are presented in Table 1 (the maximum JSD relative to the normative subset for each feature is given in bold).

Table 1. Statistical summary of acoustic features over specific dates in the monitoring period (maxima given in bold).

Record	*logEn*	*F0*	*Jt*	*Sh*	*ZX*	*HNR*	vF_1	vF_2	mF_1	mF_2	AKV
Pre (0)	0.399	0.239	0.030	0.275	0.683	0.526	**0.419**	0.135	**0.694**	0.362	**0.394**
15 days	**0.425**	0.124	**0.051**	0.320	0.636	0.604	0.073	0.173	0.674	0.655	0.327
26 days	0.258	**0.406**	0.030	0.317	0.670	0.611	0.079	**0.227**	0.665	0.352	0.264
78 days	0.366	0.208	0.045	0.363	**0.694**	**0.665**	0.124	0.101	0.658	**0.657**	0.359
109 days	0.412	0.197	0.039	**0.411**	0.693	0.661	0.187	0.107	0.532	0.387	0.214

The timely evolution of vF_1, mF_1 and AKV is graphically displayed in Fig. 6.

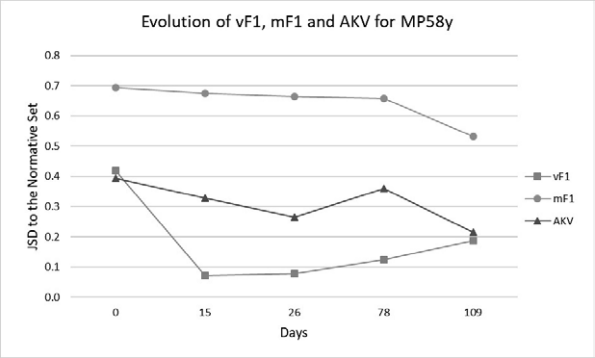


Fig. 6 Longitudinal evolution of the JSD for the first formant value (vF_1), quality factor (mF_1) and AKV from the pre- and four post-stimulus sessions from the 58-y old male participant against the normative subset.

IV. Discussion

In analyzing the results, it must be taken into account that the larger the JSD, the bigger the difference from the normative reference is. The results presented in Table 1 and Fig. 6 show that three acoustic features related with jaw movement manifest a drift towards normativity after rTMS, which may be interpreted as a beneficial effect regarding HD. In fact, it may be seen that vF_1, mF_1 and AKV show a clear trend towards the normative data (with a monotonic trend in mF_1). On its turn, vF_1 shows a decay at the first recording session (15 days) followed by a regression to worse conditions in the last three sessions. AKV shows a descent except at the fourth session (78 days). It must be mentioned that these preliminary results need further evaluation on a larger size database, and that the analysis of the prosodic profile and fluency are not covered in the present study.

V. Conclusions

The case studied shows some beneficial timely effects of rTMS on the stability and quality of the first formant. This behavior is also manifested in the joint jaw-tongue kinematics (AKV) possibly due to a stabilization of the jaw-tongue neuromotor control mechanisms. These tentative findings are to be confirmed with an extended study on more participants of both genders.

Acknowledgments

This project received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 734718 (CoBeN) and from a grant from the Czech Ministry of Health, 16-30805A, and from grants TEC2016-77791-C4-4-R (Ministry of Economic Affairs and Competitiveness of Spain), and Teca-Park-MonParLoc FGCSIC-CENIE 0348_CIE_6_E (InterReg Programme).

References

[1] L. Brabenec et al., "Non-invasive brain stimulation for speech in Parkinson's disease: A randomized controlled trial", *Brain Stimulation*, Vol. 14, 2021, pp. 571-578.

[2] P. Gómez et al., "Characterization of Parkinson's disease dysarthria in terms of speech articulation kinematics", *Biomedical Signal Processing and Control*, Vol. 52, 2019, pp. 312-320.

[3] A. Tsanas, M. Zañartu, M. A. Little, C. Fox, L. O. Ramig, and G. D. Clifford, "Robust fundamental frequency estimation in sustained vowels: Detailed algorithmic comparisons and information fusion with adaptive Kalman filtering", *The Journal of the Acoustical Society of America*, Vol. 135, No. 5, 2014, pp. 2885-2901.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, New York, Wiley-Interscience, 1991.

[5] P. Gómez, J. Mekyska, J. M. Ferrández, D. Palacios, A. Gómez, V. Rodellar, Z. Galaz, Z. Smekal, I. Eliasova, M. Kostalova, I. Rektorova, "Parkinson Disease Detection form Speech Articulation Neuromechanics", *Frontiers on Neuroinformatics*, Vol. 11, 2017, pp. 1-17.

[6] NeuSpeLab, CTB, UPM, Project MonParLoc: https://monparloc.github.io

[7] J. B. Alonso, J. Cabrera, M. Medina, C. M. Travieso, "New approach in quantification of emotional intensity from the speech signal: emotional temperature", *Expert Systems With Applications*, Vol. 42, 2015, pp. 9554-9564.

ACOUSTIC ANALYSIS OF SUSTAINED VOWELS IN PARKINSON'S DISEASE: NEW INSIGHTS INTO THE DIFFERENCES OF UK- AND US-ENGLISH SPEAKING PARTICIPANTS FROM THE PARKINSON'S VOICE INITIATIVE

A. Tsanas[1], S. Arora [2]

[1] Usher Institute, 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: **Sustained vowels have often been used to clinically assess vocal performance and infer symptoms in Parkinson's disease, with most studies focusing on cohorts from a single linguistic background. Arguably, sustained vowels are generic and language-independent, however it is not clear how findings might generalize across cohorts of people from different linguistic backgrounds. In this study, we aimed to compare phonations from UK- and US-English speaking people with Parkinson's disease using the largest known speech-Parkinson's database collected using a standard telephone network, the Parkinson's Voice Initiative (PVI). We processed 1988 sustained vowel /a/ phonations from the US-cohort and 525 phonations from the UK-cohort. We stratified data according to gender and computed the fundamental frequency (F0) as a function of age and characterized phonations using 307 acoustic measures that we have used in previous related work. There was generally very good agreement between UK- and US-English speakers in terms of F0 characteristics and traditional acoustic measures such as jitter and shimmer. However, we find pronounced cohort differences with a few of the complex nonlinear acoustic measures. These findings provide useful insights into the acoustic differences between two English speaking cohorts, which should be taken into account when generalizing findings.**

Keywords: **acoustic analysis, Parkinson's disease, speech signal processing, sustained vowels**

I. INTRODUCTION

Parkinson's Disease (PD) is a debilitating progressive neurodegenerative disorder with cornerstone symptoms which include tremor, rigidity and bradykinesia, within the broader remit of motor and non-motor symptoms [1]. PD incidence and prevalence rates have been consistently growing where there was an estimated 6.1 million of People with Parkinson's (PwP) in 2016, and this number is projected to grow further as the average life expectancy increases [2]. Vocal impairment is very common in PD [3] and is met in approximately 70-90% PwP [3].

Studies over the last two decades have demonstrated the enormous potential that speech signals have in neurodegenerative applications including PD. Indicatively, we had previously used sustained vowel /a/ phonations and demonstated: (i) differentiating PwP from age- and gender-matched controls with almost 99% accuracy [4], (ii) accurately replicating the gold standard PD symptom severity score with accuracy greater than the inter-rater variability [5]–[9], and (iii) automatically assisting voice rehabilitation [10]. More recently, we reported on the potential of speech signals towards early PD diagnosis both when using information with *LRRK*2 gene mutations [11] and also with known disease precursors such as rapid eye movement sleep behavior disorder [12]. Similarly, we have developed speech articulation kinematic models to characterize PD dysarthria to provide mechanistic insights into the underlying physiology [13]–[15], and explored PD subgroups [16], [17].

The use of sustained vowels towards the assessment of vocal performance has been well established [18] and in particular towards assessing neurodegenerative disorders [18], [19]. Most studies in the PD research literature focus on cohorts from a single linguistic background, e.g. US-English speakers. Although it could be argued that sustained vowels may be language-independent, there has not been a systematic investigation into acoustic characterization in PwP cohorts from different linguistic backgrounds. This may limit potential comparisons and insights which could be drawn when comparing PwP from different linguistic backgrounds. Motivated by the need to assess speech-PD at large, we initiated the Parkinson's Voice Initiative (PVI), an international study that collected sustained vowel /a/ phonations and basic demographic information from approximately 10,000 people [20]–[22]. This is the largest known speech-PD database and provides a unique opportunity for forming new hypotheses and exploratory analyses.

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

In this study, we aimed to compare PwP from UK- and US-English speaking linguistic backgrounds across a range of acoustic characteristics to investigate alignment at a cohort level using age and gender stratification.

II. METHODS

A. Data

The study makes secondary analysis of the PVI data focusing on the UK- and US-English speaking cohorts. We processed 1988 sustained vowel /a/ phonations from the US-cohort and 525 phonations from the UK-cohort. The speech recordings were sampled at 8 kHz and were stored at secure Aculab servers, along with basic demographic information (age, gender). For further details on PVI please see our previous work [20]–[22].

B. Acoustic analysis of sustained vowels

We computed the fundamental frequency (F0) contour using SWIPE [23], which we had previously demonstrated is very competitive in accurate F0 estimation specifically for sustained /a/ vowels [24]. We also used the Voice Analysis Toolbox (MATLAB open source code: https://www.darth-group.com/software), which provides an overview of acoustic characterization of sustained vowels across 307 acoustic measures. These have been specifically developed for PD applications [5], [6], [19], [25] and were later shown to be more broadly applicable to other settings including general voice pathology assessment [26] and forensic phonetics [27]. We compared the UK and US-English speaking cohorts in terms of average F0 and F0 trajectories stratified by age and gender to objectively illustrate overall cohort differences. Also, we compared the cohort distributions across the computed 307 acoustic measures to demonstrate how well these align in the two PwP groups.

III. RESULTS

Fig. 1 presents the average estimated F0 as a function of age, where results are stratified by gender. We observe that the general trend is similar for both cohorts,

Table 1: Indicative acoustic measures of people with Parkinson's, stratified by gender

Acoustic measure	**Brief explanation**	**US cohort (males)**	**US cohort (females)**	**UK cohort (males)**	**UK cohort (females)**
Mean F0	Mean fundamental frequency (F0) computed using SWIPE	139.61±34.03	206.84±33.24	139.17±33.79	216.25±32.98
Jitter	Average successive F0 differences (10 ms windows)	0.49±1.35	0.23±0.64	0.43±1.29	0.21±0.54
Shimmer	Average successive amplitude differences (10 ms windows)	0.10±0.04	0.09±0.04	0.09±0.04	0.10±0.05
NHR	Noise-to-harmonics ratio	0.10±0.24	0.05±0.16	0.06±0.09	0.04±0.14
GNE	Glottal to noise excitation (assessing SNR)	0.88±0.17	1.08±0.21	0.86±0.11	1.09±0.20
$VFER_{mean}$	Vocal fold excitation ratio, average frequency excitation	2.18±2.49	0.95±3.05	2.25±2.34	1.36±3.40
$VFER_{SNR-TKEO}$	Vocal fold excitation ratio, SNR energy excitation	257.40±473.70	313.12±519.29	677.63±835.43	885.88±823.73
PPE	Pitch period entropy (assessing F0 variability)	0.05±0.10	0.02±0.06	0.03±0.08	0.02±0.06
0^{th} MFCC	0th Mel Frequency Cepstral Coefficient	0.92±2.28	1.18±2.24	-0.30±2.11	0.04±2.01
1^{st} MFCC	1st Mel Frequency Cepstral Coefficient	2.10±1.74	1.32±1.67	3.97±1.69	3.40±1.28
12^{th} MFCC	12th Mel Frequency Cepstral Coefficient	0.10±0.40	-0.57±0.47	0.22±0.40	-0.28±0.49

Distributions are summarized in the form mean ± standard deviation. GNE = Glottal to Noise Excitation, MFCC = Mel Frequency Cepstral Coefficient, SNR = Signal to Noise Ratio, VFER = Vocal Fold Excitation Ratio.

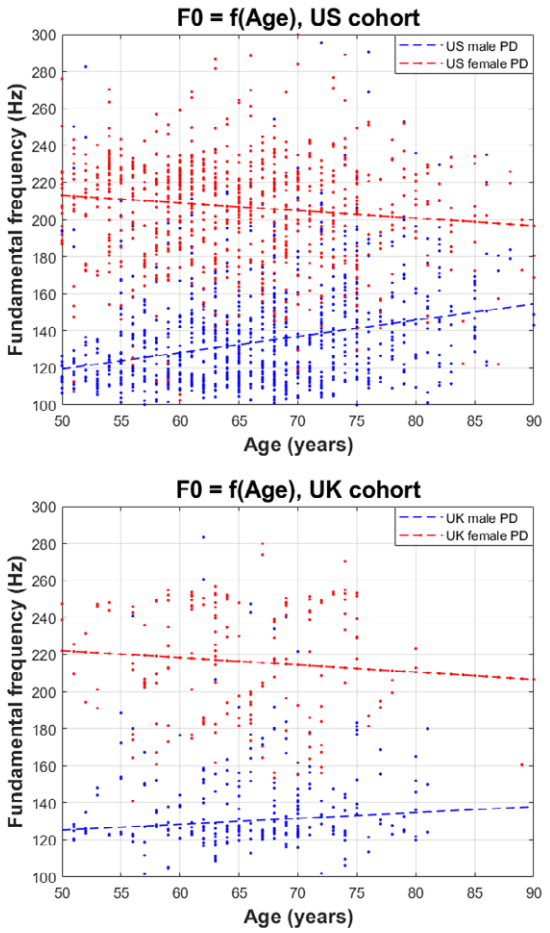


Fig. 1: *Fundamental frequency (F0) as a function of age, stratified by gender for the UK and US PwP cohorts. The best line was computed using robust linear regression fit with iteratively reweighted least squares*.

where the average F0 is increasing with age. However, for both male and female PwP the US cohort exhibit a sharper rate of change.

Table 1 summarizes indicative acoustic measures of the two PwP cohorts to facilitate a side-by-side comparison, stratified by gender. We remark that the classical acoustic measures (e.g. jitter, shimmer, NHR) were very similar. However, there were subtle and pronounced differences in some acoustic measures, in particular the Vocal Fold Excitation Ratio (VFER) measures and Mel Frequency Cepstral Coefficient (MFCC) measures.

IV. Discussion

This study investigated the use of sustained vowel /a/ phonations between speakers from UK- and US-English linguistic backgrounds across a range of acoustic measures. Overall, there was generally very good agreement between the two cohorts in terms of F0 characteristics and most of the acoustic measures investigated. This is a strong indication that clinical decision support tools developed using sustained vowel /a/ phonations in English-speaking PwP cohorts should in principle generalize to other English-speaking PwP. However, there are some subtle pronounced cohort differences with some of the acoustic measures (VFER, MFCCs), which need to be considered when generalizing findings across cohorts with different linguistic backgrounds.

VFER and MFCCs have been particularly successful in related PD clinical decision support tools that we had previously reported on using either UK- or US-English speaking cohorts [7], [12], [19]. The present study's findings could indicate that clinical decision support tools developed across either PwP cohort might need some careful tuning to be generalizable, for example exploring options with transfer learning. In turn, this could also inherently suggest that the PVI cohorts (data collected across seven countries) should be investigated separately to report on individual cohort properties and provide a cross-linguistic comparison of acoustic measure outputs and F0 changes as a function of age.

V. Conclusion

Collectively, these findings support the use of sustained vowels towards vocal assessment in PD as a robust and broadly generalizable signal modality, at least in the English-speaking cohorts. However, care needs to be exercised with some of the acoustic measures (VFER, MFCCs) which appear to differ considerably between cohorts.

References

[1] B. R. Bloem, M. S. Okun, and C. Klein, "Parkinson's disease," *Lancet*, vol. 12, pp. 2284–2303, 2021

[2] E. R. Dorsey *et al.*, "Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016," *Lancet Neurol.*, vol. 17, pp. 939–953, 2018

[3] A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates, "Speech impairment in a large sample of patients with Parkinson's disease," *Behav. Neurol.*, vol. 11, no. 3, pp. 131–137, 1998

[4] A. Tsanas *et al.*, "Novel speech signal processing algorithms for high-accuracy classification of Parkinsons disease," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1264–1271, 2012

[5] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity.," *J. R. Soc. Interface*, vol. 8, no. 59, pp. 842–

855, 2011
[6] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "New nonlinear markers and insights into speech signal degradation for effective tracking of Parkinson's disease symptom severity," in *International symposium on nonlinear theory and its applications (NOLTA)*, 2010, pp. 457–460.
[7] A. Tsanas, M. A. Little, and L. O. Ramig, "Remote assessment of Parkinson's disease symptom severity using the simulated cellular mobile telephone network," *IEEE Access*, vol. 9, pp. 11024–11036, 2021
[8] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 884–893, 2010
[9] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Enhanced classical dysphonia measures and sparse regression for telemonitoring of Parkinson's disease progression," *2010 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 594–597, 2010
[10] A. Tsanas, M. A. Little, C. Fox, and L. O. Ramig, "Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 1, pp. 181–190, 2014
[11] S. Arora *et al.*, "Investigating Voice as a Biomarker for leucine-rich repeat kinase 2-Associated Parkinson's Disease," *J. Parkinsons. Dis.*, vol. 8, no. 4, pp. 503–510, 2018
[12] S. Arora, C. Lo, M. Hu, and A. Tsanas, "Smartphone speech testing for symptom assessment in rapid eye movement sleep behavior disorder and Parkinson's disease," *IEEE Access*, vol. 9, pp. 44813–44824, 2021
[13] A. Gómez *et al.*, "A Neuromotor to Acoustical Jaw-Tongue Projection Model With Application in Parkinson ' s Disease Hypokinetic Dysarthria," *Front. Hum. Neurosci.*, vol. 15, p. 622825, 2021
[14] P. Gómez-Vilda *et al.*, "Phonation biomechanics in quantifying parkinson's disease symptom severity," in *Recent Advances in Nonlinear Speech Processing (Vol. 48 of the series Smart Innovation, Systems and Technologies)*, vol. 48, pp. 93–102, 2016
[15] A. Gómez, A. Tsanas, P. Gómez, D. Palacios-Alonso, V. Rodellar, and A. Álvarez, "Acoustic to kinematic projection in Parkinson's disease dysarthria," *Biomed. Signal Process. Control*, vol. 66, p. e102422, 2021, doi: 10.1016/j.bspc.2021.102422.
[16] A. Tsanas and S. Arora, "Assessing Parkinson's disease speech signal generalization of clustering results across three countries: Findings in the Parkinson's voice initiative study," *BIOSIGNALS 2021 14th Int. Jt. Conf. Biomed. Eng. Syst. Technol. (BIOSTEC 2021)*, pp. 124–134, 2021
[17] A. Tsanas and S. Arora, "Large-scale clustering of people diagnosed with Parkinson's disease using acoustic analysis of sustained vowels: Findings in the Parkinson's voice initiative study," *BIOSIGNALS 2020 13th Int. Jt. Conf. Biomed. Eng. Syst. Technol. (BIOSTEC 2020)*, pp. 369–376, 2020
[18] I. R. Titze, *Principles of voice production*. Iowa City: National Center for Voice and Speech, 2000.
[19] A. Tsanas, "Accurate telemonitoring of Parkinson's disease using nonlinear speech signal processing and statistical machine learning," DPhil thesis, Oxford Centre for Industrial and Applied Mathematics, University of Oxford, 2012.
[20] 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 voice," *J. Acoust. Soc. Am.*, vol. 145, no. 5, pp. 2871–2884, 2019.
[21] A. Tsanas and S. Arora, "Biomedical speech signal insights from a large scale cohort across seven countries: The Parkinson's voice initiative study," in *Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, 2019, pp. 45–48.
[22] A. Tsanas and 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,* pp. 953-956, 2019
[23] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music.," *J. Acoust. Soc. Am.*, vol. 124, no. 3, pp. 1638–52, Sep. 2008
[24] A. Tsanas, M. Zañartu, M. A. Little, C. Fox, L. O. Ramig, and G. D. Clifford, "Robust fundamental frequency estimation in sustained vowels: detailed algorithmic comparisons and information fusion with adaptive Kalman filtering.," *J. Acoust. Soc. Am.*, vol. 135, no. 5, pp. 2885–901, 2014
[25] A. Tsanas, "Acoustic analysis toolkit for biomedical speech signal processing: concepts and algorithms," in *8th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, 2013, pp. 37–40.
[26] A. Tsanas and P. Gómez-Vilda, "Novel robust decision support tool assisting early diagnosis of pathological voices using acoustic analysis of sustained vowels," in *Multidisciplinary Conference of Users of Voice, Speech and Singing (JVHC 13)*, 2013, pp. 3–12.
[27] E. San Segundo, A. Tsanas, and P. Gomez-Vilda, "Euclidean Distances as measures of speaker similarity including identical twin pairs: A forensic investigation using source and filter voice characteristics," *Forensic Sci. Int.*, vol. 270, pp. 25–38, 2017

SESSION VIII
COVID-19

THE IMPACT OF ONLINE TEACHING DURING THE COVID-19 PANDEMIC ON VOCAL FATIGUE IN UNIVERSITY PROFESSORS: SELF-REPORTS AND ACOUSTIC EVALUATION

[1]K. V. Evgrafova, [2]N. S. Sokolova, [3]N. V. Shvalev

[1] Phonetics Department, Saint Petersburg State University, Saint Petersburg, Russia
[2] the Department of English Philology and Cultural Linguistics, Saint Petersburg State University, Saint Petersburg, Russia
[3]Saint Petersburg State Pediatric Medical University, St. Petersburg, Russia
[1]evgrafova@phonetics.pu.ru, [2]nssoko@yahoo.com, [3]dr-nix99@mail.ru

Abstract: **Due to the COVID-19 pandemic, there has been a dramatic change in the work conditions of all voice professionals. In 2020 university professors around the world had to shift to online teaching. The objective of this research is to analyse the impact of this new professional reality on the vocal load of Saint Petersburg university professors and possible risks of vocal fatigue increase due to online synchronous teaching. In this study the vocal fatigue is understood as a separate phenomenon caused by excessive professional voice load. It can result in auditory perceptual and acoustic changes in the voice signal and lead to serious pathological conditions. We followed the protocol used in our pre-pandemic vocal fatigue studies to make the results comparable. The acoustic evaluation and self-reports are presented.**

Keywords: **vocal fatigue, voice disorders, teacher's voice, voice load. online synchronous teaching, COVID-19 pandemic**

I. INTRODUCTION

Vocal fatigue in voice professionals (teachers, singers, guides etc.) has been a focus of research for decades, especially regarding its symptoms and risk factors. It is frequently self-reported as a sense of increased vocal effort and a sensation of laryngeal and pharyngeal constriction caused by excessive workload [2], [11]. There is a variety of vocal fatigue symptoms which are mainly explained by the physiologic mechanisms of vocal production. There exist many studies on vocal fatigue providing various concepts of the phenomenon. However, there is no universally accepted definition. Vocal fatigue is viewed either as a voice disorder caused by other pathological voice conditions or as a separate voice problem resulting from prolonged and excessive voice use [10]. In this study the vocal fatigue is understood as a separate phenomenon caused by excessive professional voice load. It can result in auditory perceptual and acoustic changes in the voice signal and can lead to serious pathological conditions. Identifying vocal fatigue in its initial stage is important to prevent voice disorders.

Acoustically, vocal fatigue is associated with changes in tonal range, dynamic range, vocal quality, intensity, fundamental frequency. Consequently, acoustic analysis is a good objective method to evaluate voice quality under fatigue. Besides, it causes perceivable changes in pitch, loudness, pauses, and voice quality.

We presented the acoustic, auditory and clinical analysis of vocal fatigue symptoms in the professors of Saint Petersburg university (pronunciation teachers and lecturers) in a number of previous studies. [5-7]

However, due to the COVID-19 pandemic, there has been a dramatic change in the work conditions of all voice professionals. Particularly, in 2020, university professors around the world had to shift to online teaching. In [1], [9], [10] the influence of the new experience on different types of teaching professionals is described.

The objective of this research is to analyse the impact of this new professional reality on Saint Petersburg university professors and possible risks of vocal fatigue increase due to online synchronous teaching.

II. METHODS

The methodologies used across numerous vocal fatigue studies can vary [1-10]. In most studies the vocal fatigue is induced artificially as a result of reading or speaking tasks of various types. [3], [8] The conditions of our pre-pandemic experiments seem to be more realistic. A total of 20 Saint Petersburg university professors were recorded before and after their workdays. The participants were asked to read at habitual loudness a four minute phonetically

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

representative text before classes in the morning. After continuous classroom teaching during the working day they were asked to record the same text. The recordings were used later for acoustic evaluation of the vocal fatigue symptoms. [5-7] The subjects also were asked to fill in a special questionnaire before each type of the recordings. In the questionnaire they evaluated their physical state, mood and a level of activity.

The given study also employed 20 professors of Saint Petersburg State University (with the average teaching experience of 7 years) who had shifted to online synchronous teaching in 2020 due to the pandemic. The participants were involved in different types of teaching activities:

1) teachers delivering lectures on linguistics (the Department of Phonetics and the Department of English Philology and Cultural Linguistics).
2) English teachers running practical classes (the Department of English Philology and Cultural Linguistics);
3) pronunciation coaches (the Department of Phonetics).

The minimum workload a day was 3 hours while the maximum was 6 hours.

No participant reported pathological voice problems.

Given that the most of the participants had taken part in the pre-pandemic vocal fatigue experiments, we followed the same protocol:

- to make it possible to compare the results (pre-pandemic vs. pandemic) in terms of self-assessment (subjective evaluation) and acoustic analysis (objective evaluation);
- to find out if shifting to online synchronous teaching due to the pandemic caused vocal fatigue to increase;
- to upgrade the guidelines concerning the optimal working-time regime and teacher's voice-use routine.

Whereas the recordings in the previous studies had been made in the studio at the Department of Phonetics with the use of Multichannel recording system Motu Traveler, in the presented experiments the participants were asked to record themselves *before* and *after* online *class/lecture* delivery using available devices such as mobile phones.

All the subjects were asked to fill in two types of questionnaires before each type of the recordings.

We used the WAM questionnaire to evaluate psychoemotional state of the teachers before and after their work.

WAM (*wellbeing* consisting of strength, fatigue and health, *activity* comprising mobility, speed of flow of functions and *mood* compiled by the characteristics of the emotional state) is often used to assess the mental state of patients and healthy people, their psychoemotional response to loading. [6] It is presented in the form of a scale with indices (3 2 1 0 1 2 3). The subject is offered 30 pairs of words with opposite meanings (*strong - weak, active – passive, happy – unhappy* etc.). The task is to select and circle 1 digit on each scale. The selected value should most accurately reflect the state of the person as it is at the time of the test.

Each of the scales has an average score of 4. When the score exceeds 4 points the state of well-being, activity, mood is defined as favorable. For normal state assessments, a range of 5.0-5.5 points is typical.

All the subjects were also asked to fill in a self-reporting questionnaire specially designed for the study before each type of the recordings.

In the questionnaire, they described their working conditions in detail and commented on any problems with their voice *before, during* or *after* the work load.

Table 1. A fragment of the self-reporting questionnaire showing the types of questions asked

Working conditions	**The self-perception of voice** *before, during* or *after* (yes/no)
location (*home, office, classroom*)	a high level of *muscular tension/discomfort*
a type of environment (*quiet, noisy*)	*hoarse* voice quality
a type of their workload (*a lecture, a seminar, a practical class, a pronunciation training*)	*breathy* voice quality
the amount of voice load *a day/a week*	*unsteady* voice
a type of an online *teaching platform or an application* they were using	inability to maintain typical *pitch*
the *absence/presence* of a headset	*dry* throat
quality of the internet connection (speedy/slow/stable/unstable)	*sore* throat
work experience (*less than 5 years/more than 5 years*)	throat *clearing*/frequent *pausing*

III. RESULTS

The *before* (non-fatigued voice) and *after* (fatigued voice) recordings were analysed for basic acoustic parameters. We calculated (in Praat) a number of acoustic parameters based on formant values, jittter, shimmer, pitch and loudness which can help detecting the absence/presence of voice fatigue in a given speech sample. The results obtained during pandemic studies (*online teaching*) as well as the pre-pandemic studies (*classsroom teaching*) [5-7] showed a consistent dependency between acoustic parameters and vocal fatigue. After a working day F0 values were higher, the duration of vowels increased; pitch and loudness range values increased. Measuring jitter and shimmer did not give consistent results. The analysis of F0 features shows that the mean pitch value tends to be higher in fatigued speech across all the subjects. The pitch range increases significantly due to the increase of upper range value. The mean lower range value stays practically unchanged.

The pre-pandemic analysis had showed that the main tendency was the increase in the mean value of F0 in the fatigued speech. The evaluation of pandemic recordings yielded similar results. However, as it is shown in Table 2, the mean duration of laryngalized speech segments is longer in the pandemic recordings. Laryngalization which is marked by significant decrease in pitch value and pitch breaks is associated with a *creaky* voice quality. The symptom was frequently reported by the teachers during the self-assessment of voice quality.

Table 2. The ratio of laryngalized speech segments to the whole text recorded (pre-pandemic vs. pandemic material)

Pre-pandemic material, %		**Pandemic material, %**	
Non-fatigued	1,5	***Non-fatigued***	1.8
Fatigued	1,2	***Fatigued***	2.3

Table 3. The increase in vowel duration in fatigued speech (pre-pandemic vs. pandemic material)

Vowel Duration Increase in fatigued speech (ms)	
Pre-pandemic material	4.3
Pandemic material	7.2

As it is shown table 2, the mean increase in vowel duration (all vowels) in fatigued speech is more significant in pandemic recordings.

The differences in the acoustic parameters before and after vocal loading mainly seem to reflect increased muscle activity as a consequence of excessive vocal loading.

The results of the WAM questionnaire according to Wellbeing scale in all phases of measurements *before* and *after* the workload exceeded 4 points, which indicated a favorable state of the teachers (Table 4). However, on average*, before* the work load swellbeing was rated at 5.5 points which is associated with a normal psychoemotional state (whereas the maximum is 7). The *after* self-assessment showed decreased wellbeing index, but it did not fall out of the range of 4.0 points.

The results of the WAM questionnaire according to the Activity scale in all phases of measurements *before* and *after* the classes also exceeded 4 points, which indicated a favorable state. The Mood rates were similar to the Activity ones.

Table 4. The mean rates of WAM test

Wellbeing	
Before	*After*
5.5 (min. 4.3 – max. 5.8)	4.3 (min. 4 – max.5.1)
Activity	
4.3 (min. 4.1 – max. 5.5)	5.4 (min. 4.1 – max. 6.1)
Mood	
5.0 (min. 4.3 – max. 5.2)	5.3 (min. 4.9 – max. 6.3)

A total of 20 participants indicated feelings of vocal fatigue, general tiredness and psychoemotional exhaustion at the end of a day full of *online* classes and lectures (up to 6 hours of teaching).

The analysis of the self-reports revealed symptoms of a high degree of vocal fatigue *during* and *after* the work load such as

- a high level of muscular tension/discomfort (due to the *microphone* effect),
- vocal fatigue and general tiredness
- hoarse voice quality
- creaky voice quality
- breathy voice quality
- unsteady voice
- inability to maintain typical pitch
- a dry or scratchy throat
- a sore throat
- dry cough
- muscle pain in the neck and the larynx (obviously caused by inadequate posture and continuous talking while sitting)
- psychological stress (caused by a lack of auditory and visual feedback or student interaction, technical problems and online connection failures).

The pronunciation teachers reported the largest number of vocal problems then practical class teachers as pronunciation training is the most challenging in terms of vocal effort.

IV. Discussion

The comparison of the pre-pandemic results and the current ones based on acoustical analysis and self-reporting questionnaires showed that the shift to delivering classes and lectures online caused substantial vocal fatigue increase. The main symptoms included hoarseness of voice, cracked or split voice, throat discomfort, neck and dry cough. However, the voice symptoms turned to be milder in the teachers using a headset which seems to be an effective way of adjusting to the new working conditions. However, in should be noted that wearing a headset continuously during the working day in some cases caused a headache and pain in the neck in a quiet big group of the subjects.

V. Conclusion

The concern should be raised over the significant increase in the focal fatigue in university professors in comparison with the pre-pandemic time.

The guidelines concerning teacher's voice-use routine should be developed by voice pathologists according to the new working conditions. They may include special sets of vocal exercises and strategies to avoid voice overstraining by slowing the pace, taking frequent pauses, putting an emphasis on diction and consonants rather on increasing the loudness.

The optimal working-time regime should be also reconsidered both for those delivering online classes and working in a hybrid regime. It especially concerns pronunciation teachers who seem to be particularly susceptible to vocal fatigue. They have to repeat articulation drills in front of the students many times and correct continuously their pronunciation which demands a high level of vocal effort and excessive muscular tension of articulators. As a consequence of this vocal overloading, the pronunciation teachers often suffer from dysphonia and benign lesions such as nodules.

REFERENCES

[1] A. Besser, S. Lotem, V. Zeigler-Hill Psychological Stress and Vocal Symptoms Among University Professors in Israel: Implications of the Shift to Online Synchronous Teaching During the COVID-19 Pandemic. Clinics (Sao Paulo). 2021; 76: e2641. Published online 2021 Mar 19. doi: 10.6061/clinics/2021/e2641

[2] V. J. Boucher, "Acoustic Correlates of Fatigue in Laryngeal Muscles: Findings for a Criterion-Based Prevention of Acquired Voice Pathologies", in Journal of Speech, Language, and Hearing Research, vol. 51, pp. 1161–1170. October 2008.

[3] M.J. Caraty and C. Montacié, "Multivariate Analysis of Vocal Fatigue in Continuous Reading, The Proceedings of Interspeech 2010, pp. 470-473.

[4] V. A. Doskin Test differentsirovannoy samootsenki funktsional'nogo sostoyania (The Test of Differentiated Self-Assessment of Functional Status)/ V.A. Doskin, N.A. Lavrenteva, M.P. Miroshnikov, V.B. Sharay// Voprosy psikhologii 1973 №6, pp. 141-145

[5] K. Evgrafova, V. Evdokimova, "Acoustic analysis of vocal fatigue in professional voice users". MAVEBA 2011*: 153-156.*

[6] K. Evgrafova, V. Evdokimova, P. Skrelin, T. Chukaeva "Vocal fatigue in voice professionals: collecting data and acoustic analysis". DOI: 10.36505/ExLing-2016/07/0011/000270

[7] V.V. Evdokimova, K.V. Evgrafova, P.A Skrelin, T.V. Chukaeva, "The database of normal and pathological singers'voices: an approach to collecting data." *MAVEBA 2017: 23-24.*

[8] B.E. Kostyk and A.P. Rochet, "Laryngeal airway resistance in teachers with vocal fatigue: a preliminary study", in Journal of Voice, 1998, vol. 12, pp. 287–299.

[9] K, Nemr, M, Simões-Zenari, V. Cássia de Almeida, G, A, Martins and I. T, Saito COVID-19 and the teacher's voice: self-perception and contributions of speech therapy to voice and communication during the pandemic. Clinics (Sao Paulo, Brazil), 76, e2641. https://doi.org/10.6061/clinics/2021/e2641

[10] M. Patjas, H. Vertanen-Greis, P. Pietarinen, A. Geneid «Voice symptoms in teachers during distance teaching: a survey during the COVID-19 pandemic in Finland.» Eur Arch Otorhinolaryngol. 2021 Jul 4:1-8. doi: 10.1007/s00405-021-06960-w. Online ahead of print. PMID: 34219183 Free PMC article. J Voice. 2020 Jun 5 doi: 10.1016/j.jvoice.2020.05.028 [Epub ahead of print]

[11] R.C. Scherer, I.R. Titze et al, "Vocal fatigue in a professional voice user", in "Transcripts of the Fourteenth Symposium: Care of the Professional Voice", New York: The Voice Foundation, 1986, pp.124–130.

EFFECT OF PROTECTIVE MASKS ON VOICE PARAMETERS: ACOUSTICAL ANALYSIS OF SUSTAINED VOWELS

C.Manfredi[1], V.Altamore[2], A.Bandini[3], S.Orlandi[4], L.Battilocchi[5], G.Cantarella[5]

[1]Department of Information Engineering, Università degli Studi di Firenze, Firenze, Italy
[2]School of Engineering, Università degli Studi di Firenze, Firenze, Italy
[3] Scuola Superiore Sant'Anna, The Biorobotics Institute, Pisa, Italy
[4]Dipartimento di Ingegneria dell'Energia Elettrica e dell'Informazione "Guglielmo Marconi", Bologna, Italy
[5]IRCCS Ca' Granda Foundation, Ospedale Maggiore Policlinico Milano, Milano, Italy
claudia.manfredi@unifi.it, virginia.altamore@stud.unifi.it, andreabandini87@gmail.com,
orlandisilvia85@gmail.com, ludovicabattilocchi@gmail.com, giovanna.cantarella@policlinico.mi.it

Abstract: **During the last two years the use of masks as personal protective equipment became necessary and mandatory to deal with the SARS-CoV-2 epidemiological emergency with impact on the quality and efficiency of verbal communication.**
This paper compares for the first time 7 different mask configurations. The sustained vowels /a/, /i/ and /u/ emitted by Italian speakers are considered. The purpose of this work is to evaluate whether the use of different types of masks, by themselves or worn together with a protective shield, may affect the acoustical parameters and thus voice quality. This is exploited by means of acoustical analysis performed with the BioVoice tool that estimates more than 20 parameters. For each vowel, the values of the fundamental frequency F0, the first two formant frequencies F1 and F2, jitter and noise are compared among the 7 configurations.
Preliminary results show that for the three vowels there are few statistically significant differences among masks when worn alone, while the presence of the shield has a relevant impact on the signal energy above 1 kHz. Further studies are ongoing to analyze vocalic sentences in order to detect possible influence of the masks on vowel articulation.
Keywords: **Face masks, SARS-CoV-2, acoustical analysis, BioVoice, F0, formants.**

I. INTRODUCTION

Most personal protective equipments (PPEs) have a relevant impact both on the quality and the intelligibility of the voice signal especially in the case of noisy environments or hearing impairments. Moreover the inter-personal mandatory distance often leads to raise the voice, increasing voice fatigue especially for professionals that have a high daily voice load. Therefore, in the last two years the scientific community has examined the influence of masks on vocal acoustic characteristics. [1-6]. This paper compares for the first time 7 different mask configurations. The study is preliminary and limited **to** sustained vowels /a/, /i/ and /u/ emitted by Italian speakers, but further work is ongoing to analyze vocalic sentences in order to detect possible influence of the masks on vowel articulation.

The main characteristics of vowels are the fundamental frequency F0 and the formant frequencies, that is, the resonant frequencies of the vocal tract. In particular, the first two formants, F1 and F2, are related to the position of the tongue: F1 is linked to height, while F2 is linked to the front-to-back movement. Formants position may vary according to age and gender, but is also related to the language under consideration. In particular, the Italian language comprises just 7 vocalic sounds and does not make a distinction between rounded and non-rounded vowels. Figure 1 shows the vowel trapezoid of the Italian language according to the International Phonetic Alphabet (IPA):

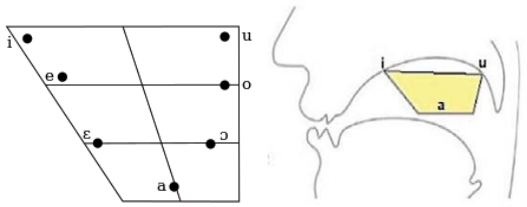


Figure 1 – Vowel trapezoid for the Italian language

In this work only the three vowels at the corners of the trapezoid are analyzed: /a/, /i/ and /u/. They roughly correspond to American English vowels /a/ ("father"), /I/ ("it") and /U/ ("foot") as reported in [7].

II. METHODS

Voice signals were recorded from 10 subjects (5 males and 5 females, mean age: 27,3, std= 1,494) of Italian mother tongue. Specifically, for females: mean=27,8; std= 0,836. For males: mean=26,8; std= 1,923.
Each subject was asked to emit the sustained vowels /a/, /i/ and /u/ at conversational amplitude for at least

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

5s. Seven configurations of different masks without or with protective shield were considered:
- No mask (Baseline)
- Surgical mask (Surgical)
- Ffp2 mask (Ffp2)
- Ffp3 mask (Ffp3)
- Surgical mask and visor (Surgical+shield)
- Ffp2 mask and visor (Ffp2+shield)
- Ffp3 mask and visor (Ffp3+shield)

Each vowel was repeated 3 times. Thus 210 recordings were collected.

Each recording was processed using the BioVoice tool that automatically distinguish among newborn cry, children, adults and singers voices, and in case of adults performs separate analyses for male and female voices [8, 9].

In this work, for each vowel the following acoustical parameters were considered: F0, F1, F2, jitter, and Adaptive Normalized Noise Energy (ANNE), along with their descriptive statistics (mean, median and standard deviation). A high-resolution method for F0 estimation is implemented in BioVoice, based on parametric AutoRegressive (AR) models applied to time-windows of varying length. AR models were also implemented to estimate the Power Spectral Density (PSD). PSD is automatically computed in the frequency range specific of each category, and normalized with respect to its maximum value, therefore the PSD range is 0dB downward. The first two PSD peaks correspond to F1 and F2.

ANNE relies on a comb filtering approach, optimised to deal with data windows of varying time duration. Large negative ANNE values correspond to good voice quality, while values close to zero reflect the presence of strong noise.

Concerning PSD, statistical analysis was implemented by dividing the frequency spectrum into 500 Hz intervals and calculating the average power over each interval for each mask configuration and each vowel, distinguishing between males and females. Also overall results (males and females) were considered and they are reported in this paper.

As data were not normally distributed, a non-parametric Friedman test was performed to find possible differences between the acoustical parameters obtained with the 7 configurations of face masks. In case of significant level of the Friedman test (p value < 0.05), a post-hoc multiple comparison was applied using the Dunn-Sidak method.

III. RESULTS

Results reported here concern male and female voices altogether. Separate analysis will be reported elsewhere.

A. F0, formant, jitter and noise

For all vowels and all configurations, F0 mean and median values are similar, both ranging between 168 Hz and 182 Hz, with a slight increase from the baseline to masks with shield. This could be related to an increasing effort in vocal emission due to protective equipments. However, no statistically significant differences were obtained for F0.

For the mean values of F1 and F2, the following ranges were found. Median values are not reported as they gave similar results.

Vowel /a/
F1: 720 Hz - 810 Hz. F2: 1080 Hz - 1180 Hz.
Vowel /i/
F1: 340 Hz - 370 Hz. F2: 1830 Hz - 2200 Hz.
Vowel /u/
F1: 400 Hz - 440 Hz. F2: 1010 Hz - 1120 Hz.

For all vowels jitter ranges between 0.50% and 1.4% and ANNE ranges between -23 dB and -27 dB.

The 6 configuration with masks were compared to the baseline with the Friedman test. Concerning jitter, no statistically significant difference was found for the three vowels.

Only the statistically significant results are reported here:

Vowel /a/
- F1 mean of FFP3 + shield was significantly lower than the baseline result.
- NNE of FFP2 + shield was significantly higher than the baseline result.
- NNE of FFP3 + shield was significantly higher than the baseline result.
- ANNE of FFP2 + shield and FFP3 + shield was significantly higher than the baseline result.

Vowel /i/
- F2 mean obtained with FFP2 + shield was significantly lower than the baseline.
- F2 median obtained with FFP2 + shield was significantly lower than the baseline.

Vowel /u/
No statistically significant differences were found with respect to the baseline.

B. Power Spectral Density

Figures 1-3 show the PSD trend in steps of 500 Hz for the three sustained vowels /a/, /i/ and /u/ and all the 7 configurations. The mean PSD of male and female values are presented, without differentiating between the two genders.

Baseline (no mask) is indicated with a solid line. Each dot corresponds to the mean value of the PSD values in each frequency step.

IV. DISCUSSION

Results show that F0 is only slightly influenced by masks and shield, while formants values exhibit statistically significant differences. This is especially true for F2 and for the vowel /a/. Indeed, F2 tends to be lowered by a back tongue constriction and raised by a front tongue constriction [7], therefore changes might be due to the presence of face mask and shield.

Also, higher ANNE values with shield with respect to the baseline might indicate a higher effort required in vowel emission. No statistically significant difference was found for jitter.

Concerning PSD: For vowel /a/ (Figure 1) the decrease in PSD is quite evident for the three configurations with shield already around 1kHz, where about 10dB of decrease is shown. Even larger differences are found from 2 kHz on for the same configurations. Less evident decrease is shown in Figures 2 and 3 that concern vowel /i/ and /u/ respectively.

It should be taken into account that these plots concern cumulative average values of men and women calculated over 500 Hz intervals, so they are somewhat different from the traditional power spectrum. In fact, when gender data were considered separately, a greater decrease and higher frequency values were observed for women. This might indicate a greater effort required to females with respect to males, possibly related to their shorter vocal tract and higher formant frequencies.

Furthermore, the vowel / u / is one of the most difficult to analyze through automatic tools, due to the position of its formants that depends on the position of the tongue and the conformation of the vocal tract which are very particular in this case.

Though preliminary, results show that masks alone have a negligible influence on the power spectral density (PSD) of sustained vowels /a/, /i/ and /u/, while the presence of the shield causes a relevant energy decreases above 1 kHz that is directly related to voice energy. This is particularly evident for vowel /a/ while vowels /i/ and /u/ show a less strong PSD decrease especially for frequencies below 2 kHz.

However, high standard deviation was found for all the configurations and vowels, baseline included. This might be related both to the mixture of male and female parameters used here and to the time of day when the recordings were made, i.e. at the end of the working day. Consequently, also the baseline parameters may have suffered from some distortion due to voice fatigue.

Work is ongoing to detect the influence of masks and shield on articulation in conversational voice and speech.

Though mandatory, the use of masks and shields might have negative impact especially in professions that make large use of voice such as teachers. These preliminary results suggest that some vocal exercise such as bubbling and face gym would be advisable at least for professionals [10].

V. CONCLUSION

This paper presents preliminary results on the impact of protective masks and shield on voice parameters estimated on the three basic sustained vowels of the Italian language /a/, /i/ and /u/. Recordings were made after a working day and concern 10 adult healthy subjects. Results confirm that voice energy decreases above 1 kHz especially when masks are worn along with the protective shield.

To the authors knowledge this is the first attempt to compare seven different masks configurations.

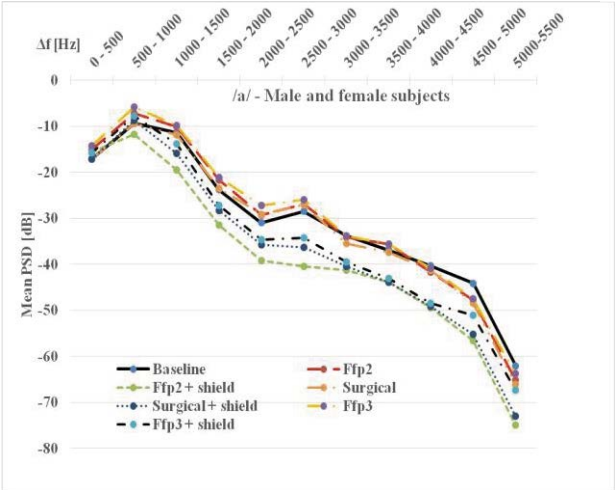


Figure 1 – /a/ vowel: dots correspond to the average PSD over 500Hz slices.

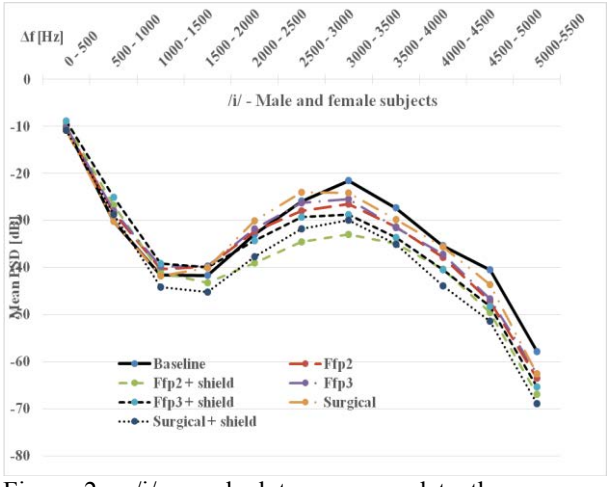


Figure 2 – /i/ vowel: dots correspond to the average PSD over 500Hz slices.

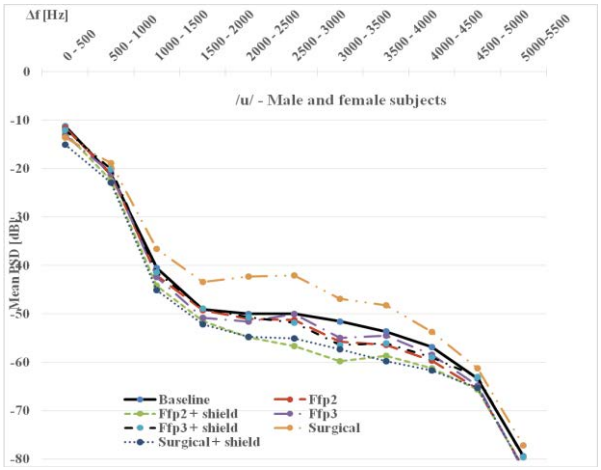


Figure 3 – /u/ vowel: dots correspond to the average PSD over 500Hz slices.

Future work will concern the analysis of vowels during articulation [11] and vocalic sentences. Moreover, self-perceptual evaluation based on a specific questionnaire will be performed.

In this work only non-invasive measures are considered, based on voice recordings and the acoustical analysis of the signal. More invasive analysis could be performed such as videokymography that might provide further helpful parameters [12].

REFERENCES

1. Pörschmann C, Lübeck T, Arend JM. Impact of face masks on voice radiation. J Acoust Soc Am. 2020 Dec;148(6):3663. doi: 10.1121/10.0002853. PMID: 33379881; PMCID: PMC7857507.
2. Goldin A, Weinstein BE, Shiman N. How do medical masks degrade speech perception? *Hearing Review*. 2020;27(5):8-9
3. Corey RM, Jones U, Singer AC. Acoustic effects of medical, cloth, and transparent face masks on speech signals. J Acoust Soc Am. 2020 Oct;148(4):2371. doi: 10.1121/10.0002279. PMID: 33138498; PMCID: PMC7857499.
4. Magee M, Lewis C, Noffs G, Reece H, Chan JCS, Zaga CJ, Paynter C, Birchall O, Rojas Azocar S, Ediriweera A, Kenyon K, Caverlé MW, Schultz BG, Vogel AP. Effects of face masks on acoustic analysis and speech perception: Implications for peri-pandemic protocols. J Acoust Soc Am. 2020 Dec;148(6):3562. doi: 10.1121/10.0002873. PMID: 33379897; PMCID: PMC7857500.
5. Cavallaro G, Di Nicola V, Quaranta N, Fiorella ML. Acoustic voice analysis in the COVID-19 era. Acta Otorhinolaryngol Ital. 2021 Feb;41(1):1-5. doi: 10.14639/0392-100X-N1002. Epub 2020 Nov 24. PMID: 33231205; PMCID: PMC7982755.
6. Magee M, Lewis C, Noffs G, Reece H, Chan JCS, Zaga CJ, Paynter C, Birchall O, Rojas Azocar S, Ediriweera A, Kenyon K, Caverlé MW, Schultz BG, Vogel AP. Effects of face masks on acoustic analysis and speech perception: Implications for peri-pandemic protocols. J Acoust Soc Am. 2020 Dec;148(6):3562. doi: 10.1121/10.0002873. PMID: 33379897; PMCID: PMC7857500.
7. Deller J R, Proakis J G, Hansen J H L. Discrete-Time Processing of Speech Signals**,** Macmillan Coll Div, 1993, ISBN 10: 0023283017ISBN 13: 9780023283017,
8. Morelli M. S., Orlandi S., Manfredi C. BioVoice: A multipurpose tool for voice analysis. Biomedical Signal Processing and Control 2021 64,102302 doi:10.1016/j.bspc.2020.102302
9. Manfredi C, Barbagallo D, Baracca G, Orlandi S, Bandini A, Dejonckere P. H. Automatic Assessment of Acoustic Parameters of the Singing Voice: Application to Professional Western Operatic and Jazz Singers, Journal of Voice, 2015, Vol.29(4), p.517.e1-517.e9 ISSN: 0892-1997, 1873-4588; DOI: 10.1016/j.jvoice.2014.09.014
10. Di Natale V, Cantarella G, Manfredi C, Ciabatta A., Bacherini C, DeJonckere P.H. Semioccluded Vocal Tract Exercises Improve Self-Perceived Voice Quality in Healthy Actors. Journal of Voice 2020. ISSN: 0892-1997,1873-4588; DOI: 10.1016/j.jvoice.2020.07.024
11. Bandini, A., Orlandi, S., Giovannelli, F., Felici, A., Cincotta, M., Clemente, D., Vanni P., Zaccara G., Manfredi, C. Markerless analysis of articulatory movements in patients with Parkinson's disease. Journal of Voice 2016, 30(6),766.e1-766.e11. doi:10.1016/j.jvoice.2015.10.014.
12. Piazza, C; Mangili, S, Del Bon, F, Gritti, F, Manfredi, C, Nicolai, P, Peretti, G.,Quantitative analysis of videokymography in normal and pathological vocal folds: a preliminary study European archives of oto-rhino-laryngology. , 2012, Vol.269(1), p.207-212 ISSN: 0937-4477 , 1434-4726; DOI: 10.1007/s00405-011-1780-y

ELIMINATION OF COMPLICATIONS OF TRACHEOESOPHAGAL BYPASSING WITH PROSTHETICS IN PATIENTS AFTER LARINGECTOMY DURING COVID-19 PANDEMIC.

E. N. Novozhilova[1,2], V. I. Popadyuk[1], A. I. Chernolev[1], N. V. Ermakova[1], I. V. Kastyro[1]
[1] Peoples' Friendship University of Russia (RUDN-University), Moscow, Russia
[2]Moscow City Oncological Hospital No. 62, Moscow Healthcare Department, Moscow, Russia
Lorval04@mail.com, acernolev@yandex.ru, n.v.ermakova@mail.ru, ikastyro@gmail.com

Abstract: **Laryngectomy for laryngeal cancer is the treatment of choice in these patients. Rehabilitation of patients with voice impairment is not an easy task. For the purpose of rehabilitation, tracheoesophageal bypass (TEB) is performed. When examining patients with TPS, medical personnel must be protected by personal protective equipment. Patients with PSI are at high risk for aspiration pneumonia. In the context of the COVID-19 pandemic, patients after laryngectomy with tracheoesophageal bypass surgery with prosthetics need to be given special attention. When infected with SARSCoV-2, these patients are at a special risk group. They need special conditions in the clinic - special care and rehabilitation.**
Keywords: **laryngectomy, rehabilitation, tracheoesophageal bypass, COVID-19, SARSCoV-2**

I. INTRODUCTION

At the end of 2019, an outbreak of a new coronavirus infection occurred in the People's Republic of China (PRC) with an epicenter in the city of Wuhan (Hubei province), the causative agent of which was given the temporary name 2019-nCoV.

In an analysis of 72,314 COVID-19 patients in China, the overall case fatality rate was 2.3%. However, for patients with severe concomitant pathology, it was equal to 7.3% (10.5% of patients with cardiovascular diseases, 7.3% - for patients with diabetes, 6.3% - with chronic respiratory diseases, 6.0% - for cancer patients) [1]

The main method of treating patients with tumors of the upper respiratory tract is usually surgery [1]. Laryngectomy for laryngeal cancer is the treatment of choice in these patients [2]. However, this type of surgery is disabling as patients lose their voice. Rehabilitation of patients with impaired voice function is not an easy problem [3-6]. For the purpose of rehabilitation, tracheoesophageal bypass surgery (TEB) is performed [7].

After laryngectomy, separation of the upper and lower respiratory tract occurs, a permanent tracheostomy is formed, and the entire biomechanism of respiration changes. Therefore, this category of patients is inevitably the most susceptible to the COVID-19 virus in a pandemic, in overcrowded hospitals, as well as in patients with severe comorbidities.

As a rule, these patients belong to the older age group, with a long history of smoking, severe manifestations of chronic obstructive pulmonary disease, and a high risk of infection against the background of mucociliary dysfunction.

Given the presence of a high viral load in the upper respiratory tract, all ENT procedures are high-risk procedures, and otorhinolaryngologists are at risk for COVID-19 infection.

All of these patients have a high risk of postoperative wound complications, as well as the risk of contracting the COVID-19 coronavirus. In case of infection, the patient himself becomes a source of transmission. Aerosol viral particles can infect surrounding health care workers and other patients, especially during airway sanitation.

II. TEB PATIENTS & COVID-19

When performing TEB with prosthetics after laryngectomy, a number of complications are possible associated with the displacement of the prosthesis and / or its course [8]. Usually, these problems can be corrected on an outpatient basis. But in the context of coronavirus infection and with an increased risk of SARS-COV-2, the patient and staff should be as safe as possible. Optimally, if in the examination room, forced ventilation with negative pressure and HEPA-filters are installed, which minimizes the risk of infection transmission [9].In patients after laryngectomy, there is no nasal breathing and untreated air through the tracheostomy directly enters the respiratory tract, which, as a rule, is accompanied by severe cough. At the same time, aerosol transmission of viral particles can significantly increase, compared with an ordinary person, when the protective function of the nose is preserved [5, 9]. So, during the outbreak of the SARS epidemic in 2003, a significant concentration of viral particles was determined in

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

tracheal aspirates. Therefore, the issue of care and contact with the patient after laryngectomy, both in inpatient and at home, is extremely important. Based on this, we recommend that any patient after laryngectomy be considered as potentially dangerous and infected with COVID-19.

We recommend a standard set of personal protective equipment for staff in contact with COVID patients to prevent infection of medical personnel when examining all patients after laryngeal surgery. It should be noted that the use of respirator No. 95 and a protective screen for the face in 100% of cases effectively protects the employee from infection [3].

In the case when an in-person consultation is absolutely necessary (examination after surgery, complications, suspicion of a relapse of the disease), it is important to "screen" these patients even before visiting the clinic. It is advisable to take a thorough history and conduct an examination for COVID-19.

It is important to note that a patient with a tracheostomy must use a respiratory heat exchanger with a viral-bacterial hygroscopic filter and cover the tracheostomy with a mask, scarf or clothing during a visit to the clinic [11].

III. TEB COMPLICATIONS

If the patient has a leak around the prosthesis, there is a risk of developing aspiration pneumonia, which can even have lethal consequences for the patient in the context of COVID-19. In the case of displacement of the prosthesis towards the trachea or esophagus, this can be diagnosed by X-ray, as well as using gastro- or tracheoscopy. It is advisable to start the study with standard X-ray images, and, if necessary, perform computed tomography (CT). Aspiration of the prosthesis into the airway is an absolute indication for urgent endoscopic intervention (regardless of the patient's COVID-19 status). It is prudent to treat all such patients as potentially infectious and to take all precautions to minimize the transmission of aerosol particles. When transporting to the operating room, it is necessary to cover the tracheostomy with a napkin, mask. Any attempt to use a filter or trachea tube in such a situation can further aggravate the cough and worsen the patient's condition.

When the patient's condition is stabilized, it is necessary to eliminate the complication as quickly as possible and, if possible, test for COVID-19. If there is a leak through the prosthesis, the patient should try to cope on his own at home. There are special plugs for the prosthesis ("like a key to a lock"), with which it is possible to block the lumen of the prosthesis. The flow will stop immediately, but the patient will not be able to talk (aphonia will occur). The patient may be advised to eat thicker food, which can also reduce aspiration.

If the voice prosthesis has completely fallen out, then at home the patient can temporarily insert a rubber catheter or a special dilator into the shunt in order to stop aspiration (the patient should be taught these procedures in advance or informed about the possibility of their own implementation). After that, the patient, in a stable condition and in safety, can already see a doctor on an outpatient basis.

In the clinic, the patient should be tested for COVID-19. Before receiving the test results, it is better to let the patient go home, and in case of a negative result, after 48 hours, invite again and replace the prosthesis.

If the test for COVID-19 is positive, then such a patient should stay at home as long as possible and undergo special antiviral treatment. Only after complete recovery from infection is it recommended to carry out procedures for replacing the prosthesis. When working with COVID-positive patients, all staff and all procedures are advised to wear a PARP respirator. If this is not possible, then use at least a respirator No. 95 and personal protective equipment (dressing gown, glasses, shoe covers).

V. CONCLUSION

In the context of the COVID-19 pandemic, patients after laryngectomy with tracheoesophageal bypass surgery with prosthetics need to be given special attention. When infected with SARSCoV-2, these patients are at a special risk group. They need special conditions in the clinic - special care and rehabilitation.

REFERENCES

[1] Wu Z McGoogan J. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA*. 2020, vol. 323, no. 13, pp. 1239-1242.

[2] Demyashkin G.A., Kastyro I.V., Sidorin A.V., Borisov Y.S. The specific immunophenotypic features of nasopharyngeal carcinoma. *Vestn Otorinolaringol.* 2018, vol. 83, no. 5, 40-44.

[3] Ganina Ch.A., Makhonin A.A., Vladimirova T.Yu., Chemidronov S.N., Ghukasyan I.M. Supracricoid partial laryngectomy for advanced laryngeal cancer. *Head and neck. Russian Journal*. 2021, vol. 9, no. 2, pp. 78–84.

Alieva S.B., Azizyan R.I., Mudunov A.M., Zaderenko I.A., Daykhes N.A., Dobrokhotova V.Z., Novozhilova

E.N., Reshulskiy S.S., Borisova T.N., Vinogradov V.V. Principles of radiotherapy for laryngeal cancer. *Opuholi Golovy i Sei*. 2021, vol. 11, no. 1, pp. 24 – 33.

[4] Kovalenko A.N., Kastyro I.V., Reshetov I.V., Popadyuk V.I. Study of the Role of Hearing Aid on the Area of the Acoustic Field of Vowels. *Doklady Biochemistry and biophysics.* 2021, vol. 497, no. 1, pp. 108–111.

[5] Popadyuk V.I., Novozhilova E.N., Fedotov A.P., Chernolev A.I., Korshunova I.A., Olyshanskaya O.V., Bitsaeva A.V. A rare observation of amyloidosis of the larynx simulating a tumor. *Vestnik Otorinolaringologii*. 2019, vol. 84, no. 3, pp. 65-67.

[6] Kastyro I.V., Kovalenko A.N., Torshin V.I., Doroginskaya E.S., Kamanina N.A. Changes to voice production caused by long-term hearing loss (HL). *Models and Analysis of Vocal Emissions for Biomedical Applications - 11th International Workshop, MAVEBA 2019*. 2019, pp. 241–244.

[7] Kryukov A.I., I.V. Reshetov, Kozhanov L.G., Sdvizhkov A.M., Kozhanov A.L. The systemic approach to the rehabilitation of the patients presenting with laryngeal cancer after the resection of the organ and laryngectomy with tracheoesophageal by-pass and endoprosthetics. *Vestn Otorinolaringol*. 2016, vol. 81, no. 4, pp. 54-59.

[8] Reshetov I.V., Fatyanova A.S., Ignatyeva M.A. Second breath: the use of heat and moisture exchangers for pulmonary rehabilitation of tracheostomized patients. Head and neck Russian Journal. 2020, vol. 8, no. 2, pp. 86–94..

[9] Popadyuk V.I., Novozhilova E.N., Chernolev A.I., Kastyro I.V., Antoniv V.F. Features of management of patients after laryngectomy during Pandemic COVID-19. *Head and neck. Russian Journal*. 2020, vol. 8, no. 4, pp. 86–91

TREATMENT OF PATIENTS AFTER LARINGECTOMY WITH COVID-19 VIRUS IN A HOSPITAL.

V.I. Popadyuk [1], E. N. Novozhilova [1,2], A. I. Chernolev[1], M. G. Kostyaeva[1], I. Z. Eremina[1], I. V. Kastyro[1]

[1] Peoples' Friendship University of Russia (RUDN-University), Moscow, Russia

[2]Moscow City Oncological Hospital No. 62, Moscow Healthcare Department, Moscow, Russia

Lorval04@mail.com, acernolev@yandex.ru, kostyaeva.71@mail.ru, irina.z.eremina@gmail.com, ikastyro@gmail.com

Abstract: **The coronavirus pandemic is spreading rapidly around the world. The health systems of all countries faced extraordinary problems in terms of the creation and distribution of medical resources, including the re-equipment and creation of new hospital beds, and the provision of personal protective equipment. The patients who undergo a laryngectomy are a special category. Given the fact that during the operation they have a separation of the upper and lower respiratory tract, in the context of the COVID-19 pandemic, such patients require special attention from oncologists and otorhinolaryngologists. Purpose of the study is to review the characteristics of patient management after a laryngectomy in a COVID-19 pandemic. Laryngectomy patients represent a unique contingent in conditions of coronavirus infection SARS-COV-2, it is advisable to focus on providing them with protective equipment. This will significantly reduce the risk of infection with their virus, which can be a deadly threat to them. Infected patients during an epidemic represent a potential source of infection for medical personnel, which requires special protective measures. All procedures associated with the replacement of the prosthesis, endoscopic manipulations, it is advisable to postpone until the normalization of the epidemiological situation. If carrying out such operations is vital, then they should be carried out, observing all necessary precautions for both the patient and medical personnel.**

Keywords: **coronavirus pandemic, COVID-19, laryngectomy**

I. INTRODUCTION

COVID-19 is caused by the SARS-CoV-2 (Severe acute respiratory syndrome-related coronavirus 2) coronavirus, which is genetically related to the SARS family and the Middle East Respiratory Syndrome (MERS) virus and is a recombinant virus between bat coronavirus and an unknown coronavirus. The genetic sequence of SARSCoV-2 is similar to the SARS-CoV sequence by at least 79% [1, 2].

In the last two years, the SARS-CoV-2 (Severe acute respiratory syndrome-related coronavirus 2) pandemic has been taking place. The transmission of infection is carried out by airborne droplets, airborne dust and contact routes [3]. The leading route of transmission of SARS-CoV-2 is airborne, which is realized when coughing, sneezing and talking at a close (less than 2 meters) distance. The contact route of transmission is carried out during handshakes and other types of direct contact with an infected person, as well as through food, surfaces and objects contaminated with the virus.

II. FEATURES OF EXAMINATION OF PATIENTS AFTER LARYNGECTOMY UNDER COVID-19 EPIDEMIC

The defeat of the pharynx and larynx by a tumor process leads to disabling consequences [4-6]. Moreover, the rehabilitation of such patients is an extremely difficult process [7-10].

Given the presence of a high viral load in the upper respiratory tract, all ENT procedures are high-risk procedures, and otorhinolaryngologists are at risk for COVID-19 infection.

The most common symptoms of coronavirus infection are cough (dry or with little sputum) in 80% of cases; shortness of breath (55%); fatigue (44%); a feeling of congestion in the chest (> 20%).

Testing for COVID-19 is most often done by swabbing the oropharynx and nasopharynx. But given that the breathing of patients after laryngectomy is carried out through a tracheostomy, it is advisable to consider testing for SARS-COV-2 by detecting the virus in tracheal aspirates and from the nasal cavity, which is consistent with the WHO recommendations.

Any diagnostic and therapeutic procedures in the upper respiratory and digestive tracts, as a rule, cause coughing and should be considered as potentially dangerous in terms of aerosol transmission for medical personnel [3]. To limit the transmission of COVID-19 and to maximize the safety of medical personnel, it is shown to use personal protective equipment (PPE),

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,
 published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

and, if possible, even to cancel or postpone a dangerous procedure (AAO-HNS).

Taking into account the peculiarities of the anatomy, as well as the volume of laryngectomy [3, 5, 9], which is associated with the formation of a tracheoesophageal fistula and voice prosthetics, all medical recommendations should be organized in such a way as to minimize the possibility of transmission of the SARS-COV-2 virus from the patient to the medical staff. In this case, the use of personal protective equipment is relevant.

In patients after laryngectomy, there is no nasal breathing and untreated air through the tracheostomy directly enters the respiratory tract, which, as a rule, is accompanied by severe cough. At the same time, aerosol transmission of viral particles can significantly increase, compared with an ordinary person, when the protective function of the nose is preserved [5, 9]. So, during the outbreak of the SARS epidemic in 2003, a significant concentration of viral particles was determined in tracheal aspirates. Therefore, the issue of care and contact with the patient after laryngectomy, both in inpatient and at home, is extremely important. Based on this, we recommend that any patient after laryngectomy be considered as potentially dangerous and infected with COVID-19.

We recommend a standard set of personal protective equipment for staff in contact with COVID patients to prevent infection of medical personnel when examining all patients after laryngeal surgery. It should be noted that the use of respirator No. 95 and a protective screen for the face in 100% of cases effectively protects the employee from infection [3].

In the case when an in-person consultation is absolutely necessary (examination after surgery, complications, suspicion of a relapse of the disease), it is important to "screen" these patients even before visiting the clinic. It is advisable to take a thorough history and conduct an examination for COVID-19.

It is important to note that a patient with a tracheostomy must use a respiratory heat exchanger with a viral-bacterial hygroscopic filter and cover the tracheostomy with a mask, scarf or clothing during a visit to the clinic [11].

III. TREATMENT OF PATIENTS WITH THE COVID-19 VIRUS IN A HOSPITAL

When a patient is admitted to a hospital and planning treatment, it is extremely important that all medical workers of the department understand the surgical anatomy of the airways in a patient after laryngectomy. The attention of the personnel should be emphasized that the use of oxygen masks and nasal catheters in such a patient will be useless, since the upper respiratory tract is "turned off" from breathing as a result of the operation, and oxygenation occurs only through the tracheostomy. Under ideal conditions, it is advisable to test all incoming patients for COVID-19.

However, if testing is not possible, all patients should be treated as potentially infected and all feasible remedies should be used. It is extremely important for patients to use heat exchangers with viral or bacterial filters attached to the tracheostomy area.

In case of severe coughing and profuse sputum secretion, special tracheotubes with powerful HEPA-filters can be used. And such a patient can be placed in a room with negative pressure and / or a closed ventilation system in order to prevent the spread of viral particles to other rooms. In some cases, it is advisable to use mechanical ventilation in auxiliary modes in order to provide a closed breathing circuit for the patient (even if his oxygenation does not suffer greatly). It is also important to use mechanical barriers over the tracheostomy (transparent blocks with holes for the doctor's hands), which is especially important at the time of intubation and extubation, when caring for the tracheotomy tube. The main thing in this situation is to prevent the spread of aerosol particles of the virus by any possible means.

Each patient after laryngectomy in the ward should have an individual suction, which the patient should be trained to use even before the operation. When caring for such patients, strict use of PPE is necessary, at least until negative tests for COVID-19 are obtained.

In case of a negative COVID-19 status for patients, it is still recommended to use HME with viral and bacterial filters from the very first hours after the operation, as well as wear a mask on the face and neck (which will provide a mechanical obstacle to the spread of the virus).

The patient should be explained that it is not necessary to touch the tracheostomy unnecessarily, and after all hygiene measures have been taken, hands should be thoroughly washed. Caring for the skin around the tracheostomy is very important to reduce airway contamination.

After laryngectomy, self-contamination (contamination with viral particles of one's own airways) is also possible during the use of a voice prosthesis and when closing the tracheostomy with a finger, therefore it is so important to focus the patient's attention on frequent hand washing. During an epidemic, the use of HANDS-FREE systems becomes extremely relevant, which allow the patient after laryngectomy not to touch the tracheostomy with a finger at all during speech load.

V. CONCLUSION

Considering the fact that patients after laryngectomy are a unique contingent in conditions of SARS-COV-2 coronavirus infection, it is advisable to focus on providing them with protective equipment (filters and heat exchangers). This will significantly reduce their risk of contracting the virus, which could pose a lethal threat to them.

In addition, already infected patients themselves during an epidemic represent a potential source of infection for medical personnel, which requires the use of special protective measures.

It is advisable to postpone all procedures related to the replacement of the prosthesis, endoscopic manipulations until the epidemiological situation normalizes. If the conduct of such operations is vital, then they should be carried out, observing all the necessary precautions for both the patient and the medical staff..

REFERENCES

[1] Zhou P., Yang X.L., Wang X.G. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020, vol. 3, iss. 579, no. 7798, pp. 270-273.

[2] Gorbalenya A.E., Baker S.C., Baric R.S., de Groot R.J. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol*. 2020, vol. 5, pp. 536–544.

[3] Popadyuk V.I., Novozhilova E.N., Chernolev A.I., Kastyro I.V., Antoniv V.F. Features of management of patients after laryngectomy during Pandemic COVID-19. *Head and neck. Russian Journal*. 2020, vol. 8, no.4, pp. 86–91

[4] Demyashkin G.A., Kastyro I.V., Sidorin A.V., Borisov Y.S. The specific immunophenotypic features of nasopharyngeal carcinoma. *Vestn Otorinolaringol*. 2018, vol. 83, no. 5, pp. 40-44.

[5] Ganina Ch.A., Makhonin A.A., Vladimirova T.Yu., Chemidronov S.N., Ghukasyan I.M. Supracricoid partial laryngectomy for advanced laryngeal cancer. *Head and neck. Russian Journal*. 2021, vol. 9, no. 2, pp. 78–84.

[6] Popadyuk V.I., Novozhilova E.N., Chernolev A.I., Kastyro I.V., Antoniv V.F. Features of management of patients after laryngectomy during Pandemic COVID-19. *Head and neck. Russian Journal*. 2020, vol. 8, no. 4, pp. 86–91

[7] Alieva S.B., Azizyan R.I., Mudunov A.M., Zaderenko I.A., Daykhes N.A., Dobrokhotova V.Z., Novozhilova E.N., Reshulskiy S.S., Borisova T.N., Vinogradov V.V. Principles of radiotherapy for laryngeal cancer. *Opuholi Golovy i Sei*. 2021, vol. 11, no. 1, pp. 24 – 33.

[8] Kovalenko A.N., Kastyro I.V., Reshetov I.V., Popadyuk V.I. Study of the Role of Hearing Aid on the Area of the Acoustic Field of Vowels. *Doklady Biochemistry and biophysics*. 2021, vol. 497, no. 1, pp. 108–111.

[9] Popadyuk V.I., Novozhilova E.N., Fedotov A.P., Chernolev A.I., Korshunova I.A., Olyshanskaya O.V., Bitsaeva A.V. A rare observation of amyloidosis of the larynx simulating a tumor. *Vestnik Otorinolaringologii*. 2019, vol. 84, no. 3, pp. 65-67.

[10] Kastyro I.V., Kovalenko A.N., Torshin V.I., Doroginskaya E.S., Kamanina N.A. Changes to voice production caused by long-term hearing loss (HL). *Models and Analysis of Vocal Emissions for Biomedical Applications - 11th International Workshop, MAVEBA 2019*. 2019, pp. 241–244.

[11] Reshetov I.V., Fatyanova A.S., Ignatyeva M.A. Second breath: the use of heat and moisture exchangers for pulmonary rehabilitation of tracheostomized patients. *Head and neck Russian Journal*. 2020, vol. 8, no. 2, pp. 86–94.

INDEX OF AUTHORS

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021*,

published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-449-6 (PDF), DOI 10.36253/978-88-5518-449-6

ISSN 2704-601X (print)
ISSN 2704-5846 (online)
ISBN 978-88-5518-448-9 (Print)
ISBN 978-88-5518-449-6 (PDF)
ISBN 978-88-5518-450-2 (XML)
DOI 10.36253/978-88-5518-449-6
www.fupress.com