



UNIVERSITÀ  
DEGLI STUDI  
FIRENZE

**DINFO**  
DIPARTIMENTO DI  
INGEGNERIA  
DELL'INFORMAZIONE

13th  
INTERNATIONAL  
WORKSHOP

MODELS AND  
ANALYSIS  
OF VOCAL  
EMISSIONS  
FOR  
BIOMEDICAL  
APPLICATIONS

September 12-13, 2023  
Firenze, Italy



# PROCEEDINGS

  
FIRENZE  
UNIVERSITY  
PRESS

PROCEEDINGS E REPORT

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



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

**13TH INTERNATIONAL WORKSHOP**

**September 12-13, 2023  
Firenze, Italy**

**Edited by  
Claudia Manfredi**

Firenze University Press  
2023



Models and Analysis of Vocal Emissions for Biomedical Applications : 13th International Workshop, September, 12-13, 2023 / edited by Claudia Manfredii. – Firenze : Firenze University Press, 2023. (Proceedings e report ; 136)

<https://books.fupress.com/isbn/9791221501469>

ISSN 2704-601X (print)

ISSN 2704-5846 (online)

ISBN 979-12-215-0145-2 (Print)

ISBN 979-12-215-0146-9 (PDF)

ISBN 979-12-215-0147-6 (XML)

DOI 10.36253/979-12-215-0146-9

Cover: designed by CdC, Firenze, Italy.

#### *Peer Review Policy*

Peer-review is the cornerstone of the scientific evaluation of a book. All FUP's publications undergo a peer-review process by external experts under the responsibility of the Editorial Board and the Scientific Boards of each series (DOI 10.36253/fup\_best\_practice.3).


#### *Referee List*

In order to strengthen the network of researchers supporting FUP's evaluation process, and to recognise the valuable contribution of referees, a Referee List is published and constantly updated on FUP's website (DOI 10.36253/fup\_referee\_list).

#### *Firenze University Press Editorial Board*

M. Garzaniti (Editor-in-Chief), M.E. Alberti, F. Vittorio Arrigoni, E. Castellani, F. Ciampi, D. D'Andrea, A. Dolfi, R. Ferrise, A. Lambertini, R. Lanfredini, D. Lippi, G. Mari, A. Mariani, P.M. Mariano, S. Marinai, R. Minuti, P. Nanni, A. Orlandi, I. Palchetti, A. Perulli, G. Pratesi, S. Scaramuzzi, I. Stolzi.

*FUP Best Practice in Scholarly Publishing* (DOI 10.36253/fup\_best\_practice)

 The online digital edition is published in Open Access on [www.fupress.com](http://www.fupress.com).

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

Metadata license: all the metadata are released under the Public Domain Dedication license (CC0 1.0 Universal: <https://creativecommons.org/publicdomain/zero/1.0/legalcode>).

© 2023 Author(s)

Published by Firenze University Press  
Firenze University Press  
Università degli Studi di Firenze  
via Cittadella, 7, 50144 Firenze, Italy  
[www.fupress.com](http://www.fupress.com)

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



# MAVEBA 2023

Firenze, Italy

The MAVEBA 2023 Workshop is sponsored by:



**Bioengineering | An Open Access Journal from MDPI**  
<https://www.mdpi.com/journal/bioengineering>



UNIVERSITÀ  
DEGLI STUDI  
FIRENZE  
**DINFO**  
DIPARTIMENTO DI  
INGEGNERIA  
DELL'INFORMAZIONE

**Dipartimento di Ingegneria dell'Informazione (DINFO),  
Università degli Studi di Firenze**  
<https://www.dinfo.unifi.it/>



UNIVERSITÀ  
DEGLI STUDI  
FIRENZE

**Università degli Studi di Firenze (UNIFI)**  
<https://www.unifi.it/>



**La Voce Artistica**  
<https://www.voceartistica.it/>



FONDAZIONE  
CR FIRENZE

**Fondazione CR Firenze**  
<https://fondazionecrfirenze.it>



**World Voice Day**  
<http://world-voice-day.org/>



# CONTENTS

Foreword .....	XI
----------------	----

## **SESSION I – BIOMECHANICS AND MODELS**

ELECTROMYOGRAPHIC ANALYSIS OF LIP AND FACE MUSCLES IN BEATBOXING .....	15
N. Henrich Bernardoni, J. Frère, A. Paroni, S. Gerber, H. Løevenbruck	

ON THE USE OF BIOMECHANICAL MODELS FOR VOCAL FOLD EDGE TRACKING IN VIDEOKYMOGRAMS .....	19
C. Drioli, G. L. Foresti	

AN ACOUSTIC ANALYSIS OF VOWELS TO PREDICT VOICE CHANGES IN A LONG READING TASK .....	23
M. Haggmuller, J. Linke, S. Lohrmann, F. Pokorny, B. Schuppler	

## **SESSION II – VOCAL FOLDS AND VOCAL TRACT**

FORMANT ANALYSIS OF VOICE AS AN EARLY BIOMARKER OF NEURODEGENERATION .....	29
A. Nacci, S. Capobianco, F. Simoni, L. Bruschini, S. Berrettini, L. Bastiani	

VOCAL FOLD IMPACT STRESS: THE KEY CONCEPT OF PHONOTRAUMA.....	33
P. H. DeJonckere, J. Lebacq	

FORMANT TRAJECTORIES IN DIFFERENT LANGUAGES .....	37
V. V. Evdokimova, M. R. Maximova	

## **SESSION III – SINGING, DRAMA AND VOICE QUALITY**

THE ROLE OF PHONIATRICES IN THE DRAMA THEATRE ACADEMIES: CLINICAL OBSERVATIONS .....	43
G. Baracca, S. Capobianco, L. Bastiani, A. Nacci	

STUDIO REPORT: RESEARCH ON THE ACOUSTICS AND CREATIVE TECHNOLOGIES OF THE SINGING VOICE IN LABMAT (LABORATORY OF MUSIC ACOUSTICS AND TECHNOLOGY, DEPARTMENT OF MUSIC STUDIES, NKUA) .....	47
A. Georgaki, A. Andreopoulou	

ON THE TRANSMISSIBILITY OF SPECTRAL DECAY RATE VOICE QUALITY PARAMETERS TO CONSONANT VOICING .....	51
W. Wokurek, M. Pützer	

## **SESSION IV – TOOLS AND METHODS FOR SPEECH AND VOICE ANALYSIS**

INFANT CRY FOR PATHOLOGIES CLASSIFICATION USING A DEEP LEARNING APPROACH ...	57
C. A. Reyes-Garcia, I. A. Valencia-Hernandez, O. F Reyes-Galaviz	

THE IMPACT OF DIFFERENT TYPES OF TEACHING MODES ON VOCAL FATIGUE IN UNIVERSITY TEACHERS

K.V. Evgrafova, N. S. Sokolova, N. V. Shvalev .....61

PERFORMANCE OF UNIVERSAL-PLATFORM-BASED VOICESCREEN APPLICATION IN AVQI MEASUREMENTS

V. Uloza, N. Ulozaitė-Stanienė, K. Pribuišis, T. Blažauskas, R. Damaševičius, R. Maskeliūnas .....65

DOMAIN ADVERSARIAL CONVOLUTIONAL NEURAL NETWORK FOR PARKINSON'S DISEASE DETECTION FROM SPEECH

E. J. Ibarra-Sulbaran, J. D. Arias-Londoño, M. Zañartu, J. I. Godino-Llorente .....69

**SESSION V – STUDENT COMPETITION**

CUMULATIVE PAIR-WISE VOWEL DISTANCE (CPVD): NEW VOWEL SPACE METRICS FOR PEOPLE WITH ATYPICAL SPEECH.....75

T. Cao, A. Favaro, T. Thebaud, J. Villalba, P. Zelasko, E. S. Oh, A. Butala, N. Dehak, L. Moro-Velazquez

EXAMINING THE DIRECTIVITY CHARACTERISTICS OF GREEK SUNG VOWELS ON FORMANT FREQUENCIES

G. Dedousis, K. Bakogiannis, A. Andreopoulou, A. Georgaki .....79

AI TECHNIQUES APPLIED TO ACOUSTICAL FEATURES OF PARALYTIC DYSPHONIA VERSUS DYSPHONIA DUE TO BENIGN VOCAL FOLD MASSES

F. Calà, L. Frassinetti, G. Cantarella, L. Battilocchi, G. Buccichini, A. Lanatà, C. Manfredi .....83

PERFORMANCE EVALUATION OF 3D NEURAL NETWORKS APPLIED TO HIGH-SPEED VIDEOS FOR GLOTTIS SEGMENTATION IN DIFFICULT CASES .....87

A. A. Dadras, P. Aichinger

HERMESPEECH RECORDER: A NEW OPEN-SOURCE WEB PLATFORM TO RECORD SPEECH TO THE CLOUD .....91

J. Park, M. Zinkus, J. Huang, A. Butala, J. Zhang, L. Clawson, S. Cust, V. Chovaz, N. Dehak, H. Wang, L. Moro- Velazquez

OPERATIC SINGING MULTI-SENSOR RECORDING PROTOTYPE: A PILOT EVALUATION STUDY .....95

E. Angelakis, K. Bakogiannis, A. Andreopoulou, M. Habela, A. Georgaki

ELECTRODERMAL ACTIVITY AND ACOUSTICAL ANALYSIS IN A WORDS/NON-WORDS READING TASK .....99

F. Calà, L. Frassinetti, P. Tarchi, V. Guarguagli, C. Manfredi, A. Lanatà

**ROUND TABLES, LABORATORY AND LECTURE**

ROUND TABLE I: ACOUSTIC AND PHYSIOLOGICAL ASPECTS OF SINGING .....105

S. Capobianco, N. Henrich Bernardoni, M. Kob (Moderator: J. Sundberg)

ROUND TABLE II: FOCUSING ON VOICE ONSET: A CRITICAL MOMENT OF PHONATION. FROM BASIC SCIENCE TO THERAPY .....	107
J. Sundberg, G. Cantarella, M. Kob, P. Aichinger (Moderator: P. H. DeJonckere)	
LECTURE: THE « VIRTUAL MUSEUM OF PHONIATRICS » A GUIDED TOUR BY THE CURATOR-IN-CHIEF .....	109
P. H. DeJonckere	
LABORATORY: A MULTIMODAL SYSTEM FOR THE CHARACTERIZATION OF PHYSIOLOGICAL DYNAMICS DURING SIFEL PROTOCOL AND WORDS/NON-WORDS READING TASKS .....	111
L. Frassinetti, F. Calà, P. Tarchi, V. Guarguagli, C. Manfredi, A. Lanatà	
INDEX OF AUTHORS .....	113





**MAVEBA**  
**2023**  
Firenze, Italy

## **FOREWORD**

This book of Proceedings includes the contributions presented at the 13<sup>th</sup> International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications – MAVEBA 2023, held in Firenze from 12 to 13 September, 2023. The previous edition of MAVEBA, held in December 2021 was celebrated in mixed mode just one year after the COVID-19 pandemic: a demonstration of strength and continuity despite the difficulties, which allowed us to meet once again from all over the World.

The series of MAVEBA International Workshops started in 1999 and is continuously proposed every two years as a multidisciplinary meeting. It concerns the study of the human voice both from the methodological point of view and its biomedical applications. 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. It welcomes contributions ranging from fundamental research and advanced technologies, with emphasis on translational research, the link with the “real” complex world of the human being.

This 13<sup>th</sup> Workshop will offer again the participants an interdisciplinary platform for presenting and sharing knowledge and recent results in this multifaceted subject that involves bioengineers, otolaryngologists, phoniaticians, neurologists, logopaedicians, linguistics, singers, actors, and any specialist in related fields, with applications ranging from the newborn to the elderly.

The papers presented at MAVEBA 2023 are divided into four Sessions:

**SESSION I – BIOMECHANICS AND MODELS**

**SESSION II – VOCAL FOLDS AND VOCAL TRACT**

**SESSION III – SINGING, DRAMA AND VOICE QUALITY**

**SESSION IV – TOOLS AND METHODS FOR SPEECH AND VOICE ANALYSIS**

The Workshop also includes two stimulating Round Tables, respectively organized by Prof. Johan Sundberg and Prof. Philippe DeJonckere:

**ROUND TABLE I: ACOUSTIC AND PHYSIOLOGICAL ASPECTS OF SINGING**

Moderator: J. Sundberg

Panelists: S. Capobianco, N. Henrich Bernardoni, M. Kob

**ROUND TABLE II: FOCUSING ON VOICE ONSET: A CRITICAL MOMENT OF PHONATION. FROM BASIC SCIENCE TO THERAPY**

Moderator: P. H. DeJonckere

Panelists: J. Sundberg, G. Cantarella, M. Kob, P. Aichinger

Moreover, a couple of Sessions are devoted to a **STUDENT COMPETITION**: the winner will be granted by a prize offered by the Journal Bioengineering (MDPI), an international, scientific, peer-reviewed, open access journal.

Last but not least, Prof. Dejonckere will give the following exciting lecture:

**THE «VIRTUAL MUSEUM OF PHONIATRICS»: A GUIDED TOUR BY THE CURATOR-IN-CHIEF P.H.DEJONCKERE**

### **ACKNOWLEDGEMENTS**

I greatly acknowledge my colleague Prof. Antonio Lanatà for his valuable contribution to the Workshop organization, PhD. Eng. Lorenzo Frassinetti and PhD. student Eng. Federico Calà, who managed and constantly updated the website, collaborated in reviewing the Proceedings and in solving the daily difficulties with patience and professionalism.

Thanks also to my Department, which supplied the congress material, to the FCRF for the economic contribution, to UNIFI, to the World Voice Day and to La Voce Artistica for the advertising on their respective websites.

Finally, a sincere and friendly thanks to the Scaramuzzi Team for their professionalism, that supported me for many years in this adventure. But above all I thank all the participants who, with their presence, wanted to be next to me once again. They 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 2023 Chair

Antonio Lanatà – Local organizing Committee

Lorenzo Frassinetti - Local organizing Committee





**SESSION I**  
**BIOMECHANICS AND MODELS**



# ELECTROMYOGRAPHIC ANALYSIS OF LIP AND FACE MUSCLES IN BEATBOXING

N. Henrich Bernardoni<sup>1</sup>, J. Frère<sup>1</sup>, A. Paroni<sup>1</sup>, S. Gerber<sup>1</sup>, H. Løevenbruck<sup>2</sup>

<sup>1</sup> Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, F-38000 Grenoble

<sup>2</sup> Univ. Grenoble Alpes, Univ. Savoie Mont-Blanc, CNRS, LPNC, F-38000 Grenoble

[Nathalie.Henrich@gipsa-lab.fr](mailto:Nathalie.Henrich@gipsa-lab.fr), [Julien.Frere@gipsa-lab.fr](mailto:Julien.Frere@gipsa-lab.fr)

**Abstract:** Muscular activations of lip and face muscles have been studied while performing beatbox sounds in comparison to speech sounds. Five trained beatboxers were recorded for various tasks involving the production of beatboxed sounds (kick, hi-hat, and rimshot) and related spoken syllables ([pu] for kick, [ti] for hi-hat, and [ka] for rimshot). Activations of orbicularis-oris lip muscle and zygomaticus-major face muscles were recorded by surface electromyography (EMG) during the production of isolated and repeated sounds in both beatboxing and speaking condition. Sound pressure level was much greater in beatboxing than in speaking. As expected, bilabial sounds kick and [p] involved higher levels of orbicularis muscle activation than the other sounds studied. This activity was significantly higher for kick than for [p]. A difference in lip and face muscular activity between beatbox and speech was not found for hi-hat and rimshot, even though the boxemes were produced louder than the consonants at same place of articulation. These results underline the value of beatbox exercises for working on labial and facial praxis, and on speech intelligibility.

**Keywords:** beatbox, electromyography, speech, consonant, orofacial muscle activity

## I. INTRODUCTION

Human Beatboxing is a recent and evolving musical practice, for which the beatboxer develops a particular motor skill in vocal instrumental playing [4, 5]. Recently, beatboxing exercises have been proposed in speech rehabilitation of articulation disorders in the case of young adults with congenital dysarthria and reduced speech intelligibility [1–3]. These studies have demonstrated that beatboxing could be an effective tool for improving articulatory deficiencies, speech production, and intelligibility. Yet, no study has ever assessed the orofacial muscular load in such sound production. The aim of the present study is to explore how muscular activity differs between producing kick/hi-hat/rimshot beatboxing effects, also called

boxemes [4], and three speech /p,t,k/ consonants with similar place of articulation uttered in syllabic consonant-vowel (CV) context.

## II. METHODS

*Subjects:* Five trained French-speaking beatboxers were recorded for various tasks involving the production of beatboxed sounds (kick, hi-hat, and rimshot). The related spoken syllables were also recorded ([pu] for kick, [ti] for hi-hat, and [ka] for rimshot). For each task, the beatboxer performed a series of twelve repeated sounds or spoken syllables. For the subsequent analyses, nine occurrences in the middle of the series were retained.

*Equipment and audio analysis:* The recordings took place in a semi-anechoic laboratory room authorized for biomedical research. Audio and electroglottographic (EGG) signals were recorded synchronously and sampled at 20 kHz on 16 bits (dual-channel electroglottograph Glottal Enterprise EG2). Sound files were annotated by means of Praat software. Burst instants (release of occlusion and start of sound) and end of sound were manually detected on audio signal. For speech, end of consonant sound corresponded to start of voicing as assessed on audio and EGG signal. Sound pressure level (SPL) was computed on the annotated sound window (either boxeme or consonant part of the spoken syllable) with Matlab software.

*Electromyographic analysis:* Electrodes were placed on the lips in correspondence with the four sections of orbicularis oris (OO) muscle and bilaterally on the face in the region of zygomaticus major (ZYG) muscles (see Fig. 1). Activation of these lip and face muscles was recorded by surface electromyography (EMG) during the production of isolated and repeated sounds (MP150 BIOPAC, 20 kHz). For each task, EMG signals were bandpass filtered (5–450 Hz) and root-mean-squared (RMS) over a 25-ms sliding window to produce a linear envelope for each muscle activity pattern. Finally, RMS values were averaged over a time window including both the sound (boxeme/consonant) and the 100 ms preceding the burst instant (see Fig. 2).

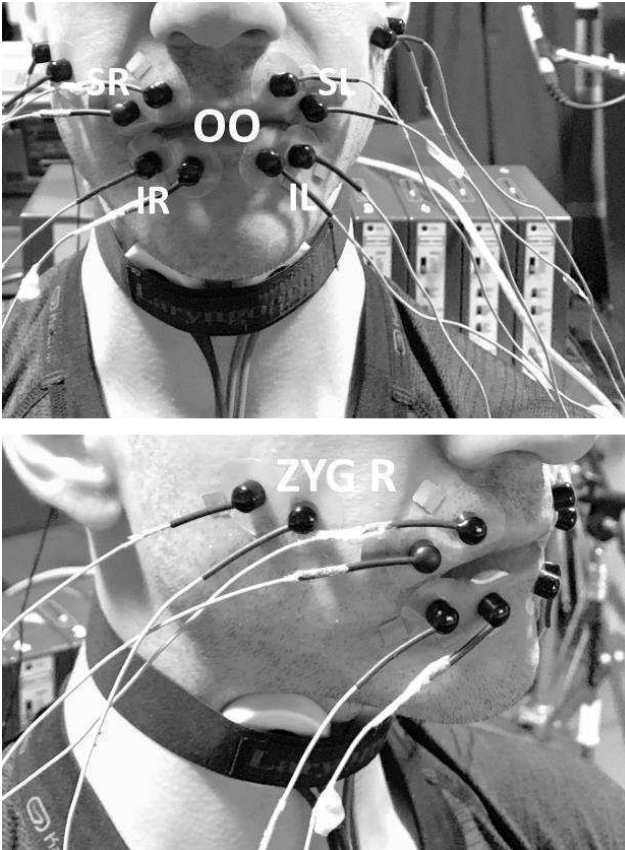


Fig. 1 : Placement of EMG surface electrodes. ZYG\_L/ZYG\_R: left and right zygomatic muscles. OO\_SR/SL: right and left superior orbicularis oris, OO\_IR/IL right and left inferior orbicularis oris.

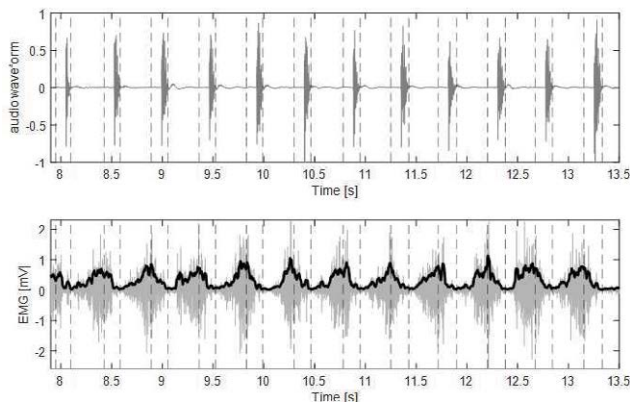


Fig. 2 : audio waveform (upper panel) and OO\_SL muscle activity (lower panel) of one participant performing a series of kick sounds. Each vertical line represents the beginning and the end of the time window of interest. Lower panel: filtered EMG signal (in grey) and RMS envelope (black bold line).

*Statistical analysis:* Statistical models were implemented in R language, using a linear mixed modeling approach (function lme of package nlme).

They aimed to explore the impact of BOXEME factor (6 modalities, p, t, k, kick, hi-hat, rimshot), MUSCLE factor (OO and ZYG) and their interaction on variation in the response variable RMS in voltage. The hypothesis that labial and facial muscular activities measured by surface EMG would be higher in beatboxing than in speech was tested.

### III. RESULTS

#### A. Sound pressure level in beatboxing and speaking.

As shown in Fig. 1 Fig. 3, sound intensity was always found to be greater in beatboxing than in speaking, whatever the place of articulation.

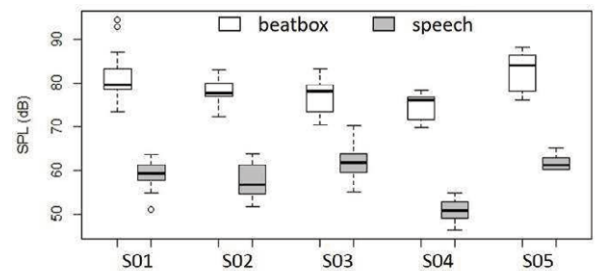


Fig. 3 : SPL values in dB for five beatboxers either beatboxing or speaking (consonant part).

#### B. Muscular activation in beatboxing and speaking

Fig. 4 presents the EMG activation amount for lip and face muscles in human beatboxing and in speech. Kick and [p] are bilabial sounds, i.e. sounds that are produced via complete occlusion and subsequent release of the vocal tract at the lips. In both cases, the occlusion is achieved by recruiting the orbicularis oris muscle. Yet, muscular activities were significantly higher for kick than for [p] ( $p < 0.001$ ) for all the studied muscles. Rimshot compared to [k] did not significantly increase the muscle activities, except for OO\_SR ( $p = 0.024$ ), even though the vocal-tract occlusion to produce these two sounds is situated far from the lips, in the back region of the oral cavity. Finally, hi-hat sound did not lead to higher level of muscle activity compared to [t], except for ZYG\_R ( $p = 0.010$ ). In such sounds, occlusion is achieved by the action of tongue pushing against upper teeth.

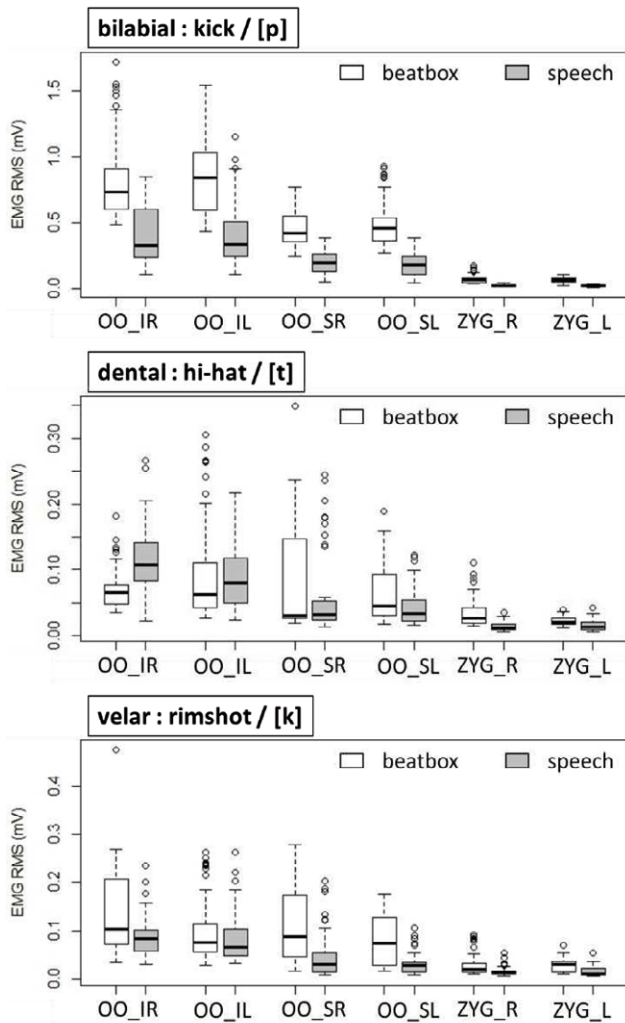


Fig. 4 : Amount of EMG RMS signal for beatboxing and speaking, depending on the place of articulation (bilabial, dental, velar). For muscles abbreviations, see Fig. 1.

#### IV. DISCUSSION AND CONCLUSION

An underlying question in this study was whether beatboxing exercises would be suitable for rehabilitation in the case of dysarthria and orofacial myofunctional disorders. Enhanced lip and facial muscular activities have been evidenced in beatboxing sounds as compared to speaking CV syllables with similar place of articulation for consonants. This has been verified for kick boxemes essentially. It has not been found for rimshot and hi-hat boxemes articulated

respectively in the velar and dental region. These sounds did not activate the lip and face muscles more than speech ones, even though they resulted in louder sounds.

An enhanced muscular contraction can be beneficial for working on labial and facial praxis in the context of dysarthria. It could contribute to muscle strengthening in the case of orofacial myofunctional disorders. However, the speech therapist who would use such beatboxing exercises should take great care in avoiding muscular overload or fatigue that may be induced. In this respect, hi-hat and rimshot boxemes are of much interest as they can be produced loud with a muscular effort comparable to speaking consonants.

Lip and face muscles were explored here. Another major articulator in beatboxing is the tongue. In a further study, it would be worth exploring whether beatboxing may require enhanced tongue muscular activation.

#### REFERENCES

- [1] M. Icht, "Improving speech characteristics of young adults with congenital dysarthria: An exploratory study comparing articulation training and the Beataalk method", *Journal of Communication Disorders*, 93, 2021, 106147.
- [2] M. Icht, "Introducing the Beataalk technique: using beatbox sounds and rhythms to improve speech characteristics of adults with intellectual disability", *International Journal of Language & Communication Disorders*, 54, 3, 2019, pp. 401–416.
- [3] M. Icht, and M. Carl, "Points of view: positive effects of the Beataalk technique on speech characteristics of young adults with intellectual disability", *International Journal of Developmental Disabilities*, 2022, pp. 1–5.
- [4] A. Paroni, N. Henrich Bernardoni, C. Savariaux, H. Løevenbruck H., 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, 2021, pp. 191–206.
- [5] 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, 2013, pp. 1043–1054.



# ON THE USE OF BIOMECHANICAL MODELS FOR VOCAL FOLD EDGE TRACKING IN VIDEOKYMOGRAMS

C. Drioli, G. L. Foresti

<sup>1</sup> Department of Mathematics, Computer Science and Physics, University of Udine, Italy.  
carlo.drioli@uniud.it, gianluca.foresti@uniud.it

**Abstract:** Videokymography (VKG) is a low-cost, high-speed imaging method for the examination of the vocal folds. We report here on the mixed use of conventional video processing techniques and biomechanical modelling of the laryngeal dynamics, to enhance the detection and tracking of the fold edges during oscillation. Analysis result examples related to data samples of different quality are discussed.

**Keywords:** Videokymography, vocal folds dynamical modelling, voice quality characterization, voice disorders.

## I. INTRODUCTION

Videokymography, introduced by Švec and Schutte in 1996 [1], is a low-cost, high-speed imaging method for the examination of the vocal folds, which allows the visualization of regular and irregular vibration patterns. Its usefulness in phonation investigation and diagnosis of voice pathologies has been documented, e.g., in [2], [3].

In this communication, we describe the processing of VKG data through the fitting of a biomechanical model within a Bayesian setting to enhance the estimation of the position of fold edges during the glottal cycle. Specifically, the method addresses the tuning of the model during the glottal open phase to fit the position of the edge of the folds, and uses in turn the tuned model to predict the observation in the next analysis windows. It is also investigated the possibility to infer the position of vocal fold edge position in those intervals of the glottal cycle in which edges are not clearly distinguishable or in which no observation data is available due to visual occlusion. E.g., when top of the vocal folds are more adducted than the bottom (convergent glottal configurations), lower edges are not visible from above. The fitting algorithm relies on a biomechanical model whose parameters are adapted so that his time evolution is coherent with the folds edge position estimated from the VKG data. We have recently investigated the use of this class of models to interpret acoustic and visual data recordings related to the fold oscillations [4], [5], [6]. In the present communication, the biomechanical model is used for the Bayesian inference as a state

transition model, with a dual role: on one hand, it models the folds edge motion to compute the likelihood of their position in given portions of the glottal cycle; on the other hand, its parameters are finely tuned to maximize the likelihood of the visual observations. The method is assessed on VKG data from healthy subjects uttering sustained vowels. It is shown that the use of a biomechanical model of the folds as a state transition model permits to accurately fit the upper and lower vocal fold edges during the intervals in which both are clearly visible, and permits to infer their position in partial or complete fold occlusion conditions of the glottal cycle.

## II. METHOD: VIDEO PROCESSING, BIOMECHANICAL MODEL, EDGE TRACKING

The video analysis procedure described in the following is aimed at estimate the motion of the vocal fold edges from a high-speed video sequence  $I(x, y, t)$  in the form of a Videokymogram, in which the vocal fold vibration is represented by the displacement trajectories in time of their upper and lower edges observed at a given medial position of the glottis. The whole process is sketched in Figure 1. Starting from top, the process is organized in three subsequent processing steps: a preliminary video analysis step, followed by a model fitting and motion prediction step, terminated by a model-to data superposition step for visualization purposes. In the upper part of the figure, the plots show the interpretation of a fragment of videokymographic data, corresponding to two glottal cycles, and sketches the image preprocessing step in which the open phase of each cycle is isolated and various spatial cues related to edges are quantified. Note that actual recorded data is characterized by clearly distinguishable romboidal-shaped regions related to the open phase, but provide barely visible information concerning the upper folds edge position during the closing interval and no information at all concerning the lower folds edge position during the opening interval (due to camera occlusion). Moreover, the VKG data is often characterized by asymmetries with respect to the L/R direction.

The middle part of the figure illustrate how the mass-spring model is used within a Bayesian framework,



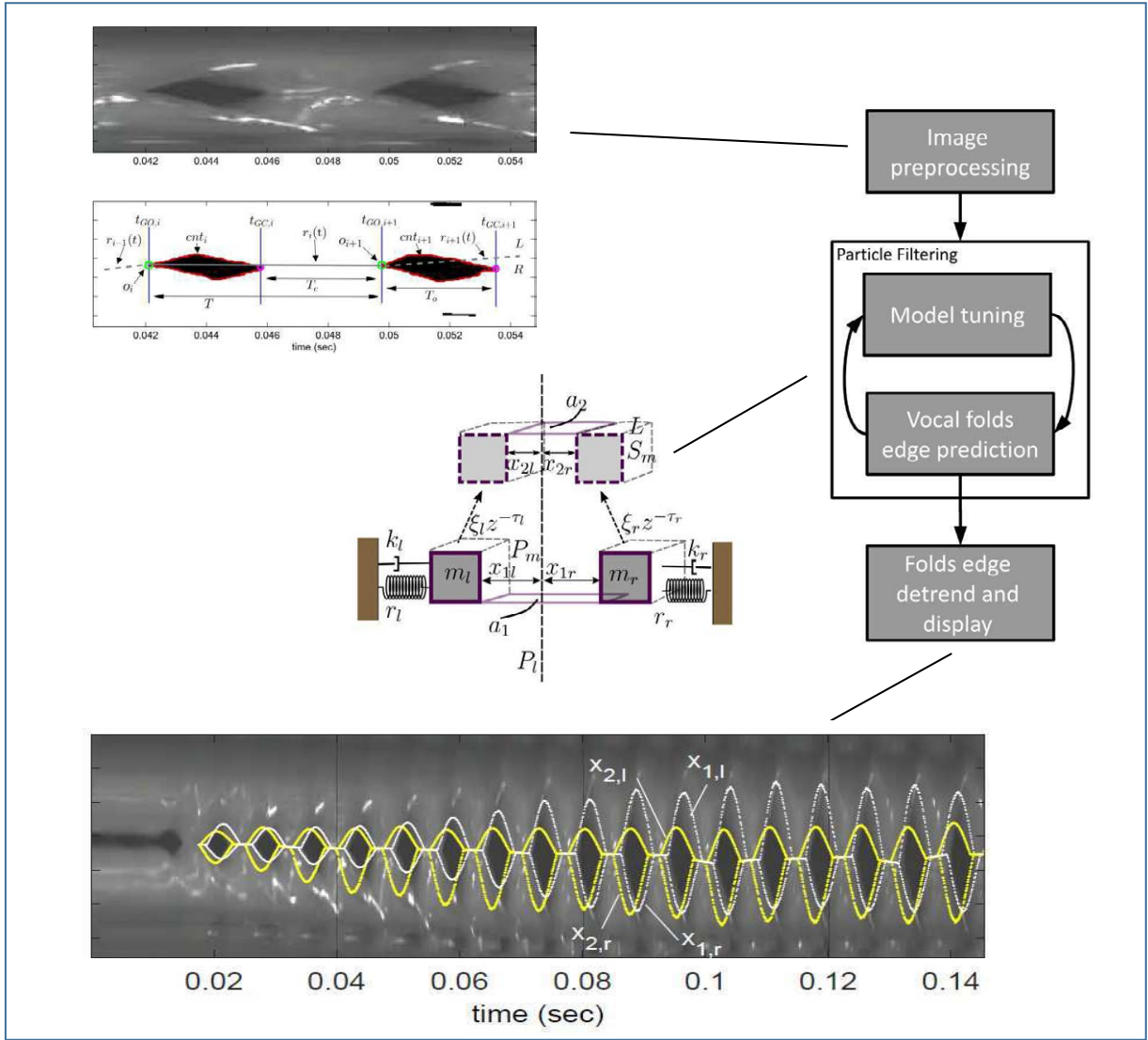


Fig. 1. Schematic view of the whole process.

namely that of particle filtering, to tune the model and to use it to predict the fold behaviour in the next analysis time window (forecast). In the model, the inferior edge of each fold is represented by a single mass-spring system with stiffness  $k$ , damping  $r$  and mass  $m$ . The L-R asymmetry is modelled by using two different single-mass systems, one for each fold. The phase difference of the vibration between the lower and the upper edge, which is essential for the modeling of self-sustained oscillations, is modeled by a delay of the displacement induced by its propagation along the cover of the fold [7], [8]. Let us call  $x_{1,l}$  the displacement of the left fold at the entrance of the glottis (lower edge), and  $x_{2,l}$  the displacement at the exit (upper edge). The displacements of the right fold are named accordingly  $x_{1,r}$  and  $x_{2,r}$ . Other details of the model can be found

in [6] and are not reported here.

The discretization of the model equations leads to a discrete-time system that is numerically solved to obtain an estimate of the glottal flow  $U_g(nT_s)$  and of the folds displacements  $x_j^{i,\alpha}(nT_s)$  and  $x_j^{s,\alpha}(nT_s)$ , at discrete time  $n$ . The model is run with sampling frequency  $F_s = (1/T_s) = 22050$  Hz, and the oscillatory patterns are visualized as if the folds were observed from above using a line scan device. The oscillation patterns obtained are then superimposed to the actual VKG data.

The fitting of the model to the visual observation is obtained by a joint model parameter estimation and model state estimation, using a Bayesian estimation process. If  $\mathbf{z}_{1:k}$  is the set of observations up to time instant  $k$ ,  $\mathbf{x}_k$  is the state of the fold edge at time

instant  $k$ , and  $\theta_k$  is the set of parameters at time  $k$ , then we are interested in the computation of the posterior probability  $p(\mathbf{x}_k, \theta_k | \mathbf{z}_{1:k})$ . This probability can be recursively computed as

$$p(\mathbf{x}_k, \theta_k | \mathbf{z}_{1:k}) = \frac{p(\mathbf{z}_k | \mathbf{x}_k, \theta_k) p(\mathbf{x}_k | \theta_k, \mathbf{z}_{1:k-1}) p(\theta_k | \mathbf{z}_{1:k-1})}{p(\mathbf{z}_k | \mathbf{z}_{1:k-1})} \quad (1)$$

where  $p(\mathbf{z}_k | \mathbf{x}_k, \theta_k)$  is the likelihood probability,  $p(\mathbf{x}_k | \theta_k, \mathbf{z}_{1:k-1})$  is the state prior,  $p(\theta_k | \mathbf{z}_{1:k-1})$  is the parameter set prior, and  $p(\mathbf{z}_k | \mathbf{z}_{1:k-1})$  is the marginal likelihood. Since it is  $p(\mathbf{x}_k | \theta_k, \mathbf{z}_{1:k-1}) p(\theta_k | \mathbf{z}_{1:k-1}) = p(\mathbf{x}_k, \theta_k | \mathbf{z}_{1:k-1})$ , joint parameter and state estimation can be achieved through augmentation of the state space by the parameter vector [9]. Assuming that the posterior pdf is available at time  $k-1$ , the prior (or prediction) pdf can be computed as

$$p(\mathbf{x}_k, \theta_k | \mathbf{z}_{1:k-1}) = \int p(\mathbf{x}_k, \theta_k | \mathbf{x}_{k-1}, \theta_{k-1}) p(\mathbf{x}_{k-1}, \theta_{k-1} | \mathbf{z}_{1:k-1}) d\mathbf{x}_{k-1}, d\theta_{k-1} \quad (2)$$

To complete the Bayesian estimation scheme, a likelihood function is required that provide a reliable measure of how well an image observation  $I(x, y, k)$  is explained by a particular hypothesis (model prediction). If we suppose that a set of video features  $\mathbf{f}(I(x, y, k))$  related to the folds edge can computed from the image frame, then we can define the likelihood  $p(\mathbf{z}_k | \mathbf{x}_k)$  at discrete instant  $k$  as

$$p(\mathbf{z}_k | \mathbf{x}_k) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{|\mathbf{f}(I(x, y, k)) - \mathbf{x}_k|^2}{2\sigma^2}\right) \quad (3)$$

In our case, we compute from the romboidal patterns in the VKG image a set of features that can be related to the observable state of the folds model, i.e. the lower and upper edge of both left and right vocal folds. The edge-related features are sketched in the top image of Figure 1, and a likelihood function is built upon these features (a detailed description can be found in [6]). The use of such function within the particle filter framework will allow to fit the folds displacement in the regions where features can be computed from the available information, and to provide an estimation of the position based on the prediction of the model in those time intervals in which information is missing.

Note that the temporal prior pdf  $p(\mathbf{x}_k, \theta_k | \mathbf{x}_{k-1}, \theta_{k-1})$  provides an estimate of the update of the state and parameters at time  $k$ , given the state and parameters at time  $k-1$ , in other words it models the dynamics of the process under observation. We propose in this case to use the biomechanical numerical model of the vocal folds as state transition model, and assume that the state vector  $\mathbf{x}_k$  is the displacement of the vocal fold as predicted by the numerical simulation of the model. For the update of the parameters, a pitch-synchronous random walk

model is assumed, i.e.

$$\theta_{Tk} = \theta_{T(k-1)} + \phi_k \quad (4)$$

where  $\phi_k \in N(0, W_\phi)$  satisfies a Gaussian distribution with zero mean and covariance matrix  $W_\phi$ . The parameters are thus assumed constant during a glottal cycle. Note that the optimization process of the parameters can be very sensitive to the initial hypothesis and to the variance of the parameter. An advantage of using a physically informed model in the process is that often a starting hypothesis can be done on a physiological basis (see, e.g. [8] for a discussion on the empirical tuning of these parameters).

The natural frequency of a mass-spring system is  $f_0 = 1/2\pi\sqrt{k/m}$ , thus its parameters  $k$  and  $m$  can be tuned accordingly when a given oscillation period of the model is desired.

A closed-form solution of eq. 1 and eq. 2 is in general not feasible, and a numerical approximation is often sought instead. We propose here the use of a Particle Filtering scheme (PF), with a Sequential Importance Resampling algorithm (SIR) to represent the posterior [10], [11], [12], [13]. The underline principle is to form a weighted particle representation of the posterior distribution, as  $p(\mathbf{x}_k, \theta_k | \mathbf{z}_{1:k}) \approx \sum_i w^{(i)} \delta(\mathbf{x}_k - \mathbf{x}_k^{(i)})$ , where  $\{(w_k^{(i)}, x_k^{(i)})\}$ ,  $i = 1, \dots, N$  is the set of particles and of the corresponding weights at instant  $k$ .

The biomechanical model is involved in the prediction step, where each particle can be considered as an independent instance of the model simulation. In the following, we will include in the estimation process three model parameters for each fold, i.e. the natural frequency  $f_\alpha$ , the vertical phase delay  $\tau_\alpha$ , and the upper-to-lower edge amplitude ratio  $\xi_\alpha$ . Hence the parameter vector is  $\theta = \{f_l, f_r, \tau_l, \tau_r, \xi_l, \xi_r\}$ .

### III. RESULTS

Some experimental results obtained using the procedure on actual VKGs data from publicly available datasets can be observed in Figs. 1, 2, and 3. In Figure 1, the bottom plot shows a fragment of approximately 140 msec, and the particle filtering fit to the observation (yellow and white scatter plot refer to x1 and x2 estimates respectively). The process is iterated on short overlapping time windows of two pulses each. The plot evidentiates the pulse-to-pulse adaptation of the model to the slowly varying open phase patterns and to the trend induced by the relative shifts of the endoscope with respect to the oscillating folds during the recording.

In Figure 2 the details of two subsequent short analysis windows are shown, evidentiating the evolution of L/R and x1/x2 asymmetries and the adaptation of the model-driven particles to the data. In Figure 3, the plots

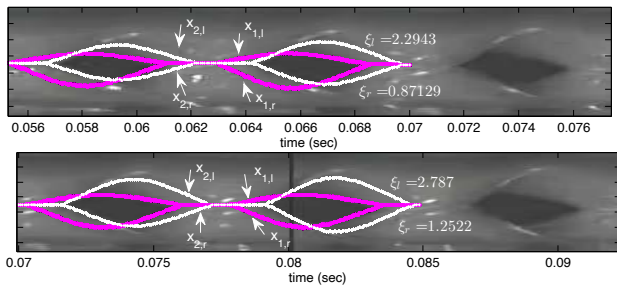


Fig. 2. VKG video analysis and vocal fold edge fitting referred to two analysis windows: asymmetries L vs R and  $x_1$  vs  $x_2$ . The scattered plots superimposed to the VKG image represent the evolution of particles related to  $x_{1,r}$  (magenta, upper portion),  $x_{1,l}$  (magenta, lower portion),  $x_{2,r}$  (white, upper portion), and  $x_{2,l}$  (white, lower portion).

show the details of a frame from a different dataset, characterized by lower resolution and noisy/blurred image quality.

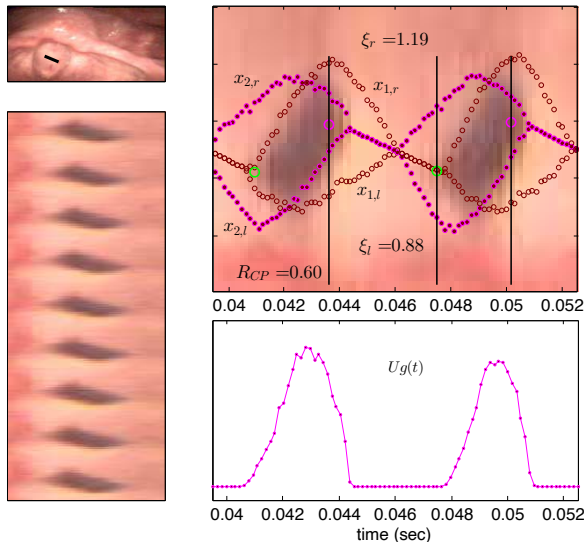


Fig. 3. Fitting results for a VKG recording from a healthy male subject, modal phonation, pitch: 160 Hz. Recordings from the database by E. Bianco and G. Degottex, IRCAM.

#### IV. CONCLUSIONS

We discussed the analysis of videokymographic data with a Bayesian estimation procedure based on the fitting of a biomechanical model of the folds to the visual data. The method is shown to be able to accurately fit the vocal folds edge displacement extracted from the videokymogram, at least in the open-phase intervals of the glottal cycle, where the edges are clearly visible, and to track the time-varying oscillatory patterns observed in the data. In the portions of the cycle where edges are not clearly visible or where partial occlusion occurs, the method provides a prediction of the fold edge position

based on the dynamics of the vocal folds model. In these particular regions, however, it is not possible to measure the accuracy of the predicted cues with the data at hand, due to the lack of a ground truth. Further investigation is thus foreseen in this direction, by using different datasets, built e.g. with highly realistic numerical models of the folds or collected by in-vitro experiments.

#### V. ACKNOWLEDGEMENTS

This work was partially supported by the DMIF's Departmental Strategic Plan (PSD) of the University of Udine – Interdepartmental Project on Artificial Intelligence (2020-25)

#### REFERENCES

- [1] J. G. Švec and H. K. Schutte, "Videokymography: High-speed line scanning of vocal fold vibration," *Journal of Voice*, vol. 10, no. 2, pp. 201–205, 1996.
- [2] H. K. Schutte, J. G. Švec, and F. Šram, "First results of clinical application of videokymography," *Laryngoscope*, vol. 108, no. 8-1, pp. 1206–10, 1998.
- [3] J. G. Švec, F. Šram, and H. K. Schutte, "Videokymography in voice disorders: What to look for?" *Annals of Otology Rhinology and Laryngology*, vol. 116, no. 3, pp. 172–180, 2007.
- [4] C. Drioli and G. L. Foresti, "Accurate glottal model parametrization by integrating audio and high-speed endoscopic video data," *Signal, Image and Video Processing*, vol. 9, no. 6, pp. 451–459, 2015.
- [5] —, "Quantitative characterization of functional voice disorders using motion analysis of highspeed video and modeling," in *Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN)*, July 2015, pp. 1–6.
- [6] —, "Fitting a biomechanical model of the folds to high-speed video data through bayesian estimation," *Informatics in Medicine Unlocked*, vol. 20, p. 100373, 2020.
- [7] 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.
- [8] C. Drioli, "A flow waveform-matched low-dimensional glottal model based on physical knowledge," *J. Acoust. Soc. Am.*, vol. 117, no. 5, pp. 3184–3195, May 2005.
- [9] S. Särkkä, *Bayesian Filtering and Smoothing*, ser. IMS Textbooks. Cambridge University Press, 2013, vol. 3.
- [10] M. S. Arulampalam, S. Maskell, and N. Gordon, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, pp. 174–188, 2002.
- [11] J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill, "Particle methods for Bayesian modeling and enhancement of speech signals," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, no. 3, pp. 173–185, Mar. 2002.
- [12] M. Vondrak, L. Sigal, and O. C. Jenkins, "Physical simulation for probabilistic motion tracking," in *Proceedings of Computer Vision and Pattern Recognition Conference (CVPR)*. IEEE Computer Society, 2008.
- [13] A. Dore, M. Soto, and C. Regazzoni, "Bayesian tracking for video analytics," *Signal Processing Magazine, IEEE*, vol. 27, no. 5, pp. 46–55, 2010.

# AN ACOUSTIC ANALYSIS OF VOWELS TO PREDICT VOICE CHANGES IN A LONG READING TASK

Martin Hagmüller<sup>1</sup>, Julian Linke<sup>1</sup>, Simon Lohrmann<sup>1</sup>, Florian Pokorny<sup>2,3</sup>, Barbara Schuppler<sup>1</sup>

<sup>1</sup>Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria

<sup>2</sup> Division of Phoniatics, Medical University of Graz, Austria

<sup>3</sup> Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

{hagmueller, linke, b.schuppler}@tugraz.at, simon.lohrmann@student.tugraz.at, florian.pokorny@medunigraz.at

**The human voice is the central instrument for many professions and keeping a healthy voice is thus of high importance for, among others, teachers, singers and call-center employees. High vocal load may lead to vocal fatigue, a phenomenon which has been investigated with respect to its acoustic characteristics and its potential to be automatically predicted from acoustic features. The features used in most of these studies, however, do not only change over the course of a speaking task given vocal fatigue, but might also change due to other aspects of speaking for a long time, such as reduced articulatory effort and/or emotional accommodation to the task. This paper analyzes different types of acoustic features (MFCCs, voice-quality related, formants) which potentially reflect different articulatory phenomena occurring in a 90-minute-long reading task. Given the speakers' perception of their own vocal effort during the task, we provide a speaker-dependent discussion about which acoustic features most reflect vocal fatigue. For most speakers, we observe a vowel space reduction during the long reading task.**

**Keywords:** vocal fatigue, voice quality features, vowel space

## I. INTRODUCTION

For many occupations, the human voice is the most important instrument, which depending on the required vocal load and subjective disposition might lead to vocal fatigue. If the speaker is keeping up the high voice use without sufficient rest, a voice disorder will develop [1]. Some even experience temporal muteness that prohibits to exercise their job. Especially affected are, among others, (kindergarten) teachers, call-center employees, and singers.

Several approaches to detect vocal fatigue have been proposed. One approach is the use of questionnaires such as the Vocal Fatigue Index (VFI) [2], another – albeit elaborate – approach is a phoniatic assessment that is usually performed at a clinic. Both approaches do not provide continuous monitoring of a speaker or singer. Voice dosimeters have been developed to measure how much a speaker is using the voice over a certain period and can give a warning in case of voice overuse [3]. Voice dosimetry is only measuring the voice usage, but does not evaluate any degradation of voice quality that might indicate negative effects of voice overuse.

Acoustic signal analysis promises to provide non-invasive easy or continuous monitoring of the voice. While many studies have used voice-related features,

such as based on fundamental frequency ( $F_0$ ) or Harmonics-to-Noise (HNR) ratio [1], [4], more general features like Mel-frequency cepstral coefficients (MFCCs) [5], [6] have been used as well. MFCCs are widely used in speech analysis for tasks such as automatic speech recognition, speaker recognition and verification, as well as emotion recognition, among others [7]. Since MFCCs describe the power spectrum of a signal, they are effective for detecting voice changes, but other phenomena are represented as well, as shown by their wide range of application areas. We therefore hypothesize that MFCCs are also sensitive to vocal fatigue, as evoked in a long reading task [4], [5].

As discussed by Caraty et al. [5], reading is a complex task that can lead to cognitive fatigue when performed over a long period of time. Cognitive fatigue has been successfully detected from the voice signal, e.g. by using MFCCs as features [8], [9]. We therefore assume that when MFCCs are used to detect vocal fatigue, they might respond also to other phenomena typical for long speaking tasks, such as for instance a change in articulation due to a reduction of the vowel space [10].

On the basis of recordings from a 90 min long reading task, this study aims to untangle voice changes potentially indicating vocal fatigue from other changes due to accommodation to the reading situation and/or cognitive tiredness. The first part (Sec. III) analyzes voice quality and MFCC features in a speaker-dependent way by comparing feature importances and classification results from random forest classifiers. The second part (Sec. IV) analyzes the vowel space changes of each speaker by comparing the first and second formant frequencies at the beginning and the end of each recording.

## II. DATASET

### A. Data and Forced Alignment with MAUS

The data of this study comes from 90 minutes-long simulated lecturer presentations from four different male speakers. For each speaker, we recorded speech data with a Tascam DRX-05 stereo field recorder in a lecture hall of the Graz University of Technology, Austria. The speakers were distanced approx. 1 m from the microphone and all of them read the same German text from a scientific book [11]. For each recording, we created audio chunks of 10–15s and orthographically

TABLE I: Used vowels and number of their occurrences (for all speakers).

vowel	#	vowel	#	vowel	#	vowel	#
/a/	453	/a:/	327	/i/	88	/i:/	401
/ɪ/	50	/e/	59	/e:/	323	/ɛ/	360
/ɛ:/	42	/u/	77	/u:/	77	/ʊ/	251
/o/	55	/o:/	178	/ɔ/	243	/aʊ/	146
/ai/	419	/ɔy/	39	/y/	22	/y:/	79
/ʌ:/	39	/ɒ/	405	/ə/	16		

transcribed those audio chunks such that the text-files contained the spoken words of the corresponding 10–15s long wav-snippets. The wav-snippets along with the Praat TextGrid-files [12] were uploaded to the WebMAUS Basic tool [13], which automatically delivered us phonetic segments of the speech material. Using these segments, we extracted approx. 1000 vowels per speaker resulting in 4149 vowel segments in total. Table I shows an overview of the number of vowels per type used for this paper (for all speakers).

### B. Speaker Characteristics

The four male speakers with IDs 101M, 102M, 103M and 104M were aged 30, 31, 30, and 50 years, respectively. None of them had a voice disorder at the time of recording. Speakers 101M and 104M lived most of their lives in Vienna and Graz, respectively (both belonging to the Eastern Austrian dialectal area) and speakers 102M and 103M in Tuttlingen and Bamberg, respectively (both Southern German). The younger speakers (101M, 102M and 103M) had little experience as presenters, whereas speaker 104M (50 years) had more speaking experience. The recording sessions were interrupted every 10 minutes to ask the speakers five short questions related to stress, fatigue, vocal roughness, excitement, alertness and concentration. These questions had to be answered verbally. For each question, speakers gave ratings which ranged between 0 (not applicable at all) and 10 (entirely applicable). In general, answers which related to stress, fatigue and vocal roughness showed that the less experienced speakers (101M, 102M and 103M) introduced higher ratings than the more experienced speaker (104M). Simultaneously, in case of answers which related to excitement, alertness and concentration, speaker 104M gave a rating of 0 at each questioning relating to excitement and higher ratings for questions relating to alertness and concentration. In case of the less experienced speakers all of those ratings were more varying while indicating descending ratings in excitement (101M and 102M), alertness (101M and 103M) and concentration (101M and 103M).

## III. ANALYSIS OF VOICE QUALITY

### A. Materials and Method

1) *MFCC Features*: We extracted the Mel-frequency cepstral coefficients (MFCCs) which relate directly to spectral characteristics while indirectly incorporating

TABLE II: Mean F1 scores from a 10-fold cross-validation for each speaker when training binary random forest classifiers (negative class: start; positive class: end) with three different feature sets.

Feature Set	101M	102M	103M	104M
VQ	0.73	0.66	0.59	0.55
MFCCs	0.59	0.70	0.69	0.62
VQ+MFCCs	0.73	0.72	0.71	0.65

also voice quality characteristics. MFCCs 1–13 were calculated by using *librosa*<sup>2</sup> [14] version 0.9.1. An FFT-window with a length of 23 ms and an overlap of 6 ms was used by applying a Hanning window. The number of filters in the Mel-filter bank was set to 128, the frequency range was from 50 Hz to 22 050 Hz. The resulting MFCCs were averaged over all FFT-windows of one vowel segment, such that we receive one mean value per vowel segment for each of the MFCCs. Those values were then normalized to have zero mean and unit variance on a single-speaker basis.

2) *Voice Quality Features*: We extracted 10 voice quality (VQ) features which assess the phonation and resonance characteristics of the voice of each speaker. We employed the toolbox outlined in [15] which utilizes the two Python toolkits *parselmouth*<sup>1</sup> (version 0.4.1) and *librosa*<sup>2</sup> in order to calculate the VQ features *Jitter*<sup>1</sup>, *Shimmer*<sup>1</sup>, *HNR*<sup>1</sup>, *mean of F<sub>0</sub> (fundamental frequency)*<sup>1</sup>, *CPP (Cepstral Peak Prominence)*<sup>1</sup>, *mean of F3 (third vowel formant)*<sup>1</sup>, *mean of RMS*<sup>2</sup>, *ZCR (Zero Crossing Rate)*, *STE (Short Time Energy)* and *H1\_H2 (energy between first two harmonics)*.

3) *Random Forest Classifiers*: We compared the first five minutes of every 90 minutes-long simulated lecturer presentation to its last five minutes on a speaker-dependent basis. For all binary classification tasks, we trained and tested a random forest classifier (RFC) with default settings given the *scikit learn* toolkit (version 1.0.2) [16] (Python 3.9.7) with a 10-fold cross-validation and present respective averaged F1 scores. We tested three different feature combinations, namely voice quality features (VQ), MFCC features (MFCCs) and a combination of both feature sets (VQ+MFCCs), leading to 12 binary classification tasks (three feature combinations x four speakers), while defining the end of a recording as the positive class. One purpose of using RFCs is its ability to provide both, classification results and impurity-based feature importances. Hence, RFCs make it possible to analyze not only which features are decisive in making a difference between the first and the last five minutes of a recording, but also how large the features' impact is relative to each other, allowing us to learn which acoustic features most represent vocal fatigue.

<sup>1</sup><https://parselmouth.readthedocs.io/en/stable/>

<sup>2</sup><https://librosa.org/doc-playground/main/index.html>

TABLE III: Vowel space areas (Bark) and their differences ( $\Delta$ ) between the beginning and the end of a recording for each speaker. Vowel space areas were measured with convex hulls derived from vowel means.

	101M	102M	103M	104M
start	11.17	9	10.19	12.12
end	11.28	7.72	8.69	12.05
$\Delta$	-0.11	1.28	1.5	0.07

## B. Results

Table II summarizes F1 scores for each speaker when training binary RFCs with three different feature sets. We observe that RFCs which were trained entirely on VQ features led to worse F1 scores in case of speakers 103M and 104M ( $\leq 0.59$ ) but to better F1 scores in case of speaker 101M (0.73) and 102M (0.66). In contrast, RFCs which were trained entirely on MFCC features achieved best F1 scores in case of speakers 102M (0.7) and 103M (0.69), whereas speakers 101M and 104M had worse F1 scores of 0.59 and 0.62. When the two feature sets were combined (VQ+MFCCs), the best overall F1 scores were achieved. Specifically, for speakers 101M, 102M and 103M, we achieved F1 scores of 0.73, 0.72 and 0.71. Yet, for the most experienced speaker 104M the incorporation of VQ features only yielded an F1 score of 0.65. Overall, classification results indicate that speaker 101M varies more over time with respect to voice quality related features, whereas the vocal changes of speakers 102M, 103M and 104M are rather represented by MFCC features.

With respect to the feature importances for the RFC with VQ+MFCCs, we observe that the 4 most important features of speaker 101M comprised *mean of  $F_0$* , *STE*, *mean of RMS* and *H1\_H2*, which all had average importances  $> 6\%$  (capturing 34% of the overall importance). In contrast, the four best features for speakers 102M and 103M were MFCCs capturing 30% (102M) and 31% (103M) of the overall importance. In case of speaker 102M, the next best feature was *mean of  $F_0$*  with an average importance of 6%. In case of speaker 103M, the best MFCC (coefficient 8) had an average importance of 13% while the next best MFCC (coefficient 2) had an average importance of 7%. The next best VQ features were *CPP*, *ZCR* and *mean of  $F_3$*  with average importances of approx. 5%. In case of speaker 104M all features had similar average importances (between approx. 3, 5%–6, 5%) where the 10 best features captured 53% of the overall importance and they comprised MFCCs (coefficients 7, 12, 2, 1, 10, 9 and 8), *ZCR*, *HNR* and *CPP*.

## IV. ANALYSIS OF THE VOWEL SPACE

### A. Materials and Method

We extracted the first (F1) and the second (F2) vowel formants in order to analyze the speaker's vowel spaces

resulting in the two features *mean of F1* and *mean of F2*. Both features were calculated with *parselmouth* with a window length of 25 ms and a formant ceiling of 5000 Hz.

We analyzed the formant features by comparing the cardinal, tense vowels (/a/,/a:/, /e:/, /i:/, /i:/, /o:/, /o:/, and /u:/, /u:/) given the means and standard deviations of *mean of F1* and *mean of F2* at the beginning and at the end of the recordings. For each speaker separately, we calculated vowel space areas by measuring convex hulls in the formant's Bark space by utilizing the *Qhull library* [17].

## B. Results

Fig. 1 shows the mean values and the standard deviations of the formants F1 and F2 of all vowels per speaker. The red values represent the means and standard deviations of the first five minutes of the speech recording and the blue values represent the last 5 minutes of the recording.

From visual inspection of Fig. 1 it is clearly visible that the two Austrians (101M and 104M) and the two Germans (102M and 103M) are more similar. Also it is visible that the most experienced speaker (104M) makes the least change in vowel space over the course of speaking. Looking at the vowel space areas in Table III one can observe a clearly smaller vowel space area at the end for speakers 102M and 103M, whereas the vowel space areas for speakers 101M and 104M don't change significantly. In general, the vowels /o/ and /u/ differed the most between the start and the end across all speakers and in most cases the mean of the vowels moved towards the central vowel. Noticeable differences from this observation are vowel /u/ of speaker 101M and vowel /a/ of speaker 104M.

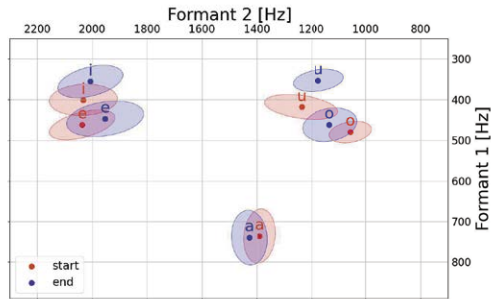
## V. GENERAL DISCUSSION AND CONCLUSION

The results of the analysis of the voice quality at the beginning and the end of the reading session showed that the addition of MFCCs for the classification does increase the F1 score either alone or in combination with the VQ features. The study of the vowel space indicates that the vowel space area decreases for three of the speakers, i.e. the vowels move to the center of the vowel space. This might be due to less accurate pronunciation that is rather related to cognitive than vocal fatigue. Even though this pilot study only contains four speakers, we want to point out that when using MFCCs or other features that represent the speech spectrum, the discriminating properties used might not be caused by vocal fatigue.

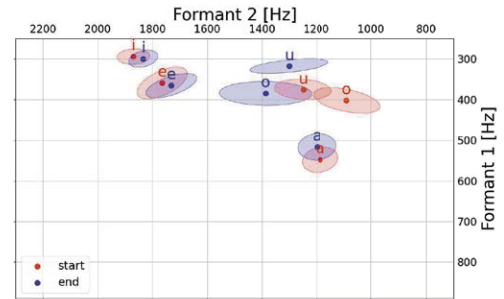
## VI. ACKNOWLEDGMENTS

We express our gratitude to the four speakers who participated in the long recording task.

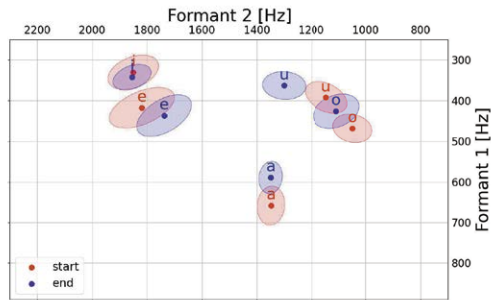




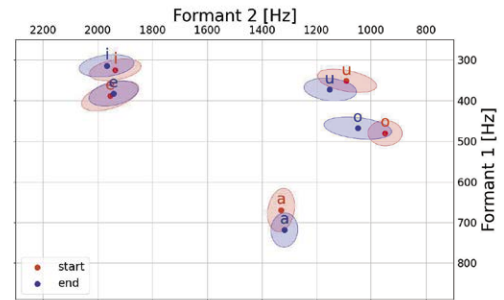
(a) Speaker 101M (Eastern Austria, little experience)



(b) Speaker 102M (Southern Germany, little experience)



(c) Speaker 103M (Southern Germany, little experience)



(d) Speaker 104M (Eastern Austria, more experience)

Fig. 1: Ellipses from combinations of the tense vowels /a/, /a:/, /e:/, /i:/, /i:/, /o/, /o:/, and /u/, /u:/ based on means and standard deviations of formants F1 and F2. Ellipses capture statistics of vowel positions for each speaker at the start (red) and the end (blue) of a recording. The speaker information provided in brackets specifies the location where the speakers have primarily resided and whether they are experienced presenters.

## REFERENCES

- [1] N. V. Welham and M. A. Maclagan, "Vocal fatigue: Current knowledge and future directions," *Journal of Voice*, vol. 17, no. 1, pp. 21–30, mar 2003.
- [2] C. Nanjundeswaran, B. H. Jacobson, J. Gartner-Schmidt, and K. V. Abbott, "Vocal fatigue index (VFI): Development and validation," *Journal of Voice*, vol. 29, no. 4, pp. 433–440, jul 2015.
- [3] P. Bottalico, I. I. Passione, A. Astolfi, A. Carullo, and E. J. Hunter, "Accuracy of the quantities measured by four vocal dosimeters and its uncertainty," *The Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. 1591–1602, mar 2018.
- [4] A. Remacle, C. Finck, A. Roche, and D. Morsomme, "Vocal impact of a prolonged reading task at two intensity levels: Objective measurements and subjective self-ratings," *Journal of Voice*, vol. 26, no. 4, pp. e177–e186, jul 2012.
- [5] M.-J. Caraty and C. Montacié, "Vocal fatigue induced by prolonged oral reading: Analysis and detection," *Computer Speech & Language*, vol. 28, no. 2, pp. 453–466, mar 2014.
- [6] S. P. Bayerl, D. Wagner, I. Baumann, T. Bocklet, and K. Riedhammer, "Detecting vocal fatigue with neural embeddings," *Journal of Voice*, feb 2023.
- [7] Z. K. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, vol. 10, pp. 122 136–122 158, 2022.
- [8] X. Gao, K. Ma *et al.*, "A rapid, non-invasive method for fatigue detection based on voice information," *Frontiers in Cell and Developmental Biology*, vol. 10, sep 2022.
- [9] H. P. Greeley, J. Berg *et al.*, "Fatigue estimation using voice analysis," *Behavior Research Methods*, vol. 39, no. 3, pp. 610–619, aug 2007.
- [10] M. W. J. Caverlé and A. P. Vogel, "Stability, reliability, and sensitivity of acoustic measures of vowel space: A comparison of vowel space area, formant centralization ratio, and vowel articulation index," *The Journal of the Acoustical Society of America*, vol. 148, no. 3, pp. 1436–1444, sep 2020.
- [11] B. Pfister and T. Kaufmann, *Sprachverarbeitung - Grundlagen und Methoden der Sprachsynthese und Spracherkennung, 2. Auflage*. Germany: Springer Vieweg, 2017.
- [12] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," Version 6.1.38, retrieved 2 January 2021, 2021. [Online]. Available: <http://www.praat.org/>
- [13] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer, Speech and Language*, vol. 45, no. C, pp. 326–347, 2017.
- [14] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015.
- [15] M. Paierl, T. Röck, S. Wepner, A. Kelterer, and B. Schuppler, "Creapy: A python-based tool for the detection of creak in conversational speech," in *Accepted for ICPHS2023*, 2023.
- [16] F. Pedregosa, G. Varoquaux *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [17] C. Barber, D. Dobkin, and H. Huhdanpaa, "The quick-hull algorithm for convex hulls," *ACM Trans. on Mathematical Software*, vol. 22, no. 4, pp. 469–483, dec 1996.

**SESSION II**  
**VOCAL FOLDS AND VOCAL TRACT**





# FORMANT ANALYSIS OF VOICE AS AN EARLY BIOMARKER OF NEURODEGENERATION

A. Nacci<sup>1</sup>, S. Capobianco<sup>1</sup>, F. Simoni<sup>2</sup>, L. Bruschini<sup>1</sup>, S. Berrettini<sup>1</sup>, L. Bastiani<sup>3</sup>

<sup>1</sup> Otolaryngology, Audiology and Phoniatries Unit, University of Pisa, Pisa, Italy

<sup>2</sup> IRCSS Ospedale Policlinico San Martino, Genoa, Italy

<sup>3</sup> Institute of Clinical Physiology of the Italian National Research Council, Pisa, Italy

[a.nacci@med.unipi.it](mailto:a.nacci@med.unipi.it); [silviacapobianco.md@gmail.com](mailto:silviacapobianco.md@gmail.com); [federicasimoni91.fs@gmail.com](mailto:federicasimoni91.fs@gmail.com); [luca.bruschini@unipi.it](mailto:luca.bruschini@unipi.it); [stefano.berrettini@unipi.it](mailto:stefano.berrettini@unipi.it); [luca.bastiani@ifc.cnr.it](mailto:luca.bastiani@ifc.cnr.it)

**Abstract:** Characterization of the first (F1) and second (F2) formants in vowels represents a reliable method to describe articulatory abilities. In patients affected by neurodegeneration tongue movement is reduced, with subsequent increase in F1 and simultaneous decrease in F2. To describe this phenomenon two parameters were used: Vowel Space Area (VSA) and Formant Centralization Ratio (FCR). 91 patients affected by neurodegenerative diseases and 174 non-dysarthric control subjects underwent voice analysis with formant characterization of the vowels /a/, /e/, /i/, /u/. By computing F1 and F2 for the vowel sounds, triangular (tVSA) and quadrangular (qVSA) VSA were obtained. Both tVSA and qVSA were shown to decrease significantly ( $p < 0.0001$ ) in dysarthric compared with non-dysarthric subjects, while FCR increased ( $p < 0.0001$ ). These changes correlated positively with dysarthria progression as described by the clinical Radbound Dysarthria Assessment (RDA), especially for the vowels /e/, /i/ and /u/. Both VSA and FCR statistically differed ( $p < 0.001$ ) between non-dysarthric and mildly dysarthric subjects, demonstrating their early alteration in disease onset. It is possible to suggest that characterization of F1 and F2 may be useful as an early biomarker of dysarthria and neurodegeneration and a possible biomarker of disease progression. **Keywords:** Neurodegeneration, Voice, Biomarkers, Acoustics, Dysarthria

Amyotrophic Lateral Sclerosis (ALS), more than 80% of patients are affected by dysarthria, which develops earlier in patients with a bulbar onset of disease, leading to anarthria in few months [10]. In 23% of ALS patients, dysarthria has been reported as the first predominant symptom in the early stage of disease, 8% more frequent than dysphagia [11]. Overall, acoustic analysis of voice may represent a good candidate as an early biomarker of neurodegeneration because it allows a transversal evaluation of neural control, it is impaired with specific patterns from the earliest stage of most NDD, it is objective, simple and non-invasive, sensible to subtle pre-clinical changes, and it can be acquired even remotely [12]. Most types of dysarthria are characterized by articulatory undershoot, as the intended place and degree of vocal tract constrictions are not fully achieved due to a reduced range of movements of articulatory movements [13] (i.e., reduced range of articulatory movements), to the extent that the intended place and degree of vocal tract constriction are not fully achieved [14]. To acoustically quantify the articulatory abilities of a subject vowel metrics are used, based on formant frequencies, which are spectral peaks produced by specific articulatory configurations of the vocal tract, with a particular role of the tongue body, whose excursion in height is inversely related the first formant (F1), and whose frontness is directly related to the second formant (F2). F1 and F2 are used to acoustically characterize different vowels, and by plotting their respective frequencies on orthogonal axes of a bi-dimensional F1-F2 plane, the Vowel Space Area (VSA) is created: the triangular VSA (tVSA) is made by the vowels /a/, /i/, and /u/, while by adding the vowel /e/ the quadrangular VSA (qVSA) results [15]. In 2010, Sapir and colleagues proposed an additional parameter to describe the articulatory range of the vocal tract, the Formant Centralization Ratio (FCR) [16], with the goal of maximizing the sensitivity of pathological vowel centralization when  $FCR > 1$  [17], and decreasing intersubject variability with respect to VSA-based parameters [18]. The overall reduction of working space for vowel articulation in dysarthric patients is likely to result in vowel formant centralization, i.e., formants that normally have high frequencies tend to

## I. INTRODUCTION

Dysarthria is an impairment of speech characterized by "abnormalities in the strength, speed, range, steadiness, tone, or accuracy of movements required for breathing, phonatory, resonatory, articulatory, or prosodic aspects of speech production due to damage of the central or peripheral nervous systems" [1]. In Parkinson's disease (PD) vocal impairments are present in up to 90% of patients [2], being reported as some of the earliest indicators of the disease, present years before the diagnosis [3-6]. Dysarthria is the most common expressive communication deficit in Multiple Sclerosis (MS) patients, with a prevalence of 45% [7-9]. In

have lower frequencies, and formants that normally have low frequencies tend to have higher frequencies. This phenomenon can be quantified by a decrease of tVSA and qVSA and an increase in FCR [14,19]. Such findings have been reported in a number of studies comparing acoustic formant-based parameters in patients affected by PD with matched healthy controls [20-26]. Acoustic analysis of voice and vowel metrics has been also applied to MS, detecting significant differences with healthy controls [27]. In patients affected by ALS a reduction in the VSA accounting for 45% of variance in speech intelligibility was described, with a faster and more prominent impact on bulbar-onset patients [10, 28, 29]. A number of studies have detected early changes in speech among highly intelligible individuals affected by ALS in the early stages of disease, however results and acoustic parameters considered were highly variable across different papers [30-32]. Although vowel metrics have been reported as potential biomarkers of dysarthria, most studies have focused on one pathology or few pathologies at the same time, the most studied disease being PD. The aim of this study is to investigate the role of acoustic analysis of voice to characterize dysarthria across a wider variety of diseases, with a focus on the potential carried by formant-based parameters as early biomarkers of neurodegeneration and as markers of disease progression and severity.

## II. METHODS

*Participants:* 265 subjects were included in the study. Among them, 91 (41 F, 50 M; mean age  $65.3 \pm 13.8$ ) were affected by neurodegenerative diseases (NDD) of both the Central (72) and the Peripheral (19) Nervous Systems. 174 non-NDD subjects represented the control group (99 F, 75 M; mean age  $52.9 \pm 16.34$ ), 87 of them being euphonic (54 F, 33 M; mean age  $49.8 \pm 17.5$ ), while 87 were perceptively dysphonic but non-dysarthric (45 F, 42 M; mean age  $56 \pm 14.54$ ). The dysphonic group underwent laryngostroboscopy to highlight functional or organic diseases of the glottis. Regarding the dysphonia perceptible grading, 49 subjects were mildly dysphonic, 27 moderately dysphonic, 11 severely dysphonic.

*Protocol:* All patients underwent a clinical evaluation of dysarthria severity according to the Radboud Dysarthria Assessment (RDA)[33], combining the different dimensions of speech production (articulation, resonance, phonation, respiration and prosody) to classify dysarthria as clinically absent ( $n^{\circ}6$ ), mild ( $n^{\circ}34$ ), moderate ( $n^{\circ}29$ ), and severe ( $n^{\circ}22$ ). In the control group of 174 non-NDD subjects (87 euphonic and 87 dysphonic), no clinically evident dysarthria was recorded. The vocal sign was recorded using a Kay Computer Speech Lab (CSL) 4500B supported by a personal computer including a Shure-Prolog SM48 microphone. Analysis of a voice sample was carried out using the 2.3 version of MDVP 5105

software. Each patient kept the vowel /a/, /e/, /i/, /u/ at a constant conversation intensity (range 55-65 dB) for at least 7 seconds; the central 4 seconds of each phonation were used for further analysis. The first and second formant frequencies for each vowel considered were extracted (F1a, F2a, F1e, F2e, F1i, F2i, F1u, F2u). The working range for vowel articulation can be visually assessed by the triangular and quadrangular Vowel Space Areas, which are constructed by the Euclidean distances between the F1 and F2 coordinates of the corner vowels /i/, /u/, and /a/ (tVSA), or the corner vowels /i/, /u/, /a/, and /e/ (qVSA) in the F1-F2 plane, according to the formulas [14,16]:

$$tVSA = 0.5 * [F1i(F2a-F2u) + F1a(F2u-F2i) + F1u(F2i-F2a)] \quad (1)$$

$$qVSA = 0.5 * [(F2iF1e + F2eF1a + F2aF1u + F2uF1i)(F1iF2e + F1eF2a + F1aF2u + F1uF2i)] \quad (2)$$

In 2010 Sapir and coworkers [16] introduced a frequency normalization parameter, the Formant Centralization Ratio (FCR), with the goal of reducing inter-subject variability [18] and maximizing the sensitivity of vowel centralization [17,] calculated according to the formula:

$$FCR = (F2u + F2a + F1i + F1u) / (F2i + F1a) \quad (3)$$

For each of the 265 subjects in this study, tVSA, qVSA and FCR were calculated. All statistical analyses were completed using SPSS Version 24 (SPSS, Chicago, Illinois) and significance was set at  $p < 0.05$ . For the age, gender, and for the clinical and instrumental evaluation (all vowel-formant elements, F1a, F2a, F1e, F2e, F1i, F2i, F1u, F2u and for tVSA, qVSA, FCR), the difference between groups was performed using parametric (independent T test or Anova) and non-parametric statistics (Mann-Whitney U or Kruskal-Wallis test). In both neurological and non-neurological groups parametric (Pearson) and non-parametric (Spearman) correlation coefficient was performed to evaluate the relation between age, triangular vowel space area (tVSA), quadrilateral vowel space area (qVSA) and formant centralization ratio (FCR).

## III. RESULTS

**3.1 Neuro vs Non-Neuro:** In the comparison between the control ( $n^{\circ}174$ ) and the study group ( $n^{\circ}91$ ), differences were found for the three composite indicators measured. For both tVSA (p-value 0.000) and qVSA (p-value 0.000) the means and medians were lower in the neurological group. Also considering FCR the value of the mean and the median in the neurological group was significantly different from the non-neurological group (p-value 0.000).

### 3.2 Euphonic vs Dysphonic vs Neurological:

Considering the whole sample (n°265) in the classification between euphonic (n°87), dysphonic (n°87) and neurological (n°91) subjects, and performing ANOVA analysis for tVSA, qVSA and FCR, statistically significant differences ( $p < 0.001$ ) were observed in the comparison of neurological subjects with both the euphonic group and the dysphonic group for all three parameters, tVSA, qVSA and FCR. On the contrary, the comparison between the euphonic and dysphonic subgroups of patients not affected by NDD did not highlight any statistically significant difference.

### 3.3 Dysarthria severity in the neurological group:

Evaluating only the sample of patients affected by NDD (n°91), classified according to the clinical evaluation of dysarthria severity by the Radbound Dysarthria Assessment (RDA) scale [33] in absent (n°6), mild (n°34) moderate (n°29) and severe (n°22), by performing ANOVA analysis for tVSA, differences were observed between moderate vs absent ( $p$ -value 0.05), and severe vs moderate ( $p$ -value 0.01). For qVSA, differences were observed only between severe vs absent ( $p$ -value 0.01), while for FCR, statistically significant differences were recorded between absent vs severe ( $p$ -value 0.01), and between mild vs severe ( $p$ -value 0.01).

**3.4 Mild-dysarthria vs non-neurological:** In the comparison between the non-neurological (n°174) and mild dysarthric patients (n°34), differences were found for all the three vowel metrics calculated. For both tVSA and qVSA the means and medians were lower among mildly dysarthric patients than non-NDD subjects ( $p$ -value 0.001). Also for the others two indices, qVSA and FCR, the values of the mean and the median among mildly dysarthric patients were significantly different from the non-neurological group (qVSA  $p$ -value 0.001/FCR  $p$ -value 0.001).

### 3.5 Correlation between age, tVSA and qVSA in non-neurological and neurological groups:

Evaluating the correlation between age, tVSA ( $p$ -value 0.021) and qVSA ( $p$ -value 0.001) both in the non-neurological and in the NDD group, a significant correlation was observed only in the non-neurological group, while in the NDD no correlation between age, tVSA and qVSA was recorded.

**3.6 Dysarthria type:** Evaluating by ANOVA analysis 89 patients affected by NDD presenting a clinical picture of dysarthria classified according to the Mayo perceptive classification system (Darley, Aronson and Brown)[1] in hypokinetic, spastic, flaccid and ataxic, no statistical significance was detected for tVSA and for qVSA between Hypokinetic (n°34), Spastic (n°31) Flaccid (n°18) and Ataxic (n°6). 2 patients presenting hyperkinetic kinetic dysarthria were removed from this

analysis as this class did not provide enough power to perform statistical comparisons. A statistical significant difference was recorded in FCR between the Flaccid and Ataxic groups ( $p$ -value  $< 0.05$ ).

## IV. DISCUSSION

The aim of this study was to investigate the role of formant-based vowel metrics (tVSA, qVSA and FCR) to characterize dysarthria across a wide variety of neurodegenerative diseases (NDD), by comparing a population of 91 subjects affected by NDD of both the CNS and PNS with a control group of 174 subjects (87 euphonic and 87 dysphonic), with a total of 265 subjects considered.

By comparing the study group with the control group as a whole (considering both dysphonic and euphonic subjects), it was possible to highlight a statistically significant reduction in tVSA ( $p < 0.0001$ ) and qVSA ( $p < 0.0001$ ), and an increase in FCR ( $p < 0.0001$ ) in the neurological group. These parameters have already been demonstrated sensible to differences in speech between healthy controls and patients affected by various NDD, especially Parkinson's [21-24,26], Multiple Sclerosis [24] and Amyotrophic Lateral Sclerosis [28]. Also in the present study, considering a more heterogeneous population comprising patients affected by a wide variety of NDD, tVSA, qVSA and FCR proved to be reliable parameters in differentiating dysarthric patients from healthy controls. Similar results were obtained when comparing the neurological group with the euphonic and the dysphonic groups separately. As expected, the comparison between the control subgroups (euphonic vs dysphonic) did not highlight any statistically significant difference. In fact, dysphonic patients present a wide variety of organic and/or functional diseases of the glottis (i.e. the voice source) but not at the acoustic filter represented by the vocal tract. Inversely, dysarthric patients, though euphonic in their glottal emission, present alterations in the filtering action of the vocal tract. By considering only the neurological group and performing ANOVA analysis comparing different severity grades of dysarthria, it was possible to highlight how tVSA, qVSA and FCR represent all reliable markers of disease progression, as reported also in other studies, considering mostly PD in a longitudinal setting [21,22]. When comparing mildly dysarthric patients with non-neurological patients, tVSA, qVSA and FCR showed statistically significant differences ( $p < 0.001$ ). Therefore, tVSA, qVSA and FCR may have a role as early markers of dysarthria and possibly of neurodegeneration, as previous studies have reported acoustic alterations in voice in the early stages of numerous neurodegenerative diseases [25,26,32]. While among patients affected by NDD VSA-based

parameters (tVSA and qVSA) were not affected by age, in the control group a significant correlation was found, as tVSA and qVSA significantly reduce as the articulatory abilities of subjects physiologically decrease with age. According to the Mayo classification system (Darley, Aronson and Brown) [1], perceptively dysarthria can be classified as spastic (n°31), ataxic (n°6), hypokinetic (n°34), hyperkinetic (n°2), flaccid, and mixed. When considering 89 dysarthric patients, no statistical differences were recorded in terms of tVSA, qVSA and FCR values with regard to different perceptive classes of dysarthria.

#### V. CONCLUSION

The aim of the present study was to characterize dysarthria across a wide variety of neurodegenerative diseases by means of formant-based parameters (tVSA, qVSA and FCR), which differed significantly between the study and the control group, in the former being also markers of disease severity. Moreover, these parameters significantly differentiated mildly dysarthric patients from non-dysarthric controls, possibly inferring their role as early biomarkers of neurodegeneration. Overall, it is possible to suggest that characterization of F1 and F2 may be useful as an early biomarker of dysarthria and neurodegeneration and a possible biomarker of disease progression.

#### REFERENCES

- [1] J.R. Duffy: *Motor speech disorders: Substrates, differential diagnosis, and management*. St. Louis, MO: Elsevier, 2013.
- [2] S.B.O. Sullivan, T.J. Schmitz, and G. Fulk: (2013). *Physical rehabilitation*, 5th ed. USA: FA Davis Company, 2013.
- [3] J.W. Tetrad: "Preclinical Parkinson's disease: detection of motor and nonmotor manifestations," *Neurology*, vol. 41, suppl. 2, pp. 69–71, 1991.
- [4] C. Stewart, L. Winfield, A. Hunt, S.B. Bressman, S. Fahn, A. Blitzer, and M.F. Brin: "Speech dysfunction in early Parkinson's disease". *Mov. Disord*, vol. 10, pp. 562-565, 1995.
- [5] R.B. Postuma, A.E. Lang, J.F. Gagnon, A. Pelletier, and J.Y. Montplaisir: "How does parkinsonism start? Prodromal parkinsonism motor changes in idiopathic REM sleep behaviour disorder," *Brain*, vol. 135, pp. 1860–1870, 2012.
- [6] S. Perez-Lloret, L. Nègre-Pagès, A. Ojero-Senard, et al: "Oro-buccal symptoms (dysphagia, dysarthria, and sialorrhea) in patients with Parkinson's disease: preliminary analysis from the French COPARK cohort," *Eur J Neurol*, vol.1, pp. 28-37, 2012.
- [7] D.G. Theodoros, B.E. Murdoch, and E.C. Ward: "Perceptual features of dysarthria in multiple sclerosis", in: *Speech and language disorders in multiple sclerosis*, London: Whurr Publishers, 2000.
- [8] G. Noffs, T. Perera, S.C. Kolbe, et al: "What speech can tell us: A systematic review of dysarthria characteristics in Multiple Sclerosis," *Autoimmun Rev*, vol.17, no.12, pp. 1202-1209, 2018.
- [9] J.E. Sussman, and K. Tjaden: "Perceptual measures of speech from individuals with Parkinson's disease and multiple sclerosis: intelligibility and beyond," *J Speech Lang Hear Res*, vol.55, no.4, pp.1208-1219, 2012.
- [10] B. Tomik, and R.J. Guiloff: "Dysarthria in amyotrophic lateral sclerosis: A review," *Amyotroph Lateral Scler*, vol.11, no. 1-2, pp. 4-15, 2010.
- [11] B.J. Traynor, M.B. Codd, B. Corr, et al: "Clinical features of amyotrophic lateral sclerosis according to the El Escorial and Airlie House diagnostic criteria: A population-based study," *Arch Neurol*, vol. 57, no. 8, pp. 1171-1176, 2000
- [12] M. Magee, D. Copland, and A.P. Vogel: "Motor speech and non-motor language endophenotypes of Parkinson's disease," *Expert Review of Neurotherapeutics*, 2019.
- [13] G. Weismer, J.Y. Jeng, J.S. Laues, R.D. Kent, and J.F. Kent: "Acoustic and intelligibility characteristics of sentence production of neurogenic speech disorders," *Folia Phoniatr Logop*, vol.53, pp.1–18, 2001.
- [14] R. Kent, and Y. Kim: "Toward an acoustic typology of motor speech disorders," *Clin. Linguist. Phonetics*, vol. 17, pp.427–445, 2003.
- [15] M.J. Ball, and M. Muller: "Phonetics for Communication Disorders," Psychology Press, 2014.
- [16] S. Sapir, L.O. Raming, J.L. Spielman, and C. Fox: "Formant Centralization Ratio: A Proposal for a New Acoustic Measure of Dysarthric Speech," *J Speech Lang Hear Res*, vol.53, no.1, pp. 114-125, 2010.
- [17] F.L. Lansfors, and J.M. Liss: "Vowel Acoustics in Dysarthria: Mapping to Perception," *J Speech Lang Hear Res*, vol.57, no. 1, pp. 68-80, 2014.
- [18] P. Adank, R. Smits, and R. van Hout R: "A comparison of vowel normalization procedures for language variation research," *Journal of the Acoustical Society of America*, vol. 116, no. 5, pp. 3099-3107, 2004.
- [19] S. Sapir, J.L. Spielman, L.O. Ramig, B.H. Stroy, and C. Fox: "Effects of intensive voice treatment (the Lee Silverman Voice Treatment [LSVT]) on vowel articulation in dysarthric individuals with idiopathic Parkinson disease: Acoustic and perceptual findings," *J. Speech Lang. Hear. Res*, vol. 50, pp. 899–912, 2007.
- [20] S. Skodda, W. Visser, and U. Schlegel: "Vowel articulation in Parkinson's disease," *J Voice*, vol. 25, no. 4, pp. 467-472, 2011.
- [21] S. Skodda, W. Grönheit, and U. Schlegel: "Impairment of vowel articulation as a possible marker of disease progression in Parkinson's disease," *PLoS One*, vol. 7, no. 2, 2012.
- [22] S. Skodda, W. Grönheit, N. Mancinelli, and U. Schlegel: "Progression of voice and speech impairment in the course of Parkinson's disease: a longitudinal study," *Parkinsons Dis*, 2013.
- [23] D. Mirarchi, P. Vizza, G. Tradigo, N. et al: "Signal Analysis for Voice Evaluation in Parkinson's Disease," in *Proc of the IEEE International Conference on Healthcare Informatics (ICHI)*, 2017.
- [24] P. Vizza, G. Tradigo, D. Mirarchi, et al: (2019). "Methodologies of speech analysis for neurodegenerative diseases evaluation," *Int J Med Inform*, vol. 122, pp. 45-54.
- [25] M. Duranovic, N. Salihovic, A. Ibrahimagic, and N. Toromanovic: "Characteristics of voice in individuals with multiple sclerosis," *Materia Socio-Medica*, vol. 23, no. 1, p. 23, 2011.
- [26] J. Ruzs, R. Cmejla, H. Ruzickova, E. Ruzicka: "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease," *J Acoust Soc Am*, vol. 129, no. 1, pp. 350-367, 2011.
- [27] J. Ruzs, R. Cmejla, T. Tykalova, et al: "Imprecise vowel articulation as a potential early marker of Parkinson's disease: effect of speaking task," *J Acoust Soc Am*, vol.134, no.2, pp.2171-2181, 2013.
- [28] G.S. Turner, K. Tjaden, and G. Weismer: "The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis," *J Speech Hear Res*, vol. 38, pp. 1001-13, 1995.
- [29] G.S. Turner, and K. Tjaden: "Acoustic differences between content and function words in amyotrophic lateral sclerosis," *J Speech Lang Hear Res*, vol. 43, pp. 769-81, 2000.
- [30] B. Tomik, J. Krupinski, L.B. Glodzik-Sobanska, M. ala-Slodowska, W. Wszolek, M. Kusiak, et al: "Acoustic analysis of dysarthria profile in ALS patients," *J Neurol Sci*, vol. 169, pp.35-42, 1999.
- [31] D. Robert, J. Pouget, A. Giovanni, J.P. Azulay, and J.M. Triglia: "Quantitative voice analysis in the assessment of bulbar involvement in amyotrophic lateral sclerosis," *Acta Otolaryngol*, vol. 119, pp. 724-31, 1999.
- [32] A.K. Silbergleit, A.F. Johnson, and B.H. Jacobson: "Acoustic analysis of voice in individuals with amyotrophic lateral sclerosis and perceptually normal vocal quality," *J Voice*, vol.11, pp. 222-31, 1997.
- [33] S. Knuijt, J.G. Kalf, B.M.G. van Engelen BGM, et al: "The Radboud Dysarthria Assessment: Development and Clinimetric Evaluation," *Folia Phoniatr Logop*, vol.69, no.4, pp. 143-153, 2017

# VOCAL FOLD IMPACT STRESS: THE KEY CONCEPT OF PHONOTRAUMA

P. H. DeJonckere<sup>1</sup>, J. Lebacqz<sup>2</sup>

<sup>1</sup> Federal Agency for Occupational Risks, Brussels, Belgium

<sup>2</sup> Institute of Neurosciences, University of Louvain, Brussels, Belgium

[Ph.DeJonckere@outlook.com](mailto:Ph.DeJonckere@outlook.com) ; [Jean.Lebacqz@uclouvain.be](mailto:Jean.Lebacqz@uclouvain.be)

**Abstract:** Mechanical impact stress on the vocal fold surface, particularly when excessive, has been postulated to cause the so-called phonotraumatic tissue lesions, such as nodules and polyps. The collision stress between the vocal folds is a function of the vocal fold velocity at the time of impact. Combining a precise photometric measurement of glottal area and simultaneous measurements of translaryngeal impedance (electroglottogram) for identifying the time of the maximum rate of increase of vocal fold contact allows computing the vocal fold collision speed in a wide range of loudnesses. The vocal fold collision speed is - for modal voicing - always smaller than the maximum vocal fold velocity during the closing phase, but it strongly increases with intensity. Moreover, this increase shows a biphasic pattern, with a significant enhancement from around 78 dB on.

**Keywords:** Vocal fold collision – Voice intensity – Glottal area – EGG.

## I. INTRODUCTION

Mechanical impact stress on the vocal fold (VF) surface, particularly when excessive, has been postulated to cause the so-called phonotraumatic tissue lesions [1]. VF nodules and polyps are the best-known examples [2]. The maximum area declination rate (MADR) in the closing phase of the glottis during VFs' vibration has been reported as a measure of the impact stress loading the VFs during collision [4], thus, as a relevant parameter when considering biomechanical economy of phonation [5]. The collision stress between the VFs can be estimated from basic physical principles [6]: Assuming the mass of a tissue element at the medial surface of the VF edge to be

$$m = \rho \Delta x \Delta y \Delta z \quad (1)$$

where  $\rho$  is tissue density (1040 kg/m<sup>3</sup>) and  $\Delta x \Delta y \Delta z$  is a small volume, then, from Newton's second law, the average collision force over an impact interval  $\Delta t$  is

$$F = m \Delta v / \Delta t \quad (2)$$

where  $\Delta v$  is the change in velocity during impact. Jiang & Titze [7] estimated the impact interval to be of the order of

$$\Delta t = T_0 / 10 \quad (3)$$

where  $T_0$  is the fundamental period. The velocity change in Eq. (2) can be estimated by assuming a sinusoidal motion of amplitude  $A$  and radian frequency  $\omega = 2\pi F_0$ . The maximum velocity, which occurs near impact, is

$$\Delta v = \omega A = 2\pi F_0 A \quad (4)$$

This velocity is reduced to zero during the collision interval, such that

$$v = v - 0 = 2\pi F_0 A \quad (5)$$

Substituting (1), (3), and (5) in (2),

$$F = 20\pi A F_0^2 \rho \Delta x \Delta y \Delta z \quad (6)$$

when the impact starts from the phase when the VF velocity is at its maximum value. If  $\Delta y \Delta z$  is the impact surface and  $\Delta x$  the depth of the vibrating tissue, then the collision stress is

$$\sigma = F / \Delta y \Delta z = 20\pi A F_0^2 \rho \Delta x \quad (7)$$

To set these ideas on an example corresponding with modal male speech, for an amplitude of vibration of  $10^{-3}$  m, a depth of vibration of  $10^{-3}$  m, and a  $F_0$  of 120 Hz, the stress is 9.4 hPa.

Direct measurements on human subjects also yielded a range of 1–5 kPa.

Eq. 7 shows that collision stress increases with  $F_0^2$ ; however, both amplitude and depth of vibration are expected to decrease with  $F_0$ , making the exact stress uncertain. At habitual speaking frequency, when  $F_0$  remains within a limited range, the velocity change

during impact is the essential parameter (Eq. 2). As velocity is reduced to zero during the collision interval, the peak velocity and the moment at which this peak occurs with respect to the impact are major determinants. Titze hypothesized that the maximum velocity occurs near impact [4]. This is indeed what can be expected in the case of a closed quotient of 0.5, i.e., when the closed phase and the open phase have the same duration. It is interesting to know what happens when the closed quotient departs from 0.5.

Hence it is relevant to more precisely know the actual VFs velocity at the time they collide, and the relationship between impact velocity and voicing intensity.

It would indeed be desirable to weigh increases in vocal loudness against increases in tissue stress to obtain a cost and/or benefit ratio for certain vocal productions like teaching, acting or speaking in public, and consequently to define the level absolutely requiring electrical amplification.

## II. METHODS

### Glottal area and sound signal

The glottal area was derived from a photometric record obtained by transilluminating the trachea, as described in previous work [8,9]. The light flux was detected by a nondirectional photovoltaic transducer positioned as dorsally as possible in the pharynx (photoglottography PGG).

The photoglottographic signals are more accurate than those provided by image processing from high-speed video; the high sampling frequency allowing adequate time resolution for computation of the derivative much better than that of the imaging techniques.

By inspecting still stroboscopic pictures at the time of maximal opening, we found that the contour of the glottal image could be well fitted with an ellipse, the major and the minor axes of which were the ventrodorsal length and the maximal width of the glottis picture respectively.

The sound signal, measured in  $V_{rms}$ , was first calibrated in dB by recording series of short (~ 4–5 s) voice utterances at stable SPL (controlled by visual feedback) at intervals of 5 dB, from ~ 55 dB on.

The correlation coefficient between the maximum glottal width and the SPL is 0.98. As the length of the glottis is constant, maximal width and glottal length were used to calculate the area (in  $mm^2$ ) of the equivalent ellipse by applying the simple geometrical equation of the area of the ellipse, and this value is equivalent to the maximum value of the PGG signal

during each cycle, or 100% of the glottal area. All values of the PGG signal, expressed in % of the maximum area, were then transformed into units of area; from area, the half width of the glottis was finally calculated by the equation of the ellipse. Similarly, all values of the derivative of the PGG signal were expressed in rate of change of area (in  $mm^2/s$ ), from which the speed of the edge of each VF (in m/s) was obtained.

This approach makes it easily possible to calculate the speed of each VF edge at its middle length, with the assumption that, in normal conditions, VFs are vibrating approximately symmetrically.

### Translaryngeal impedance (Electroglottography EGG)

The EGG-signal, used as a reference for monitoring the contact surface changes, was detected using a portable electroglottograph (Laryngograph Ltd, London, UK) Model EG90. As for the photoglottogram, the very high sampling frequency makes it possible to accurately compute the derivative. The positive peak of the EGG-derivative indicates the maximum rate of increase in VF contact. [10].

Time delays due to electronic circuitries were measured and the necessary corrections applied: 0.102 ms for PGG and 0.056 ms for EGG).

### Vocal material

A corpus of about 140 recordings was created with short sustained vocal emissions on /ə/ with the photoglottograph in situ, and simultaneous EGG and sound monitoring, at spontaneous speaking pitch ( $F_0$  between 95 and 125 Hz) in a large range of loudnesses, Out of this corpus, 32 records were suited for detailed analysis (criterion: full display of all traces in the central part of the recording).

## III. RESULTS & DISCUSSION

Fig.1 shows an example of an original raw tracing with the three signals: PGG (glottal area), EGG (translaryngeal electrical impedance) and microphone signal. Intensity is moderate (70.24 dB), as is the closed quotient (0.35).

The horizontal axis is time (ref. = 2 ms). The y axis represents the calibrated glottal area (increasing upwards, ref. = 10  $mm^2$ ) for the PGG, the translaryngeal electrical impedance (decreasing upwards) for the EGG and the acoustic pressure (microphone). Fundamental frequency is about 115 Hz, and intensity 70.24 dB.

An estimate of the maximum glottal area during one cycle of a sustained phonation (28.9  $mm^2$ ) is obtained by using videokymography and



videostroboscopy in similar voicing conditions and in the same subject.



Fig. 1

Fig. 2 is as Figure 1 but at a higher intensity (82.90 dB) with a larger closed quotient (0.58). Fundamental frequency is about 117 Hz. The estimate of the maximum glottal area during one cycle of a sustained phonation is here 46.3 mm<sup>2</sup>.

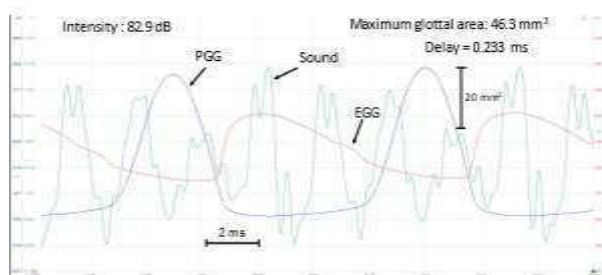


Fig. 2

The most relevant information is the actual velocity of one single VF (m/s) at the time of collision (cf. Eq.2). This velocity depends on (1) the maximum velocity, given by the magnitude of the negative PGG peak, and (2) the extent of the reduction (in %) of this maximum velocity during the interval between the time this maximum closing velocity is reached and the VF collision peak. The extent of the deceleration depends itself on this delay, but also on the shape of the closing phase, which is not linear and involves, at its terminal phase, aspects like tissue compression and deformation, i.e., what occurs between the first contact of the VF and the maximum rate of increase in VF contact when VF collide, which has been considered as

the collision peak. At low voicing intensities, the collision velocity is not more than 5%–25 % of the maximum velocity, while at higher intensities the percentage promptly becomes 25%–70%.

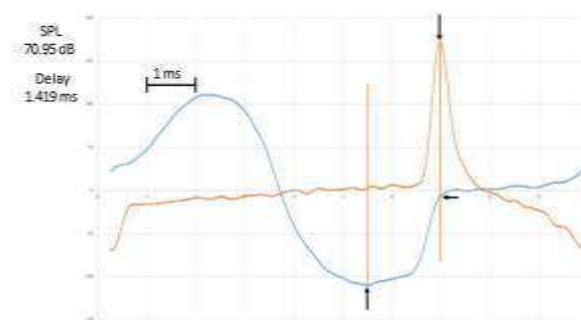


Fig. 3

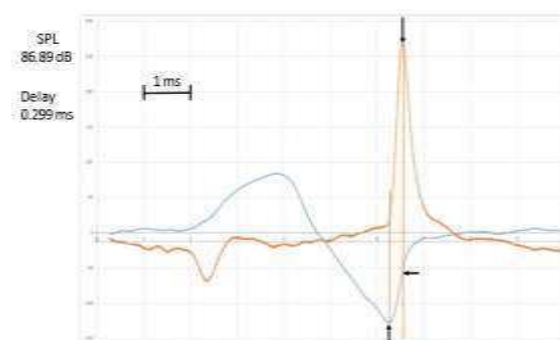


Fig. 4

Examples of the first derivatives of the PGG (dPGG/dt) and EGG (dEGG/dt) signals, calculated from two original tracings corrected for the respective time delays of 0.102 ms and 0.056 ms can be compared in Figures 3 and 4. The vertical arrows indicate the positive peak for EGG (i.e., the max. rate of increase in VF contact when VF collide, considered as the collision peak) and the negative peak for PGG (i.e., the max. glottal closing velocity). At the lower intensity (70.95 dB) (Figure 3), the delay between the two peaks is quite large (1.419 ms). At the higher intensity (86,89 dB) (Figure 4), the delay between the two peaks becomes much shorter (0.188 ms). When a vertical straight line is drawn through the peak of the EGG-derivative, it can be seen that the time of this peak (collision peak) corresponds in Figure 3 (horizontal arrow) to a value of 6.9% of the maximum negative amplitude of the PGG-derivative (= max. glottal



closing velocity), giving a calculated VF velocity of 0.04 m/s. Similarly, in Figure 4, the peak of the EGG-derivative (collision) corresponds (horizontal arrow) to 44.2% of the maximum negative amplitude of the PGG-derivative and to a calculated VF velocity of 1.02 m/s.

To oversimplify, if we compare the oscillation pattern of the VF edge with a sinusoidal motion clipped at half-height (one-mass model), clipping begins at the time of maximum velocity. This is the equivalent of a closed quotient of 0.5. In a situation where VFs freely oscillate without making contact, there is no glottal closure, the closed quotient is 0 and the velocity is also 0 at the time the VF edges are closest to each other. Between these extremes, every percentage (0%–100%) of the maximum velocity at the moment of clipping (contact) is possible.

The actual velocity of one single VF at the collision peak is plotted as a function of intensity (dB) in Figure 5. Globally, the velocity at the time of impact clearly increases with intensity. Yet the relation is not linear, and two different patterns can be identified: from about 78 dB on, the regression slope becomes substantially steeper, even if the correlation is strong and highly significant both in the range 65–78 dB and in the range 78–87dB. Covariance analysis demonstrates that the difference in slope between the two regression lines is highly significant ( $t = -3.3029$ ;  $p = 0.0026$ ).

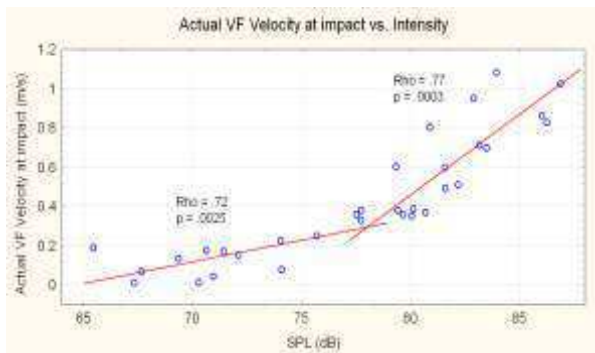


Fig. 5

## CONCLUSION

At modal speaking pitch, the actual VF collision velocity is significantly lower than the maximum closing velocity, and the extent of the deceleration effect strongly depends on the intensity of voicing. Moreover, the relationship between collision velocity and intensity shows a biphasic shape: the deceleration (“braking”) effect is increasingly reduced at loud voicing, from about 78 dB on. Hence the MADR in the closing phase of the glottis during VFs’ vibration may not be considered as a measure of the impact stress

loading the VFs during collision. Mechanical stress has been considered as the key to the etiology of VF nodules [7]. In depth understanding of physiological variables that influence VF collision forces provides relevant insight into the pathophysiology and the prevention of voice disorders associated with phonotraumatic vocal hyperfunction

## REFERENCES

- [1]. Li Z, Bakhshae H, Helou L, et al. Evaluation of contact pressure in human vocal folds during phonation using high-speed video-endoscopic, electroglottography, and magnetic resonance imaging. *Proc Meet Acoust (Acoustical Society of America)*. 2013;19:1–8. <https://doi.org/10.1121/1.4800732>. 060306.
- [2]. DeJonckere P H, Kob M. Pathogenesis of vocal fold nodules: new insights from a modelling approach. *Folia Phoniatr Logop*. 2009;61: 171–179.
- [3]. Horáček J, Laukkanen A M, Sidlof P, et al. Comparison of acceleration and impact stress as possible loading factors in phonation: a computer modeling study. *Folia Phoniatr Logop*. 2009; 61:137–145.
- [4]. Titze I R. Theoretical analysis of maximum flow declination rate versus maximum area declination rate in phonation. *J Speech Lang Hear Res*. 2006 ;49: 439–447.
- [5]. Titze I R, Laukkanen AM. (2007) Can vocal economy in phonation be increased with an artificially lengthened vocal tract? A computer modeling study. *Logoped Phoniatr Vocol*. 2007;32:147– 156.
- [6]. Titze I R. Mechanical stress in phonation. *J Voice*. 1994;8:99–105.
- [7]. Jiang J J, Titze I R. Measurement of vocal fold intraglottal pressure and impact stress. *J Voice*. 1994;8:132–144.
- [8]. DeJonckere P H, Lebacqz J. In Vivo quantification of the intraglottal pressure: modal phonation and voice onset. 2019 *J Voice*. 2020;34:645. e19–645.e39. <https://doi.org/10.1016/j.jvoice.2019.01.001>. Epub 2019 Jan 16.
- [9]. DeJonckere P H, Lebacqz J, Titze I R. Dynamics of the driving force during the normal vocal fold vibration cycle. *J Voice*. 2017;31:649– 661.
- [10]. Sarvaiya J N, Pandey P C, Pandey V K. An impedance detector for glottography. *IETE J Res*. 2011;55:100–105.

# FORMANT TRAJECTORIES IN DIFFERENT LANGUAGES

V. V. Evdokimova<sup>1</sup>, M. R. Maximova<sup>2</sup>

<sup>1</sup> Saint Petersburg State University/Department of Phonetics, Saint-Petersburg, Russian Federation

<sup>2</sup> Saint Petersburg State University/Department of Phonetics, Saint-Petersburg, Russian Federation  
v.evdokimova@spbu.ru, st076821@student.spbu.ru

**Abstract:** This paper compares formant transitions in English, French and German monophthongs after labial and lingual (tip, front and back) consonants. The monophthongs were extracted from speech samples read by female speakers. The formant values were measured in 9 points within each vowel. The line graphs representing formant transitions were plotted. The derivatives of the resulting functions at 9 points were calculated to compare the slopes of different formant trajectories. The formant values were converted from Hertz to Bark scale in order to evaluate the formant structure variability. F2 values were the highest in German vowels and the lowest in English monophthongs. However, English vowels demonstrated the highest F1 values. Differences in the steepness of the first transition region were observed in French rounded and unrounded monophthongs. The region in which the greatest slope occurred was the same for English and German monophthongs, but different for French vowels. The highest variability of formant structure was observed in F1 of English vowels. Overall, the formant structure of German vowels was the least variable. In English and German, formant structure variability was lower in the back monophthongs compared to the front vowels. However, French monophthongs showed the opposite tendency.

**Keywords:** Speech acoustics, phonetics, formant transitions, Bark scale, formant slopes

## I. INTRODUCTION

The quality of a speech sound is determined by the movement of articulators. Another important factor is the regions of the vocal tract that influence the sound spectrum. The spectrum is formed as a sound passes through the upper vocal tract (i.e., the oral cavity and the nasal cavity). The cavities of the upper vocal tract perform as resonators increasing those sound frequencies that are equal or close to natural resonating frequencies of the resonators. The increased frequencies are referred to as formants. A natural

resonating frequency of a cavity is determined by its shape, which can be changed by certain movements of articulators located in this cavity [2].

Since 1950s, there has been a number of studies that have reported a crucial role of the first two formants in vowel quality. The value of the first formant (F1) correlates with vowel height, while the second formant (F2) correlates with vowel backness. These parameters are essential for vowel discrimination and, consequently, for automatic speech recognition [2].

Beside formant values, formant trajectories (or, formant transitions) are also important for vowel discrimination [3]. In connected speech, vowels rarely occur in isolation. Therefore, the acoustic features of vowels can change depending on the characteristics of neighboring consonants or vowels. Vowel quality is primarily affected by the loci of neighboring consonants and by secondary consonantal articulations (e.g., palatalization, labialization, etc.) that modify the shape of the vocal tract [5].

Watson and Harrington [9] argue that vowels can also be characterized by the slope of their formant trajectories. This feature is determined by the distance between the consonantal locus and vowel target and, consequently, by the formant values of the consonant and the vowel. However, likewise mean formant values, slopes cannot possibly be used as a single parameter for vowel recognition as the slopes of most monophthongs are similar.

The aim of this study is to investigate formant trajectories in English, French and German monophthongs after labial and lingual (tip, front and back) consonants. This paper compares formant transitions of vowels in different languages in similar consonantal contexts.

## II. METHODS

The monophthongs in labial and lingual (tip, front, and back) consonantal contexts were obtained from sentences, words and texts read by female speakers. Only female informants were considered for the reason that more female speakers were found than male ones. Male voices were not included to the data for this

research as male formant values are, on average, lower than formant values in female speakers.

The English speech samples were taken from the LUCID corpus (London UCL Clear Speech in Interaction Database), designed by Baker and Hazan [1], and also from the audio recordings for the manual «MyGrammarLab: Advanced C1/C2» [4]. The French data was drawn from the audio application for the manual «EDITO C1 Méthode de français» [7]. The German recordings were gathered from the audio course which was a part of the manual «Wir-3» [6].

The combinations of 12 English, 10 French and 16 German phonemes with labial, tip, front and back consonants were selected for analysis. In total, 100 samples per vowel in each context were examined. Diphthongs and nasal vowels were not included in this set for the reason that diphthongs are not presented in the phoneme inventory of French, while nasal vowels do not exist as phonemes in English and German.

To plot formant trajectories, F1 and F2 values were measured in 9 points within each vowel using a script from the SpeCT (The Speech Corpus Toolkit for Praat). After that, the derivatives of the resulting functions were calculated in order to compare the slopes of formant transitions in the three languages. To evaluate formant structure variability, the formant values were converted from Hz to Bark.

The formant trajectories in English, French and German monophthongs were compared on the basis of shape, slope, and the variability of formant structure.

### III. RESULTS

#### A. The shapes of formant transitions

To illustrate the formant transitions that are characteristic of each vowel in all the three languages in each of the contexts, averaged plots were designed. The formant values in the plotted trajectories were averaged over all the samples of the given vowel in the given context.

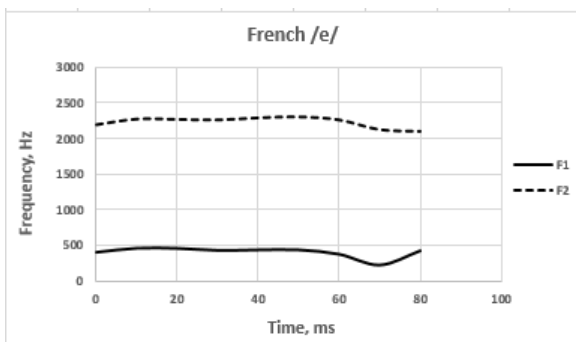


Fig. 1: An averaged plot of the formant trajectories of the French vowel /e/ after labial consonants.

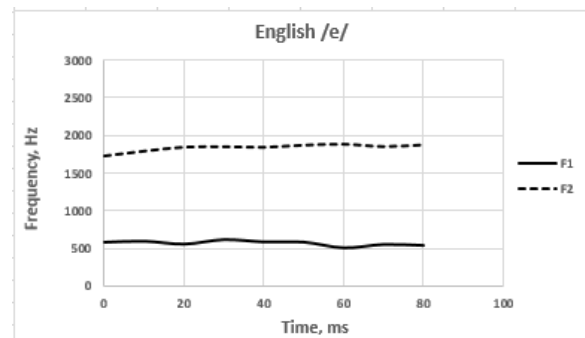


Fig. 2: An averaged plot of the formant trajectories of the English vowel /e/ after labial consonants.

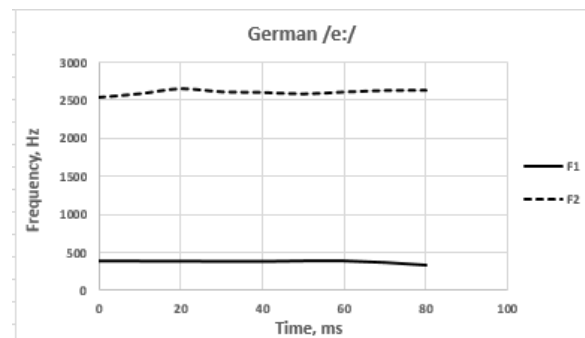


Fig. 3: An averaged plot of the formant trajectories of the German vowel /e:/ after labial consonants.

The results indicated that formant transitions differed the most after nasal consonants. Interestingly, formant values were observed to decrease as well as to increase in nasal context, although formant values generally fall when surrounded by nasal consonants [5]. F2 values in German vowels in each context were higher than in the other languages. Relative to German and French monophthongs, F2 values in English vowels were the lowest. Nevertheless, F1 values were as a rule the greatest in English monophthongs (see Fig. 1, Fig. 2, Fig. 3).

Regarding French vowels, a difference was observed between rounded and unrounded phonemes. French unrounded monophthongs were characterized by less significant formant movements in the first transition region compared to English and German unrounded vowels. However, French rounded vowels showed more steep contours of the first transition region in comparison with rounded monophthongs of the other languages.

#### B. The slopes of formant trajectories

To calculate the derivatives of F1 and F2 functions, trend lines were constructed in Excel. The roots of the trend line equations were then identified. These values

were plotted in order to illustrate the rate of change of the formant values.

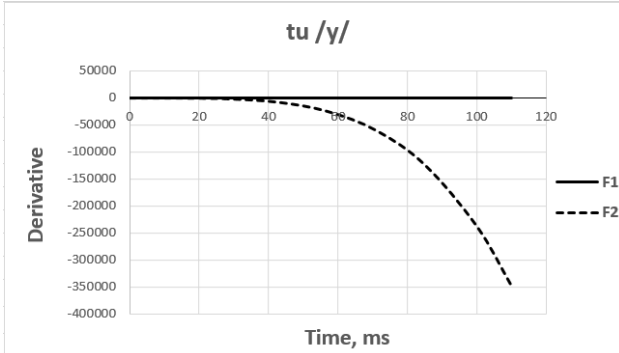


Fig. 4: The slopes of F1 and F2 trajectories of the French vowel /y/.

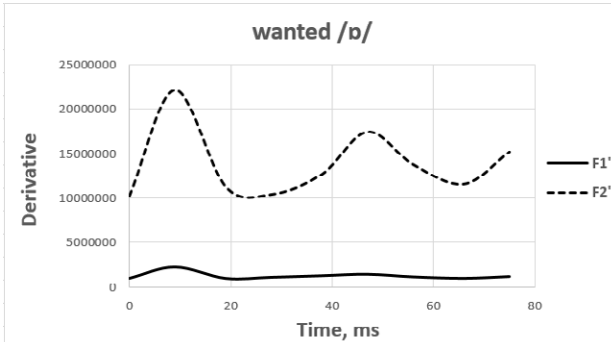


Fig. 5: The slopes of F1 and F2 trajectories of the English vowel /ɒ/.

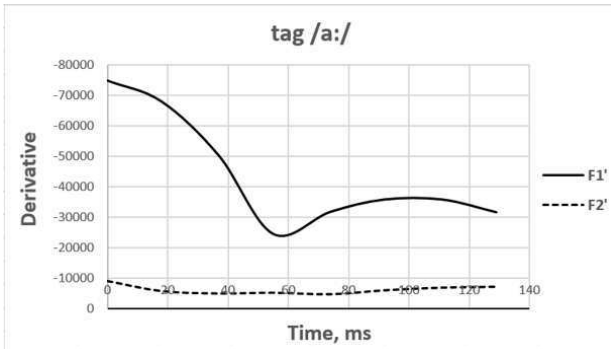


Fig. 6: The slopes of F1 and F2 trajectories of the German vowel /a:/.

The results revealed that F2 slopes were generally higher than F1 slopes in every language (see Fig. 4 and Fig. 5). However, the languages differed from each other to a certain extent. Firstly, in English and German, higher slopes were observed in back vowels in all the contexts (see Fig. 5 and Fig. 6). On the contrary, in French, front close vowels showed higher slopes. The highest slopes in this language occurred in vowel center or near the end of a formant trajectory (see Fig. 4). In English and German vowels, by

contrast, the highest slopes occurred more often in the transition region from a preceding consonant to a vowel (see Fig. 5 and Fig. 6).

### C. Formant structure variability

In order to evaluate the formant structure variability in the monophthongs, the formant values calculated before in Hz were converted into values on the psychoacoustic Bark scale. The following formula was used [8]:

$$6 * \sinh^{-1} \left( \frac{\text{Hz}}{600} \right)^{40} \quad (1)$$

Table 1: Averaged F1 and F2 values of the French vowel /ɔ/ in labial context given in Hz and Bark.

French /ɔ/			
F1		F2	
Hz	Bark	Hz	Bark
694	5.9	1208	8.7
689	5.9	1204	8.7
689	5.9	1216	8.7
684	5.9	1270	9.0
688	5.9	1280	9.0
721	6.1	1380	9.4
717	6.1	1377	9.4
724	6.1	1397	9.5
723	6.1	1329	9.2

Table 2: Averaged F1 and F2 values of the English vowel /ɒ/ in labial context given in Hz and Bark.

English /ɒ/			
F1		F2	
Hz	Bark	Hz	Bark
490	4.5	1086	8.1
418	3.9	1091	8.2
582	5.2	1078	8.1
573	5.1	1094	8.2
579	5.1	1105	8.2
538	4.8	1119	8.3
446	4.1	1140	8.4
425	4.0	1144	8.4
555	5.0	1128	8.3

Table 3: Averaged F1 and F2 values of the French vowel /ɔ/ in labial context given in Hz and Bark.

German /ɔ/			
F1		F2	
Hz	Bark	Hz	Bark
598	5.3	1013	7.8
646	5.6	1045	7.9
692	5.9	1082	8.1
754	6.3	1099	8.2
735	6.2	1174	8.5
744	6.2	1283	9.0
739	6.2	1326	9.2
730	6.2	1396	9.5
698	6.0	1346	9.3

Formant variability was the most significant in F1 of English vowels. Overall, F1 variability was higher than that of F2. Contrary to English monophthongs, German vowels demonstrated the lowest variability.

In addition, English and German back vowels showed less formant variability in comparison with front vowels in these languages. The opposite tendency was observed in French monophthongs. The formant structure variability increased from front to back French vowels.

#### V. CONCLUSION

This study has shown that both similarities and differences could be observed in the formant trajectories of English, French and German vowels. Monophthongs differ from each other not only in terms of formant transitions, but also in terms of formant structure variability.

As it is known from many theoretical works, acoustic features of vowels are determined by their articulatory characteristics [2, 6]. Some findings of this research illustrate this correlation. For example, the formant values of back vowels were more variable than those of front vowels in French. Similarly, the steepness of the first transition regions varied between French rounded and unrounded vowels. It was also revealed that F1 values are overall more variable than F2 values. This implies that tongue position in oral cavity from back to front changes more frequently than tongue location in the vertical plane. However, the slopes of F2 trajectories appeared to be higher than those of F1 trajectories. This indicates that vowel height generally changes more significantly within monophthongs than backness does.

#### REFERENCES

- [1] Baker, Rachel & Hazan, Valerie. LUCID: A corpus of spontaneous and read clear speech in British English. Proc. DiSS-LPSS Joint Workshop (DiSS 2010), pp. 3-6, 2005.
- [2] Bondarko L.V.: Osnovy obshhej fonetiki: uchebnoe posobie dlya studentov lingvisticheskikh i filologicheskikh spetsial'nostej [Basics of phonetics — a textbook for students of linguistics and philology]. Moscow: Akademiya, 2004.
- [3] Broad, D. J., & Clermont, F. Target-locus scaling methods for modeling families of formant trajectories. *Journal of Phonetics*, 38, pp. 337–359, 2010.
- [4] Foley M., Hall D. MyGrammarLab. Advanced. C1-C2. Book with key and MyEnglishLab. Harlow: Pearson Education Limited, 2012.
- [5] Kodzasov S.V., Krivnova O.F. Obshchaya fonetika [General phonetics]. Moscow: Russian State University for Humanities Press, 2001.
- [6] Motta, G.: Wir 3 Lehrerbuch. Stuttgart: Klett, 2007.
- [7] Pinson C., Bourmaysan A., Cros I. Edito. Méthode de français. C1 - Livre + DVD-Rom. Paris: Didier, 2018.
- [8] Wang, S., Sekey, A. and Gersho, A. “An objective measure for predicting subjective quality of speech coders”. *IEEE Journal on Selected Areas in Communications*, 10 (5), pp. 819–829, 1992.
- [9] Watson, C. I., and Harrington, J. Acoustic evidence for dynamic formant trajectories in Australian English vowels. *Journal of the Acoustical Society of America* 106, pp. 458–468, 1999.

**SESSION III**  
**SINGING, DRAMA AND VOICE QUALITY**



# THE ROLE OF PHONiatrICS IN THE DRAMA THEATRE ACADEMIES: CLINICAL OBSERVATIONS

G. Baracca<sup>1</sup>, S. Capobianco<sup>2</sup>, L. Bastiani<sup>3</sup>, A. Nacci<sup>2</sup>

<sup>1</sup> Conservatory Cesare Pollini, Padova, Italy; Accademia Teatrale Veneta, Venezia, Italy

<sup>2</sup> ENT Audiology Phoniatric Unit, University of Pisa, Pisa, Italy

<sup>3</sup> CNR Institute of Clinical Physiology, Pisa, Italy.

[giovanna.baracca@gmail.com](mailto:giovanna.baracca@gmail.com), [silviacapobianco.md@gmail.com](mailto:silviacapobianco.md@gmail.com), [a.nacci@med.unipi.it](mailto:a.nacci@med.unipi.it), [luca.bastiani@ifc.cnr.it](mailto:luca.bastiani@ifc.cnr.it)

**Abstract:** Voice professionals constitute more than 40% of phoniatric patients. Among them there is a peculiar group of patients composed of actors, specialized in speaking voice.

As in other groups of voice users, vocal folds problems can lead the actors to have emotional consequences, technical impairment and occupational difficulties. Recent studies underscore the fact that functional and inflammatory vocal fold diseases have higher prevalence in acting students than in the general population. The data collected in some of the most important drama schools in Italy confirm the higher risk for voice impairment in acting students.

Close collaboration is therefore required between theater schools and voice specialists, as phoniatricians and voice therapists. This is necessary to increase awareness of healthy voice use in young acting students and to recognize and manage the presence of voice problems.

**Keywords:** acting voice, speaking voice, dysphonia

## I. INTRODUCTION

Theatre actors constitute a specific population of artists, whose characteristics in terms of voice disorders have not still been fully investigated.

They are generally included in the group of professional voice users by authors dealing with occupational voice, but they constitute a small percentage of the materials inherent in these studies. As all professional voice users, actors are a population at risk for voice disorders [1-4]. Indeed, even a slight dysphonia in this group of workers can have serious professional consequences that produce occupational, emotional and morale problems. This unique population requires a specific voice usage and a strong vocal demand: for example, they must adjust their voice production to theatres of varying size or outdoor

stages, while maintaining the ability to express the entire range of human emotions, their performance is at times characterized by extreme vocal behaviors and sudden emotional outbursts like screaming, shouting, crying and sobbing. Furthermore, they are expected to portray various characters (eg young or old, healthy or unhealthy, loud and aggressive or soft spoken) to meet the artistic demands of their role and have to combine voice projected emissions [1]. The actor adjusts his or her voice in order to produce the required vocal quality authentically and may possibly introduce damaging effects to the vocal mechanism. If actors constitute a population at risk for vocal disorders, a drama student must learn to pay attention to his or her voice. In drama schools, students receive an intensive training in voice technique, singing, acting voice, physical activity (about 8 hours per day), and are expected to participate in stages, rehearsals and full performances. They often participate in highly demanding vocal activities before they acquire the required knowledge and techniques to preserve their voices. Students undergo an extensive vocal loading before they can learn the necessary awareness of their vocal health [5]. Consequently, it would be recommendable to assess their vocal status to preserve voice health conditions. We present some clinical data collected in professional acting schools.

## II. METHODS

All data are collected by students attending the first year of a professional drama academy. This group of subjects is composed of 56 students (27 males and 29 females), between the ages of 20 to 27 years. We do not consider any exclusion criteria. The control group is composed of 60 subjects (26 males and 34 females), between the ages of 21 and 44 years, without voice problems.

All subjects underwent voice assessment:



- Interview regarding voice use, vocal disturbances, exposition to risk factors (smoking, laryngopharyngeal reflux symptoms, allergies, immune or hormonal problems)
- GRB evaluation from the GRBAS scale by 3 independent judges (1 voice teacher, 2 phoniatricians and/or voice therapists)
- Videolaryngostroboscopic recordings evaluated by 2 independent judges (classification of the vocal folds abnormalities in: presence of masses, inflammatory aspects, dysfunctional aspects and presence of scars)

*Statistical analysis is still in progress*

### III. PRELIMINARY OBSERVATIONS

#### A. Voice interview

In the study group we found 33 smokers corresponding to 59% of the subjects, 20 students with laryngopharyngeal reflux symptoms (17 with mild symptoms corresponding to 31% of the sample and 3 with moderate symptoms corresponding to 5% of the sample), 9 subjects with pollinosis corresponding to 16% of the sample. Nobody reported immune or hormonal problems.

#### B. GRB perceptual evaluation

Concerning GRB perceptive evaluation in the study group, we observed that:

- G parameter: there was no dysphonia in 50% of the students, mild dysphonia in 46.4% of the students, moderate dysphonia in 3.6% of the students
- R parameter: there was no roughness in 64.3% of the voices, mild roughness in 32.1% of the voices, moderate roughness in 3.6% of the voices
- B parameter: there was no breathiness in 71.4% of the voices, mild breathiness in 28.6% of the voices

#### C. Videolaryngostroboscopic evaluation

We observed a prevalence of vocal folds abnormalities in 37 students, corresponding to 68.5% of the study group. More in detail, 18 cases were categorized as presence of masses on the vocal folds, 14 students showed inflammatory aspects like edema and hyperemia, 13 students showed dysfunctional characteristics such as glottal incompetence, 1 student had a scar. Some students showed more than one

pathological aspect. Data analyses and comparison with the control group are in progress.

### IV. DISCUSSION

Actors are considered a professional category at risk for voice problems, even they are not deeply investigated. The most common vocal symptoms reported by actors include hoarseness, voice breaks, vocal weakness and fatigue [1,2], increased effort during phonation, difficulties in producing high-pitch tones and reduction in pitch range [6], physical complaints including shortness of breath, dry throat, laryngeal discomfort, strain, pain and physical tension [1,7]. Laryngeal findings in actors include altered vocal folds vibratory pattern, decreased mucosal wave, vocal fold edema and abnormal vascularity patterns [8], non infective laryngitis, asthenicity, nodules and upper respiratory infections [9].

Acoustic analysis of actors' voices is characterized by high perturbances values and high noise-to-harmonic ratio values [8]. The high prevalence of voice disturbances in actors leads to consider drama students as a population that needs voice education and assessment. Our preliminary clinical observations in drama academies are in agreement with data from the literature. We confirm that frequently acting students have poor hygiene habits: they often smoke too much and they tend to have poor eating habits eliciting reflux problems [10]. Furthermore, they are used to perform or watch performances in the evenings and do not have a regular rhythm for eating or sleeping. They use to shout and speak excessively outside the school environment, because they do not consider ordinary vocal activity as being of the same vocal folds. They could underestimate the risk of vocal damage compared to singers, because the voice is considered merely one of the instruments inherent to acting. The prevalence of alterations of the vocal folds observed through the videolaryngostroboscopic examination could be related to the lack of awareness of voice cure and attention. The most important point is that theatre actors are at risk for developing vocal disorders from a young age. In addition, due to the dependance of these performers on their vocal quality and capacity, an education regarding vocal hygiene and vocal training starting from the first year of drama school should be recommended.

### V. CONCLUSION

We observed a high prevalence of vocal problems during the first year of attendance in professional drama schools, with a reduced self-awareness of healthy vocal use. Voice assessment and education

should start at the beginning of the theater school to prevent vocal damage and acquire healthy voice conditions.

#### REFERENCES

- [1] M.Z. Lerner, S. Paskhover, L. Acton, N. Young, "Voice disorders in actors", *J Voice*, vol. 27(6), pp. 705-708, 2013.
- [2] A. Novak, O. Dlouha, B. Capkova, M. Vohradnik, "Voice fatigue after theater performance in actors", *Folia Phoniatr (Basel)*, vol. 43(2), pp. 74-78, 1991.
- [3] S.V. Stager, S.A. Bielamowicz, J.R. Regnell, A. Gupta, J.M. Barkmeier, "Supraglottic activity: evidence of vocal hyperfunction or laryngeal articulation?", *J Speech Lung Hear Res*, vol. 43, pp.229-238, 2000.
- [4] E. D'haeseleer, F. Quintyn, I. Kissel, T. Papeleu, I. Meerschman, S. Claeys, K. Van Lierde, "Vocal Quality, Symptoms, and Habits in Musical Theater Actors", *J Voice*, vol. 36(2), pp. 292.e1-292.e9, 2022.
- [5] O. Amir, A. Primov-Fever, T. Kushnir, O. Kandelshine-Waldman, M. Wolf, "Evaluating voice characteristics of first-year acting students in Israel: factor analysis", *J Voice*, vol. 27(1), pp 68-77, 2013.
- [6] B.N. Raphael, R.C. Scherer, "Voice modifications of stage actors: acoustic analyses", *J Voice*, vol. 1, pp. 83-87, 1987.
- [7] E. D'haeseleer, I. Meerschman, S. Claeys, C. Leyns, J. Daelman, K. Van Lierde, "Vocal quality in theater actors", *J Voice*, vol. 31(4), pp. 510.e7-510.e14, 2017.
- [8] B. Hoffman-Ruddy, J. Lehman, C. Crandell, D. Ingram, C. Sapienza, "Laryngostroboscopic, acoustic, and environmental characteristics of high-risk vocal performers", *J Voice*, vol. 15(4), pp. 543-552, 2001.
- [9] N.A. Punt, "Applied laryngology to singers and actors", *J Laryngol Otol Suppl*, vol. 6, pp.1-24, 1983.
- [10] B. Timmermans, M.S. De Bodt, F.L. Wuyts, A. Boudewijns, G. Clement, A. Peeters, P.H. Van de Heyning, "Poor voice quality in future elite vocal performers and professional voice users", *J Voice*, vol. 16, pp. 372-382, 2002.



# STUDIO REPORT: RESEARCH ON THE ACOUSTICS AND CREATIVE TECHNOLOGIES OF THE SINGING VOICE IN LABMAT (LABORATORY OF MUSIC ACOUSTICS AND TECHNOLOGY, DEPARTMENT OF MUSIC STUDIES, NKUA)

Anastasia Georgaki<sup>1</sup>, Areti Andreopoulou<sup>2</sup>

<sup>1</sup> Music Department, National and Kapodistrian University of Athens, Greece

<sup>2</sup> Music Department, National and Kapodistrian University of Athens, Greece

[georgaki@music.uoa.gr](mailto:georgaki@music.uoa.gr)

[aandreo@music.uoa.gr](mailto:aandreo@music.uoa.gr)

**Abstract:** Since 2016, the Laboratory of Music acoustics and Technology (LabMAT) at the Department of Music Studies, NKUA has carried various research and projects on the acoustic analysis of the singing voice applied in computational acoustic musicology as also in the fields of vocal performance, digital creation, and pedagogy through the development of new technologies related to room acoustics. This studio report presents the research carried on the analysis of the Byzantine singing, the Cretan singing idiomatic technique, the somatosensory education of singers through a multimodal system, the phonological aspects of singing in different languages, the education of choral singing using XR technologies, and the ASMA project.

**Keywords:** singing voice, vocal performance, choral singing, XR technology, vocal pedagogy.

## I. ACOUSTIC ANALYSIS OF THE SINGING VOICE IN SYSTEMATIC MUSICOLOGY: EXPLORING DIFFERENT SINGING STYLES THROUGH PRAAT

The richness of the different singing styles that exist in the Greek music heritage has led us to better understand the voice in different singing performance practices, having as a starting point musicological problems that remain to be solved. In order to structure the common conception of the elements of historical musical reconstruction, it is necessary to understand the acoustic cues of different signing styles that exist mostly in the Greek Territory, or even to reconstruct lost models of Prosody used in drama performance.

In these preliminary studies we analyze the performance acoustic parameters of experts in singing, in order to compose a database with the melodic contours, ornamentation, timber quality (formants), and voice positioning, and to explore the micromelismatic characteristics of the Greek singing voice.

### A. Analysis of the Prosodic style in ancient Greek tragic poetry

From an historical point of view, we have been interested in listening to the reconstructed ancient Greek prosody in tragic poetry by analyzing different recitation of experts in order to understand the intonation, rhyme, and articulation/phonological techniques which are believed to work in the Erasmian context. [18]

### B. Exploring the performance modes of Byzantine Chant

Following this historical there have also been studies on the singing performance style in Byzantine hymnology, by examining the microtonality system in different modes (systaltikon, diastaltikon, isychastikon). This work is realized through data extraction from extended measurements and recordings of Byzantine chant. [16]

### C. Exploring Cretan singing

The particularities of Cretan singing have been approached through computational ethnomusicology methods, focusing on the analysis of the idiomatic singing style of the Cretan ritzitiko song, by analyzing the vocal characteristics of various singers through Praat and formant tuning. [15][14]

## II. MUSIC TECHNOLOGY FOR THE VOCAL PEDAGOGY: FROM ELEMENTARY SCHOOL TO THE OPERA SINGER

The development for digital technologies assisting/supporting vocal pedagogy of lyrical singers but also young children has been a focal research point for LabMAT over the past ten years. This work has been realized through extended singing voice recordings, the exploration of visual feedback, visualization technologies, and interactive

technologies, as a means of assessment helping users better understand pitch accuracy, timber quality, and intonation, in order to improve under the informed guidance of their teachers.

#### *A. Analytical approaches for the pedagogy and performance of classical singers*

Research is being carried on the development of a prototype system which lead to a multifaceted approach to lyrical singing pedagogy, through the collection and analysis of biofeedback from lyrical singers (respiration, vocal cords, positioning of the vocal cavity and articulation) through the use of different somatosensory singing technologies

The somatosensory pedagogy of the voice of lyrical singers concerns with a newly designed multi-sensor recording prototype for operatic singing which employs sensors that record acoustic and electroglottographic data, breathing kinetic actions, and data regarding pertinent postural and body movement behavior. It was recently utilized for the recording of 28 operatic singers in a controlled experiment setup. [3][4][5]

Research is also being carried on the phonological analysis of the lyrical singing voice in different registers and different languages. Currently, the focus lies on examines the degree to which the substantial knowledge of a foreign language (French) can assist a Greek-speaking classical singer in performing authentically in it. [13][10]

#### *B. Developing visual feedback technologies for the vocal pedagogy in primary education*

We have also developed different techniques for vocal pedagogy to young students in primary school through the ASMA project. ASMA (Assistance for students in Singing and Music Aesthetics) was a nationally funded research project which proposed through a theoretical/practical substantiation and development interactive applications, the support of singing instruction to elementary school music teachers, using applied scientific approaches and digital tools of visual feedback. ASMA also proposed solutions to problems related to the signing voice quality (timber breathiness, nasality) as also singing skills of primary education students.

Teaching singing in primary school is important to be seen through the lens of voice applied science and technology in order to help students understand and control better the mechanisms of their voice, improve their performance, correct their errors, and better express themselves in different singing styles. The main outcomes of ASMA include an interactive voice

guide with exercises, and tools that improve timbre, pitch accuracy, tempo, and other vocal qualities.[1] [2]

### III. DEVELOPING CREATIVE TECHNOLOGIES FOR EXTENDED VOCALITY IN ANCIENT GREEK DRAMA

The field of interactive technologies in the performing arts is increasingly drawing attention, both from the perspective of directors and the performers. When thinking of interactive /creative technology research input on vocal pedagogy, questions like, how can new technologies provide tools to increase the expressiveness of the voice or how can new technologies assist performers improvise and dynamically change the final outcome, arise.

#### *A. Drama tools*

Under this scope, LabMAT has been concerned with the development of interactive tools used in the context of ancient Greek Drama and Prosodic recitation. The designed Drama tools are based on the rules of Prosody and on the theories of ancient Greek music. Thus, they transform individual elements of the ancient Greek language and transposed ancient music theories, such as the curve of “logodes melos” of Aristoxenus, into an interactive process.[17]

#### *B. Kinesthesia tool*

Kinestisia presents a novel human-centered gestural system for vocal improvisation in drama, “Kinesthesia”, to be used in new opera and musical theatre, and redraws the relationship between music composition, gesture, and programming. The system takes advantage of multimodal interaction techniques through the invisible interface (signal processing), which examine the use of electroacoustic techniques in the human voice. It explores the use of live or recorded, digitally processed voice as a sound source for the development of music cues for playback through a multi-channel speaker system. [12]

### IV. ACOUSTIC STUDY OF THE SINGING VOICE RELATED TO ROOM ACOUSTICS AND EXTENDED REALITY

The importance of room acoustics on accurate vocal performance simulations, the impact of the acoustic conditions on a singer’s (amateur or professional) performance, as well as the possibility of leveraging from extended reality technology educational tools for vocal training are also of interest to LabMAT. On that end there has been extensive research conducted on the directivity characteristics and vocal projection qualities

of the Greek singing voice, as well as on the development of immersive, individualized, acoustically accurate applications supporting vocal training.

*A. Analytical examination of the spherical directivity characteristics and formant analysis of the Greek singing voice*

This work is part of a larger study investigating the sound projection and directivity characteristics of a wide variety of traditional Greek musical instruments and professional singers, performing in various common Greek music genres and in realistic performance scenarios, i.e., in places where musicians would be expected to perform and/or be recorded. Vocal directivity and projection analysis is based on data collected from professional and amateur singers as well as children in various singing styles (classical, byzantine, modern) in the Greek language. Unlike previous works focusing mainly on the horizontal plane, this study reports results on four elevation angles (+90°, +30°, 0°, and -30°), captured using a 29 semi-spherical microphone array. The collected data consists of short song excerpts and vowel sounds at different pitches.

*B. Studying the impact of room acoustic conditions on the singers' performance quality, using immersive audio and extended reality (XR) techniques.*

The broader goal of this work is to understand the impact of on-stage acoustic impression on the performers' musicality and performance quality. On-stage acoustic conditions vary among performance spaces, and, more often than not, between the latter and rehearsal spaces. As a result, musicians develop certain strategies to overcome difficulties arising during a performance due to the said acoustic mismatches. Virtual and Augmented reality technology has been suggested as a means for studying the effects of room acoustics and on-stage acoustic impressions on one's performance, and as a tool for helping performers adapt to the acoustic conditions of a performance hall without being physically present in it. Stemming from this work, LabMAT is focusing on the development and assessment of a tool aiming to virtually place users in various spots within a virtual choir on a virtual stage, by augmenting audio recordings with auditory spatialization and room-acoustic cues.

#### REFERENCES

- [1] A. Andreopoulou, N. Kotsani, G. Dedousis, and A. Georgaki, "Evaluating the vocal characteristics of elementary school students: basic assessment tools and methodology," in *Interaction Design and Children*, in IDC '21. New York, NY, USA: Association for Computing Machinery, Jun. 2021, pp. 216–223. doi: 10.1145/3459990.3460720.
- [2] E. Angelakis, A. Andreopoulou, and A. Georgaki, "Multisensory biofeedback: Promoting the recessive somatosensory control in operatic singing pedagogy," *Biomedical Signal Processing and Control*, vol. 66, p. 102400, Apr. 2021, doi: 10.1016/j.bspc.2020.102400.
- [3] E. Angelakis and A. Georgaki, "Towards a Somatosensory Training Digital Environment for Lyric Singing Pedagogy," in *Models and Analysis of Vocal Emissions for Biomedical Applications: 10th International Workshop*, Firenze, Italy, Dec. 2019, p. 51.
- [4] E. Angelakis, A. Georgaki, and P. Velianitis, "'Match Your Own Voice!': A Software Tool to Assist Singing Practice on the Somatosensory Motivation," in *PEVOC12*, Ghent, Belgium, 2017.
- [5] E. Angelakis, G. Kosteletos, A. Andreopoulou, and A. Georgaki, "Development and Evaluation of an Audio Signal Processing Educational Tool to Support Somatosensory Singing Control," presented at the Audio Engineering Society Convention 145, Audio Engineering Society, Oct. 2018. Accessed: Jul. 03, 2023. [Online]. Available: <https://www.aes.org/e-lib/browse.cfm?elib=19842>
- [6] E. Angelakis, N. Kotsani, and A. Georgaki, "Towards a Singing Voice Multi-Sensor Analysis Tool: System Design, and Assessment Based on Vocal Breathiness," *Sensors*, vol. 21, no. 23, p. 8006, 2021.
- [7] E. Angelakis, P. Velianitis, A. Andreopoulou, and A. Georgaki, "'Match Your Own Voice!': An Educational Tool for Vocal Training," presented at the Audio Engineering Society Convention 143, New York: Audio Engineering Society, Oct. 2017. Accessed: Jul. 03, 2023. [Online]. Available: <https://www.aes.org/e-lib/browse.cfm?elib=19323>
- [8] K. Bakogiannis, G. Dedousis, Y. Malafis, and A. Andreopoulou, "On the spherical directivity and formant analysis of the singing voice; a case study of professional singers in Greek Classical and Byzantine music," presented at the Audio Engineering Society Convention 153, Audio Engineering Society, Oct. 2022. Accessed: Jun. 21, 2023. [Online]. Available: <https://www.aes.org/e-lib/browse.cfm?elib=21960>
- [9] G. Dedousis, A. Andreopoulou, and A. Georgaki, "The Impact of Room Acoustics on Choristers' Performance: From Rehearsal Space to Concert Hall,"

presented at the 5th Stockholm Music Acoustic Conference, Stockholm, Sweden, Jun. 2023.

[10] N. Kotsani, E. Angelakis, and A. Georgaki, "Evaluating the nasalization of the singing voice," in *Models and Analysis of Vocal Emissions for Biomedical Applications: 12th International Workshop*, Firenze, Italy: Firenze University Press, Dec. 2021, p. 119.

[11] N. Kotsani, G. Dedousis, E. Angelakis, A. Andreopoulou, and A. Georgaki, "The ASMA Tool-Suite: Augmenting singing instruction of elementary school students," presented at the *Dictionary for Multidisciplinary Music Integration*, Trento, Italy, Nov. 2022.

[12] T. Moussas, N. Kotsani, and A. Georgaki, "Kinesthesia: An Interactive Voice Software for Intertextual Improvisation in Theatre Performance | Zenodo," in *Proceedings of the 19th Sound and Music Computing Conference*, Saint-Etienne, France, Jun. 2022. Accessed: Jul. 03, 2023. [Online]. Available: <https://zenodo.org/record/6797798>

[13] G. Papadimitriou, A. Andreopoulou, and A. Georgaki, "Preliminary acoustic analysis of articulation differences in spoken and sung French language by Greek classical singers," presented at the 5th Stockholm Music Acoustic Conference, Stockholm, Sweden, Jun. 2023.

[14] S. Kalozakis, A. Georgaki, G. Kouroupetroglou (2021) "*Formant Tuning In Cretan Rizitiko Singing*". 12th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications MAVEBA 2021. December 14th– 16th, 2021, (page: 127)

[15] S. Kalozakis-A. Georgaki: (2018) "*Acoustical Characteristics And Vocal Timbre Nuances Of The Cretan Rizitika*", Fifth International Conference On Analytical Approaches To World Music, Thessaloniki, 2018

[16] Georgaki A. Chaldaiakis, Tzevelekos P. (2013). Parameterization of the Byzantine Chant Ethos through Acoustic Analysis: from theory to praxis, in *SMAC 2013 in Stockholm Music Acoustics conference 2013 proceedings*, Stockholm

[17] G. Petras, P. Tsangarakis, A. Georgaki (2019): Extended Drama Prosodic Tools: Design and Aesthetics in *International Journal of Music Science, Technology and Art (IJMSTA.com)*

[18] A. Georgaki, S. Psaroudakes, M. Carle , PanagiotisTzevelekos : "Prosody Model of Attic tragic poetry : from Logos to mousike", in *SMC09 Proceedings*, University of Porto, Porto 2009

# ON THE TRANSMISSIBILITY OF SPECTRAL DECAY RATE VOICE QUALITY PARAMETERS TO CONSONANT VOICING

W. Wokurek<sup>1</sup>, M. Pützer<sup>2</sup>

<sup>1</sup> Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Germany

<sup>2</sup> FR Sprachwissenschaft und Sprachtechnologie, Universität des Saarlandes, Saarbrücken, Germany  
[wokurek@ims.uni-stuttgart.de](mailto:wokurek@ims.uni-stuttgart.de), [puetzer@coli.uni-saarland.de](mailto:puetzer@coli.uni-saarland.de)

**Abstract:** Phonatory quality of consonant voicing is compared to that of the surrounding vowels (i.e. VCV context) using voice quality parameters (VQPs). The parameters are based on the decay rate of the short term spectrum. The results indicate that (i) the analysis procedure is also appropriate to analyze voiced consonant segments and that (ii) the direction of the VQP vector changes about 15 degrees when the phonatory quality stays the same as for consonant surrounding vowels. E.g. phonatory quality of the nasal consonant /n/ in comparison to that of the vowel /a/. Statistical analysis shows no significant change in the spectral gradient VQPs. However, the relative bandwidth of the first formant VQP IC changes significantly. Leaving out IC from the VQP vector definition only slightly reduces some of the angles by one degree.

**Keywords:** Voice quality parameters, phonatory quality of consonant voicing and vowels

## I. INTRODUCTION

Voice quality is a suprasegmental property of voiced sounds, usually attributed to prosody beside intensity and fundamental frequency (pitch). A neutral, relaxed voice is named modal. Deviations from that may contain the psycho-acoustic dimensions like roughness, breathiness, hoarseness. This holds for healthy and pathological voice qualities. These dimensions are usually rated by experienced listeners using ordinal rating scale protocols with four levels of expression (none, low, mid, high). Voice quality parameters (VQPs) are sets of numeric quantities, computed from the speech recording by signal processing. An aim of many works is to predict the perceptual assignments by instrumental analytical results (i.e. VQPs) or at least find reasonable correlates. In contrast to that, the present study extends a method that was developed for vowels, to consonants. In particular, whether and how the parameters change between the consonant and the two adjacent vowels. This is a pilot study with recordings from a single male speaker and it focuses on modal phonatory quality of the nasal consonant /n/ and the vowel /a/.

## II. METHODS

VQPs have been defined on the sound pressure signal [1] or on the electroglottogram (EGG) [2], in time domain [2] or in frequency domain [1]. Frequency domain VQPs are used here.

### A. Voice Quality Parameters

The voice quality parameters (VQP) were originally defined in [1]. These voice quality parameters have been shown to be noise robust [3]. The version used here modifies the amplitude ratios (decibel differences, e.g. H1-A1) to spectral decay rates (decibels per octave) by division by the frequency ratios ( $\text{ld } F_{1p} / \text{ld } F_0$ ).  $F_{1p}$  denotes the frequency of the harmonic peak nearest to  $F_1$ . This modification was introduced to make the voice quality parameters less dependent of  $F_0$  changes. It is indicated by a G (for gradient) appended to the original parameter name.

The harmonic peak amplitudes used in the voice quality parameter definitions include inverse filtering to reduce the influence of voice quality, since a microphone in front of the speaker is used. Stevens compensated only for the most prominent neighboring formants, and did not include their bandwidths. Here linear prediction estimates of the first four formants including their bandwidths are used to compensate for articulation by subtracting the transfer function estimate from all used spectral peak amplitudes. This is indicated by the trailing i (for inverse filtered) in the parameter names.

Tab. 1: Voice Quality Parameters Names

OQGi	Open Quotient	(H1i-H2i,T0Gi)
GOGi	Glottal Opening	(T1Gi)
SKGi	Skewness	(T2Gi)
RCGi	Rate of Closure	(T3Gi)
T4Gi	Triangle to the 4 <sup>th</sup> formant	
IC	Incompleteness of Closure	(B1/F1)



Tab. 1 shows the abbreviations, the classical names of the voice quality parameters used, and some hints. These classical names arouse interest that can hardly be verified when a large number of speakers (some hundred) is analyzed. More neutral names based on the geometric construction of the gradient parameters are proposed in the third column. OQGi (T0Gi) is special because it involves the first two harmonics, that are always separated by an octave. Hence, the decay triangle matches the original definition.

Incompleteness of Closure IC is not defined by a spectral decay triangle but by the normalized bandwidth of the first formant. The amount of glottal opening introduces energy loss of the first formant oscillation to the subglottal system and increases its bandwidth.

### B. Vector of Voice Quality Parameters, Angle

To enable angle measurements between averaged voice quality parameters of sound segments, the vector of voice quality parameters  $v$  is defined in (1).

$$v = (\text{OQGi}, \text{GOGi}, \text{SKGi}, \text{RCGi}, \text{T4Gi}, \text{IC})^* \quad (1)$$

All vector coordinates are normalized quantities and therefore dimensionless. Since the aim of this vector is averaging and angle measurements an euclidean metric is used. The angle  $\alpha_{12}$  between the vectors  $v_1$  and  $v_2$  results from (2).

$$\alpha_{12} = 180/\pi * \arccos( |v_1|/|v_1| * |v_2|/|v_2| ) \quad (2)$$

The  $*$  in the argument of the arcus cosine is an inner product and  $|v|$  is the length (2-norm) of the vector  $v$ .

### C. Speech Analysis

The computation of the VQP is automatized with ESPS programs and PERL scripts [4]

Since the VQPs are based on harmonic peak amplitudes and their frequencies, the procedure starts with a short time spectrum using a window that is long enough to effectively contain two or more fundamental periods of the voiced segments (i.e. a 25.6 millisecond long hamming window using `fft`).

A fundamental frequency estimate, and the probability of voicing is obtained by `get_f0`.

By means of `formant`, linear prediction is used for estimates of the first 4 formants, i.e. their center frequencies and bandwidths.

### D. Inverse Filtering

The influence (i.e. the magnitude of the transfer function) of a single formant with center frequency  $F$  and bandwidth  $B$  on the amplitude of the source signal at frequency  $f$  is modeled by (3), Fant, Stevens.

$$\text{Formant}(f, F, B) = \frac{(F^2 + (B/2)^2)}{\sqrt{((f-F)^2 + (B/2)^2) * ((f+F)^2 + (B/2)^2)}} \quad (3)$$

The inverse filtering on a harmonic peak measured at frequency  $f$  (usually near an integer multiple of  $F_0$ ) in decibels  $H$  is done by subtracting the result of (3) converted to decibels for all 4 formants.

### E. Speech Material

The CV-syllable repetitions *'tatata/*, *'dadada/*, *'nanana/* utterances were selected to be recorded with rough, modal and breathy voice quality to study the VQPs. Only the production one modal *'nanana/* is analyzed in this pilot study. Of particular interest was (i) whether the VQPs are applicable in the voiced segments of the consonant and (ii) whether they differ from the surrounding vowel segments. The vowel was restricted to *a* because our previous studies showed that the VQPs change significantly with the vowel quality and it is not clear whether this is an analysis artefact or caused by a change of the glottal pressure and volume velocity waveforms.

### F. Speech Recordings

The recordings were made in the sound treated room of the IMS (Institut für Maschinelle Sprachverarbeitung). The sound pressure signal was recorded with an omnidirectional capacitor microphone AKG-CK62-ULS. The microphone was located in front of the speaker in a distance of about 40cm slightly out of the speaking direction. This distant setup without popkiller was chosen because the speaker held additional recording equipment, a rothenberg mask at his face and an EGG and acceleration sensor at his neck.

### G. Statistical Analysis

The data were analyzed using SPSS version 29. For each of the parameters as a dependent variable a MANOVA and/or a nonparametric test (Mann-Whitney-U tests, depending on normal distribution) were carried out for the effects of conditions vowel /a/ and nasal /n/.

## III. RESULTS

### A. Sound Pressure Signal

Fig. 1 shows the microphone signal (sound pressure) of the /'nanana/ recording. The sound segments in milliseconds are:  $n_1=30-70$ ,  $a_1=80-180$ ,  $n_2=190-290$ ,  $a_2=300-380$ ,  $n_3=390-520$ ,  $a_3=530-730$

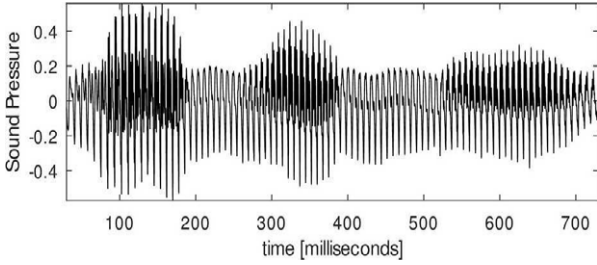


Fig. 1: Sound pressure of /'nanana/

### B. VQP Contours

The analysis procedure yields a new set of estimates every 10 milliseconds. Fig. 2 shows all VQPs for the /'nanana/ utterance.

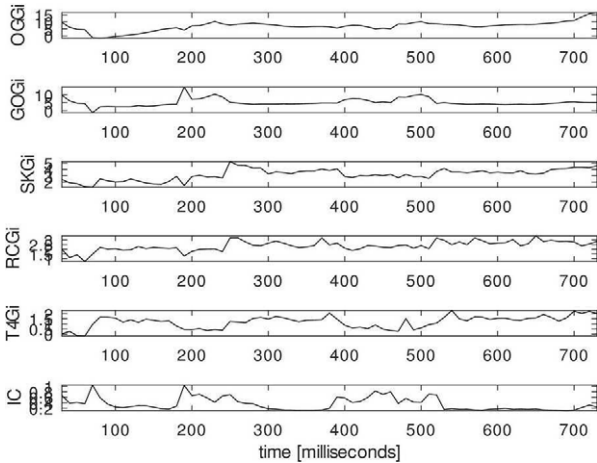


Fig. 2: VQPs of /'nanana/

None of the VQPs is unchanged along the utterance.

### C. Angles

The observation that SKGi, RCGi, and T4Gi are smaller during some n-s than in their surrounding a-s, supports a closer look to the direction of the VQP vectors.

Tab. 2: Angles between the averaged voice quality vectors of the six sounds of /'nanana/ in degrees.

	$a_1$	$n_2$	$a_2$	$n_3$	$a_3$
$n_1$	29	6	18	4	21
$a_1$	0	27	26	26	32
$n_2$	27	0	12	2	16
$a_2$	26	12	0	14	7
$n_3$	26	2	14	0	18

Tab. 2 shows the angles between all sounds of /'nanana/. The VQP direction of the first vowel /a1/ differs strongest from that of all subsequent sounds by 26 degrees and more. The vowels /a2/ and /a3/ are only 7 degrees apart. All three /n/ directions are closer together. The /n/-/a/ changes with 29, 12, 18 degrees are comparable to that at the CV-CV boundaries /a/-/n/ 27, 14. The largest angle is between the first and last vowel /a1/ and /a3/.

Tab. 3: Angles between the averaged voice quality vectors without IC of the six sounds of /'nanana/ in degrees.

	$a_1$	$n_2$	$a_2$	$n_3$	$a_3$
$n_1$	29	<b>5</b>	<b>17</b>	4	<b>20</b>
$a_1$	0	27	<b>25</b>	26	32
$n_2$	27	0	12	2	16
$a_2$	<b>25</b>	12	0	14	7
$n_3$	26	2	14	0	18

The statistics in the next section D reveals no significant difference of the spectral gradient VQPs between nasals and vowels, but significant difference of IC. The obvious question: do the angles shrink, if IC is left out in the VQP vector definition? The modified vector of voice quality parameters  $vd$  is defined in (4).

$$vd = (OQGi, GOGi, SKGi, RCGi, T4Gi)^* \quad (4)$$

Tab. 3 shows that only few of the angles shrink by one degree. Those angles, that differ from Tab. 2 are

displayed boldface. A closer look reveals that the first /na/ gets now a little closer to the subsequent two.  
*D. Statistics*

Only for the parameter IC (incompleteness of closure) a significant difference between vowel and nasal productions can be found. The parameter refers to the bandwidth of the first formant. It is significantly larger for the nasal /n/ than for the vowel /a/ ( $p < 0.01$ ). A physiological explanation for this result can be seen in the fact, that adding the nasal tract as a second resonant cavity increases the bandwidth of the first formant. The bandwidth reflects a lower or higher loss of acoustic energy depending on whether the nasal tract is added or not. A change of subglottal pressure during the nasal may also influence the glottal opening and hence, the IC. In particular, a higher loss of acoustic energy at the glottal level increases the bandwidth of the first formant.

#### IV. DISCUSSION

The VQP space currently lacks cartographic maps. Previous studies compared utterances of different phonation qualities or pathologies and identified significantly differing sets of VQPs. But a mapping of a given VQP vector to phonation qualities is not available.

One reason for that situation is the lack of manually and continuously phonation quality labeled speech recordings – not even to wish for corpora.

The study is planned to be continued by labeling and analyzing all the recorded utterances. This will show whether the first observations reported here hold.

Then, the extension to rough and breathy voice quality will be tried.

Applying the analysis to the nasal microphone of the rothenberg mask may show whether the VQPs are applicable there.

The dadada recordings add a stop closure, but /d/ may contain some sonorant frames of sufficient duration to allow the VQPs to be computed. Here the comparison of the VQPs between the mouth microphone of the rothenberg mask and the microphone in front of the speaker may reveal a further aspect of robustness of the VQPs.

#### V. CONCLUSION

The spectral gradient VQPs do not change significantly between nasal and vowel segments in the sound pressure recording of /'nanana/ that was analyzed in this pilot study. However, the VQP IC, the relative bandwidth of the first formant, changes significantly. Leaving out IC from the VQP vector definition only slightly reduces the angles between the sounds of the first /na/ to the subsequent ones by one degree.

#### REFERENCES

- [1] K. Stevens, H. Hanson, "Classification of glottal vibration from acoustic measurements," in *Vocal Fold Physiology*, O. Fujimura and M. Hirano Eds. Cambridge MA: Hiltop University Press 1998, pp. 147-170.
- [2] K. Marasek, "Electroglottographic Description of Voice Quality," Habilitationsschrift, Stuttgart, 1997.
- [3] M. Lugger and B. Yang and W. Wokurek, "Robust Estimation of Voice Quality Parameters Under Realworld Disturbances," IEEE ICASSP 2006, pp. 1097--1100.
- [4] W. Wokurek and M. Pützer, "Automated corpus based spectral measurement of voice quality parameters," Proc. 15th ICPhS (Barcelona 2003), pp. 2173--2176.

**SESSION IV**  
**TOOLS AND METHODS FOR SPEECH**  
**AND VOICE ANALYSIS**



# INFANT CRY FOR PATHOLOGIES CLASSIFICATION USING A DEEP LEARNING APPROACH

Carlos A. Reyes-Garcia<sup>1</sup>, Ingrid A. Valencia-Hernandez<sup>2</sup>, and Orion F Reyes-Galaviz<sup>3</sup>

<sup>1,2</sup> Instituto Nacional de Astrofísica Óptica y Electrónica, Puebla, México

<sup>3</sup> Laivly, Winnipeg, Canada

<sup>1</sup>[kargaxxi@inaoep.mx](mailto:kargaxxi@inaoep.mx) <sup>2</sup>[ingrid.valencia.1243@gmail.com](mailto:ingrid.valencia.1243@gmail.com), [orionfausto.reyes@laivly.com](mailto:orionfausto.reyes@laivly.com)

**Abstract:** Crying in babies is a primary communication function, governed directly by the brain; any alteration of the normal functioning of the babies' body is reflected in the cry. Based on the information carried by the cry's wave, the infant's physical state can be determined; and even pathologies in very early stages of life detected. To perform the identification of pathologies, a Deep learning algorithm was developed and applied. The input features are presented as spectrogram and Mel Frequency Cepstral Coefficient (MFCC) image representations. The combination of the deep learning model with the image representation of the acoustic features brings a very high classification accuracy around 95%.

**Keywords:** Infant Cry, Pathologies Classification, Deep Learning.

## I. INTRODUCTION

The number of researchers interested in the study of infant cry has exponentially increased in recent years, as well as the areas of interest, technological and medical approaches, and experimented methodologies. Two main things have motivated the fast growing of the interest in this field: one is to uncover the unambiguous interpretation of the rich information hidden in the crying wave. The second is to develop robust intelligent recognition systems as the main components of supporting tools to help medical specialists make accurate and objective diagnostics based on the crying waves representations.

It is well known that during the early days of life the Central Nervous System (CNS) is in charge of all vital functions. Among those functions, one that is even essential for survival is infant cry. Given its reliance on the CNS, any changes in the baby's physical and emotional state will manifest in the form of alterations in the crying pattern. If the needed knowledge or the right technological tools are available, the information carried by the crying wave can be extracted for its decoding and interpretation. In general, the automatic infant cry analysis is focused on determining either the cause or the type of crying. When attempting to determine the cause of a baby's cry, the focus is on identifying the underlying reason, whereas determining

the type of cry aims to distinguish between normal and pathological crying. In this context, innovative intelligent methodologies have emerged to not only recognize pathological cries but also classify the specific conditions affecting the baby. The present work is driven by this objective. The analysis of newborn crying, conducted through spectrogram observations, has played a crucial role in defining its key characteristics. Starting from the seminal studies of Michelsson and Wasz-Hockert [1] on healthy infants and those with asphyxia in the 1960s, advancements have been made in automated methods for cry analysis. Traditionally, crying can be examined from two perspectives: quantitative analysis and qualitative analysis.

### 1.1. Related Works

In earlier studies focused on acoustical analysis of infant crying, significant distinctions were observed among different cry types, such as healthy cries, cries of pain, and pathological cries. These distinctions were made possible through the utilization of classification methodologies based on Self-Organizing Maps [2], neural networks (NN) [3], and spectral analysis [4]. In a particular study conducted by Petroni, Neural Networks [5] were employed to differentiate between pain and no-pain crying. Cano directed several works devoted to the extraction and automatic classification of acoustic characteristics of infant cry. In a notable study conducted in 1999, Cano demonstrated the effectiveness of Kohonen's Self-Organizing Maps in classifying Infant Cry Units [6]. More recently, in [7] our team reported the classification of cry samples from deaf and normal babies with feed-forward neural networks. In 2004 Cano and his group further investigate the presence of CNS diseases, employing a radial basis network (RBN) [8]. Additionally, in [9] we showcased the implementation of a Fuzzy Relational Neural Network (FRNN) for Detecting Pathologies through Infant Cry Recognition.

In the connectionist approach (ANN), pattern classification is done with a multi-layer neural network. A weight is assigned to every link between neurons in contiguous layers. In the input layer, each neuron receives one of the features present in the input pattern vectors. Each neuron in the output layer

corresponds to each speech unit class (word or sub-word). The neural network associates input patterns to output classes by modeling the relationship between the two pattern sets. The pattern is estimated or learned by the network with a representative sample of input and output patterns [10]. A very important stage in the search for more robust classification models was the development of a new approach that has been known as deep learning. Deep neural networks are powerful machine learning models with successive layers of nonlinear processing to extract features from the data [11].

### 1.2. The Baby Chillanto Data Base

The Baby Chillanto Data Base is a collection of Mexican samples recorded by medical doctors who were instructed to capture the crying sounds of infants. After each recording, the doctors were required to capture the cause or type of crying, enabling us to accurately identify and label them within the digital database. The data base comprises cry samples collected from 98 babies, including six babies suffering from asphyxia, six with deafness, and the rest from normal babies. For cases of deafness and normal cry multiple samples were recorded from the same baby during different crying episodes, resulting in a total of 53 complete samples from deaf infants. The recordings were obtained from babies aged between 2 days up to six months. The duration of these recordings ranged from 7 seconds to 3 and a half minutes. It is important to note that cry samples associated with full hunger and full pain were classified within the Normal Cry category. The recordings were conducted in a controlled environment, specifically in a closed room, with the only source of contamination being the noise generated by the air conditioning system. Next, the original recordings were segmented into one second segments, each of which is taken as a training sample. It is worth to mention that this database is nowadays the standard infant cry database more used as a reference in the works.

## II. METHODS

The crying recordings underwent a pre-processing process to ensure the consistency of the data and facilitate efficient processing. Plus, our main goal was to avoid the need of additional tools beyond Python libraries. To pre-process the data, the Librosa library was used, which is used for audio processing. This library was developed by Brian McFee [12]. The tools from this library allowed us to read the audio data and convert them into a NumPy data structure. When converting audio data into vectors, the default process involved converting the sample rate to 22.05 KHz, normalizing the data to a bit depth that ranges from -1

to 1, and flattening the channels into a single (mono) channel.

### 2.1. Feature Extraction

The next step involved extracting the necessary features for model training. To achieve this, spectrograms were created as visual representations of each audio sample. This allows us to derive features for classification. Thus, note that the model was trained not on the audio features themselves, but on the visual representations of the spectrograms treated as images. Spectrograms are a valuable technique used to visualize the frequency spectrum of a sound temporal variations. Furthermore, another representation similar to spectrograms, known as Mel Frequency Cepstral Coefficients (MFCC), was extracted, which have been widely used in crying analysis. The difference between these two representations is that a spectrogram uses a linear spaced frequency scale (ensuring even frequency distribution), while MFCC use a quasi-logarithmic spaced frequency scale, which is more similar to how the human auditory system perceives and processes sounds. For each audio file in the data set, MFCC features are extracted and converted into images, which are then stored alongside their respective class labels in an array.

### 2.2. Deep Learning Model

For the classification process, the subsequent step involved constructing and training a deep learning neural network using the image sets. The implemented neural network is a Convolutional Neural Network (CNN), renowned for its proficiency in image classification tasks. CNN models excel at extracting features and performing classification within their architecture, enabling them to effectively learn the shapes and spatial patterns inherent in the input images. The model was sequentially built using the Keras library, which is based on TensorFlow [13]. It comprises four convolutional layers (Conv2D) and one dense output layer. The output layer has 5 nodes, corresponding to five infant cry classes: asphyxia, deafness, hunger, normal and pain.

Each convolutional layer has filters that extract features from the sound wave representations captured in the images. The first layer has 16 filters, followed by 32 in the second layer, 64 in the third, and 128 in the fourth. These extracted features play a pivotal role in classifying the different types of crying. To aid in feature selection, each convolutional layer is accompanied by a maximum pooling layer, which computes the maximum value within each “patch” of the feature map (sections extracted by each filter). Additionally, a dropout layer is included after each convolutional layer. This layer introduces noise during

training, serving as a regularization technique to mitigate the risk of overtraining or overfitting the model. Figure 1 illustrates the proposed CNN architecture.

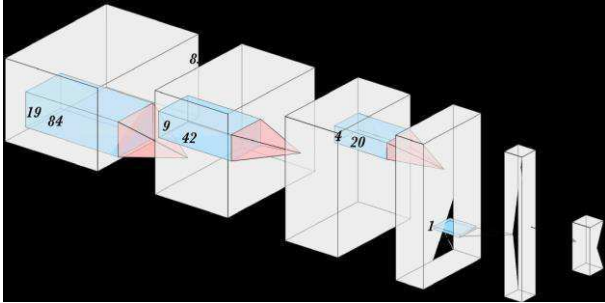


Figure 1. Convolutional Neural Network architecture.

### III. RESULTS

To assess the effectiveness of our model, we used various evaluation parameters, including a categorical cross-entropy loss function, and an accuracy metric (i.e., MSE) to gauge network performance based on validation data. To optimize the model we used Adam, which is a stochastic optimization method.

The model underwent training for 70 epochs, with the classification accuracy being verified and measured at the end of each epoch. Only if the model demonstrated improvement compared to the previous epoch, it automatically saved. To gain further insights, we used confusion matrices to identify any troubles encountered by the final model in distinguishing between the expected classes and examine the nature of these misclassifications.

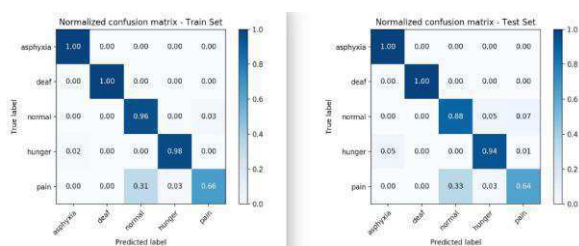


Figure 2. Confusion matrix of the Train Set (left) and of the Test Set (right)

The initial untrained model exhibited a loss of 6.35 and a validation accuracy of 16.52%. After training, the model's loss significantly improved to 0.2271, accompanied by a validation accuracy of 95.31%. Analyzing the confusion matrix for the Test Set, it is evident that the model demonstrated generalization

across all five classes. Notably, it achieved a 100% classification accuracy for the asphyxia and deafness classes, 94% for the hunger class, and 88% for the normal class. However, the pain class exhibited a lower classification percentage due to the model occasionally confusing it with normal crying. This can be attributed to the fact that the pain cry was emitted by babies who are otherwise normal and do not exhibit any specific pathologies; they are simply experiencing a painful episode.

### IV. CONCLUSION

We have successfully demonstrated the feasibility of automating the classification of infant cry. And that, once a robust classification system is developed, the results can potentially assist pediatricians, nurses, or general doctors in identifying certain pathologies, such as deafness or asphyxia, in recently born babies. A system like the one described here, is not intended to substitute the medical specialist; to the contrary, it is thought as a non-invasive tool to warn doctors of possible malfunctions or pathologies present in babies. By providing timely warnings regarding such pathologies, doctors can pay special attention on suspicious cases, in order to detect the extent of a pathology as early as possible. This early diagnosis facilitates the application of appropriate therapies, preventing learning delays, future disabilities, and even potential mortality. We also showed that CNN is a reliable classification model, which offers very acceptable performance results. Compared to other classifiers we have tested; a CNN offers some advantages including the ability to directly process the spectrograms of the recordings without having to extract additional acoustic features.

### REFERENCES

- [1] Wasz-Hockert et al, The Infant Cry: A Spectrographic and Auditory Analysis, William Heinemann Medical Books Ltd, 1968.
- [2] Sergio D. Cano, Daniel I. Escobedo y Eddy Coello, El Uso de los Mapas Auto-Organizados de Kohonen en la Clasificación de Unidades de Llanto Infantil, Grupo de Procesamiento de Voz, 1er Taller AIRENE, Universidad Catolica del Norte, Chile, 1999, pp 24-29.
- [3] Marco Petroni, Alfred S. Malowany, C. Celeste Johnston, Bonnie J. Stevens. Identification of pain from infant cry vocalizations using artificial neural networks (ANNs), pp.729-738. The International Infant Cry Research Group. Applications and Science of Artificial Neural Networks. The International Society for Optical Engineering. Volume 2492. Part two of two. Paper #: 2492-79. 1995.



- [4] O. Wasz-Hockert, J. Lind, V. Vuorenkoski, T. Partanen y E. Valanne, *El Llanto en el Lactante y su Significación Diagnóstica*, Ed. Científico-Médica, Barcelona, 1970.
- [5] Petroni, M., Malowany, A. S., Johnston, C., and Stevens, B. J., (1995),. Identification of pain from infant cry vocalizations using artificial neural networks (ANNs), *The International Society for Optical Engineering*. Volume 2492. Part two of two. Paper #: 2492-79. pp.729-738.
- [6] Cano, Sergio D, Escobedo, Daniel I., and Coello, Eddy (1999), *El Uso de los Mapas Auto-Organizados de Kohonen en la Clasificación de Unidades de Llanto Infantil*, Grupo de Procesamiento de Voz, 1er Taller AIRENE, Universidad Católica del Norte, Chile, pp 24-29.
- [7] Orozco, J., Reyes, C.A. (2003), Mel-frequency Cepstrum Coefficients Extraction from Infant Cry for Classification of Normal and Pathological Cry whit Feed-Forward Neural Networks, *Proc. of ESANN*, Bruges, Belgium.
- [8] Cano, Sergio D, Escobedo, Daniel I., Ekkel, Taco (2004) A Radial Basis Function Network Oriented for Infant Cry Classification, *Proc. of 9th Iberoamerican Congress on Pattern Recognition*, Puebla, Mexico.
- [9] Suaste, I., Reyes, O.F., Diaz, A., Reyes, C.A. (2004) Implementation of a Linguistic Fuzzy Relational Neural Network for Detecting Pathologies by Infant Cry Recognition, *Proc. of IBERAMIA*, Puebla, Mexico , pp. 953-962.
- [10] Morgan, D.P., and Scofield, C.L. (1991), *Neural Networks and Speech Processing*, Kluwer Academic Publishers, Boston.
- [11] S. Dong, P. Wang, and K. Abbas, “A survey on deep learning and its applications,” *Computer Science Review*, vol. 40, p. 100379, 2021.
- [12] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015, July). *librosa: Audio and music signal analysis in python*. In *Proceedings of the 14th python in science conference* (Vol. 8, pp. 18-25).
- [13] Dillon, J. V., Langmore, I., Tran, D., Brevdo, E., Vasudevan, S., Moore, D., ... & Saurous, R. A. (2017). *Tensorflow distributions*. arXiv preprint arXiv:1711.10604.

# THE IMPACT OF DIFFERENT TYPES OF TEACHING MODES ON VOCAL FATIGUE IN UNIVERSITY TEACHERS

K.V. Evgrafova<sup>1</sup>, N. S. Sokolova<sup>2</sup>, N. V. Shvaley<sup>3</sup>

<sup>1</sup> Phonetics Department, Saint Petersburg State University, Saint Petersburg, Russia

<sup>2</sup> the Department of English Philology and Linguoculturology, Saint Petersburg State University, Saint Petersburg, Russia

<sup>3</sup> Saint Petersburg State Pediatric Medical University, Saint Petersburg, Russia

**Abstract:** Vocal fatigue in university teachers is an important problem which has been focused on in both acoustic and medical studies. It results in perceptual and acoustic voice changes and can lead to serious pathological conditions. The shift to online synchronous teaching due to the onset of COVID-19 pandemic in 2020 brought a new challenge causing the significant increase in vocal fatigue. The goal of this study is to analyse the impact of different teaching modes on vocal fatigue in university professors. We compared acoustic and clinical data obtained during pre-pandemic years (classroom teaching mode); pandemic semesters 2020-2021 (online synchronous teaching); post-pandemic semesters 2021-2022 (a hybrid teaching mode); post-pandemic semesters 2022-2023 (classroom teaching mode).

**Keywords:** vocal fatigue, teacher's voice, voice load, online synchronous teaching, COVID-19 pandemic.

## I. INTRODUCTION

Vocal fatigue in voice professionals is an important multifaceted issue which has been focused on in many studies [1-4], [6-17]. It has clinical symptoms (associated with different types of dysphonia) shown by laryngoscopy and self-reporting complaints such as a sense of increased vocal effort and a sensation of laryngeal and pharyngeal constriction. There are also physiological and psychological symptoms which can appear as a result of vocal overloading. Another aspect of the phenomenon is an acoustic one. The vocal fatigue can be manifested in the variation of fundamental frequency.

We performed acoustic and clinical assessments as well as psychometric evaluation of self-report questionnaires to compare the symptoms and degree of vocal fatigue in the professors of Saint Petersburg State university (pronunciation teachers and lecturers) working in different types of teaching modes.

We compared the data obtained during

- 1) pre-pandemic years (*classroom teaching mode*);
- 2) pandemic semesters 2020-2021 (*online synchronous teaching*);
- 3) post-pandemic semesters 2021-2022 (a mixture of distant and classroom activities);
- 4) post-pandemic semesters 2022-2023 (*classroom teaching mode applied only*).

The goal of this paper was to compare and describe the impact of each type of teaching modes on the level of the vocal fatigue in university professors.

## II. METHODS

To obtain the data during the post-pandemic semesters, we stuck to the protocol used in the pre-pandemic and pandemic vocal fatigue studies [6-10]. The experimental design, tasks and recording material were kept, although the set of subjects and recording conditions differed.

In the pre-pandemic studies 20 male and female subjects were recorded. We had involved pronunciation teachers employed at the department of Phonetics (Saint Petersburg University) with average work experience of 7 years. The recordings were made in the recording studio at the Department of Phonetics, Saint-Petersburg State University. Multi-channel recording system Motu Traveler, capacitomicrophone AKG and WaveLab program were used. The recordings had a sample rate of 44100 Hz and a bit rate of 16 bits.

However, in pandemic and post-pandemic studies 10 female teachers currently employed at the period at the Department of Phonetics and the Department of English Philology and Cultural studies were involved. They performed different types of teaching activities (i.e.: lecturing on linguistics; running practical English classes, and pronunciation coaching). The minimum workload a day was 3 hours while the maximum was 6 hours.

Due to Covid-19 restrictions, experimental recordings at the studio were not available. The teachers were instructed to record their voices *before* and *after* teaching activity using their mobile phones.

Nevertheless, smart phone microphones are assessed in terms of the reliability of acoustic voice parameters in a number of relevant studies. It is shown that measures obtained from voice recordings using regular microphones in a sound-proof room and smartphone microphones have no statistically significant difference. [17] In order to obtain reliable acoustic data for subsequent acoustic analysis, the participants were provided with a set of recording guidelines. During the post-pandemic semesters 2022-2023 the recordings were made in the recording studio at the Department of Phonetics again.

At the moment of the experiments in each study (*pre-pandemic/pandemic/post-pandemic*) participants did not report any chronic voice pathologies. All of them had been undergoing regular laryngeal exams.

The subjects read a four minute phonetically representative text in Russian (their native tongue). They were asked to read at habitual loudness before classes in the morning. After continuous classroom/online teaching during the working day they were asked to record the same text.

In all the studies the **WAM** questionnaire was used to evaluate psychoemotional state of the teachers before and after their work. **WAM** (wellbeing, activity, mood) is used to assess the mental state of subjects, their psychoemotional response to loading. [5] The WAM questionnaire has the form of the scale with indices (3 2 1 0 1 2 3) and 30 pairs of words with opposite meaning (active - passive, strong-weak, cheerful-sad). Besides, each participant wrote a detailed report describing their self-perception of voice, mood, physical condition, type of voice activity, working conditions, and platforms used for online synchronous teaching.

All the participants had the laryngoscopy of vocal cords done regularly during the period of 2021-2023.

### III. RESULTS

#### A. Acoustic results

The main acoustic parameter which tended to vary were mean F0 values. Besides, the ratio of laryngealization passages to the whole text was also different.

The values of these parameters in non-fatigued and fatigued female speech in pre-pandemic, pandemic and post-pandemic recordings are presented in Table 1 below.

Table 1. Mean F0 variation (female voice)

Female subjects	F0, Hz	Pitch max, Hz	Laryngealization, %
<b>Classroom teaching mode (pre-pandemic)</b>			
N/F	209	351	1,5
F	212	445	1,2
<b>Online synchronous teaching (pandemic)</b>			
N/F	239	365	1,8
F	251	468	2,3
<b>Hybrid mode of teaching (post-pandemic)</b>			
N/F	217	360	1,4
F	222	452	1,9
<b>Classroom teaching mode (post-pandemic)</b>			
N/F	205	348	1,3
F	208	447	1,5

F0 increase is noticeable in the fatigued speech across all types of the recordings, but the pandemic maximum pitch value tends to be the highest. Meanwhile, the post-pandemic values appear to have returned to the pre-pandemic ones.

The mean ratio of laryngealized speech segments to the whole text is the biggest during the pandemic period and also has reduced in the post-pandemic material

Laryngealization (occurs typically in the end of an utterance before a pause) which is marked by significant decrease in pitch value and pitch breaks is associated with a creaky voice quality. This symptom was frequently reported by the teachers during the self-assessment of voice quality. The mean ration of laryngealized speech segments to the whole text is the longest during the pandemic period and also has reduced in the post-pandemic material.

#### A. Psychometric evaluation of self-report questionnaires

The WAM questionnaires showed that in all types of the studies *before* and *after* the workload **Wellbeing** scale exceeded 4 points (Table 2).

However, on average, the *after* self-assessment showed decreased wellbeing index, but it did not fall out of the range of 4.0 points. According to the

**Activity** scale the rates exceeded 4 points. The **Mood** rates increased *after* the workload.

In total, the results of pre-pandemic and post-pandemic tests look similar, whereas well-being, activity and mood rates are lower in pandemic data.

These results are compliant with the complaints in the self-reports presented in the pandemic period.

Table 2. The mean rates of WAM test

Before	After
<b>Classroom teaching mode (pre-pandemic)</b>	
<b>Wellbeing</b>	
5.9 (min. 5.3 – max. 5.8)	5.8 (min. 5.2 – max. 6.1)
<b>Activity</b>	
4.8 (min. 4.1 – max. 6.5)	5.5 (min. 5.1 – max. 6.2)
<b>Mood</b>	
6.0 (min. 4.3 – max. 6.7)	6.3 (min. 5.9 – max. 6.7)
<b>Online synchronous teaching (pandemic)</b>	
<b>Wellbeing</b>	
5.5 (min. 4.3 – max. 5.8)	4.3 (min. 4 – max. 5.1)
<b>Activity</b>	
4.3 (min. 4.1 – max. 5.5)	5.4 (min. 4.1 – max. 6.1)
<b>Mood</b>	
5.0 (min. 4.3 – max. 5.2)	5.3 (min. 4.9 – max. 6.3)
<b>Hybrid mode of teaching (post-pandemic)</b>	
<b>Wellbeing</b>	
5.8 (min. 4.9 – max. 6.7)	5.5 (min. 5.0 – max. 5.9)
<b>Activity</b>	
4.7 (min. 4.2 – max. 6.2)	5.8 (min. 5.2 – max. 6.5)
<b>Mood</b>	
5.9 (min. 4.5 – max. 6.2)	6.1 (min. 4.9 – max. 6.3)
<b>Classroom teaching mode (post-pandemic)</b>	
<b>Wellbeing</b>	
5.7 (min. 5.4 – max. 5.9)	5.9 (min. 5.2 – max. 6.0)
<b>Activity</b>	
4.5 (min. 4.1 – max. 5.6)	5.4 (min. 4.1 – max. 6.1)
<b>Mood</b>	
5.3 (min. 4.3 – max. 5.4)	5.35 (min. 4.9 – max. 6.4)

### C. Clinical results

The clinical analysis showed that online synchronous teaching mode caused the most alarming voice fatigue symptoms. The self-reports included the following complaints: hoarse voice quality, creaky/fry voice, breathy voice, unsteady pitch, dry/scratchy throat, frequent throat clearing, sore throat, dry cough. It is evident that vocal overload, inadequate posture and continuous talking while sitting, lack of auditory and visual feedback/student interaction, technical problems, online connection failures lead to psychological stress and difficulties in voice production.

The laryngoscopy analysis showed hypotonic dysphonia in several subjects. It is marked by decrease in the density of closure of the true vocal folds, linear and oval fissure in all parts of the range, visibility of the ventricles of the larynx, absence of stroboscopic comfort. Besides, one severe case was observed in which voice overloading brought the pre-nodule condition of vocal cords [10]

## IV. DISCUSSION

We believe that online synchronous teaching mode triggers excess voice use due to the necessity to use remote microphones which leads to forced manner of voice production often resulting in vocal fry/creaky voice.

Besides, stress, asthenia, and general decrease in physical activity imposed by COVID-19 isolation were found to be additional factors of the hypotonic dysphonia development. Fast vocal fatigue and overall lack of energy are often seen as subjective manifestations of MTD [12], [14-16].

Switching to hybrid mode of teaching and returning to entirely classroom teaching in the post-pandemic period along with developing adaptation mechanisms brought the relief in voice fatigue to the educators. No microphone use and visible audience follow-up and students' reaction has had a positive impact on the vocal functions which prevented possible pathological changes in the larynx.

The data obtained during the period of hybrid mode of teaching demonstrated that the teachers were able to develop specific voice strategies, which prevented voice fatigue. They included reducing rate of speech, increasing vocal pauses in connected speech, focusing on clear articulation to avoid increasing loudness.

## V. CONCLUSION

The analysis of acoustic and clinical data on vocal fatigue in university teachers shows that online synchronous teaching activity is the most challenging

one. The most alarming symptoms of dysphonia due to voice overstrain were present during pandemic semesters. However, the adaptation strategies were developed which helped the educators to cope with the excess voice use during the hybrid mode period. The return to the regular working conditions has had a positive effect on the vocal functions, although the vocal quality and the clinical picture still do not resemble the pre-pandemic data which may be related to the long covid syndrome in some of the subjects.

In the further study we plan to analyse the impact of long covid syndrome on vocal endurance in our subjects.

#### REFERENCES

- [1] N.A. Agadzhanian, I.S. Vasilenko, and A.I. Smirnova. "The effect of phonational load on parameters of cardiorespiratory system in hypotonic dysphonia". *Vestn Otorinolaringol.*; (4):15-7. PMID: 16091714 in Russian. 2005.
- [2] 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).76: e2641. Published online 2021 Mar 19. doi: 10.6061/clinics/2021/e2641 2021.
- [3] 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.
- [4] M.J. Caraty, C. Montacié, "Multivariate Analysis of Vocal Fatigue in Continuous Reading, The Proceedings of Interspeech, pp. 470-473, 2010.
- [5] V.A. Doskin, "Test differentsirovannyi samoosneno funktsional'nogo sostoyaniya" (The Test of Differentiated Self-Assessment of Functional Status)/ Doskin, V.A. Lavrenteva, N.A. Miroshnikov, M.P. Sharay V.B. // *Voprosy psikhologii* 1973 №6, 141-145, 1973.
- [6] K.V. Evgrafova, V.V. Evdokimova "Acoustic analysis of vocal fatigue in professional voice users". *MAVEBA* 2011, pp. 153-156.
- [7] K.V. Evgrafova, V.V. Evdokimova P.A. Skrelin, T.V. Chukaeva "Vocal fatigue in voice professionals: collecting data and acoustic analysis". DOI: 10.36505/ExLing-2016/07/0011/000270
- [8] K.V. Evgrafova, N.S. Sokolova, N.V. Shvaley. "The Impact of online teaching during the COVID-19 on vocal fatigue in university professors: self-reports and acoustic evaluation". *Maveba* 2021, 167-170. DOI: 10.36253/978-88-5518-449-6
- [9] 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
- [10] K. Evgrafova, N. Sokolova, N. Shvaley "The Adaption to Online Synchronous Teaching and Voice Fatigue: Acoustic and Clinical Data." *J Voice*. 2023 Mar 27:S0892-1997(23)00086-3. doi: 10.1016/j.jvoice.2023.02.029. Epub ahead of print. PMID: 36990863; PMCID: PMC10041334.
- [11] 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>
- [12] O. S. Orlova, Yu. S. Vasilenko, A. F. Zakharova, L. O. Samokhvalova, and P. A. Kozlova "Prevalence, causes and features of voice disorders in teachers". *Vestnik otorinolaringologii*. 2000;5:18-21. (In Russ.)
- [13] M. H. Patjas, P. Vertanen-Greis, A. 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]
- [14] A. Sama, P. N. Carding, S. Price, P. Kelly, Wilson J. A. The clinical features of functional dysphonia. *Laryngoscope*. 2001;111:458-63.
- [15] E. Van Houtte, K. Van Lierde, S. Claeys Pathophysiology and treatment of muscle tension dysphonia: A review of the current knowledge. *J Voice*. 2011;25:20-27.
- [16] Yu. S. Vasilenko "Golos. Foniatricheskie aspekty". M.: Dipak, 2013. (In Russ.)] *Rossiiskaya otorinolaringologiya*.
- [17] Uloza, Virgilijus et al. "Accuracy of Acoustic Voice Quality Index Captured With a Smartphone - Measurements With Added Ambient Noise." *Journal of voice : official journal of the Voice Foundation* (2021): n. pag. DOI:10.1016/j.jvoice.2021.01.025

# PERFORMANCE OF UNIVERSAL-PLATFORM-BASED *VOICESCREEN* APPLICATION IN AVQI MEASUREMENTS

Virgilijus Uloza<sup>1</sup>, Nora Ulozaitė-Stanienė<sup>1</sup>, Kipras Pribušis<sup>1</sup>, Tomas Blažauskas<sup>2</sup>, Robertas Damaševičius<sup>2</sup>, Rytis Maskeliūnas<sup>2</sup>

<sup>1</sup> Department of Otorhinolaryngology, Lithuanian University of Health Sciences, Kaunas, Lithuania

<sup>2</sup> Faculty of Informatics, Kaunas University of Technology, Kaunas, Lithuania

**Abstract:** This study aimed to assess the reliability of the VoiceScreen app in measuring the Acoustic Voice Quality Index (AVQI) and distinguishing between normal and pathological voices. The study included 135 adult participants, consisting of 49 individuals with normal voices and 86 patients with pathological voices. The "VoiceScreen" app was utilized on five iOS and Android smartphones to estimate AVQI. The AVQI values obtained from voice recordings using a reference studio microphone were compared with those obtained from smartphones. The accuracy of the app in distinguishing between normal and pathological voices was evaluated using receiver-operating characteristics. The study found a nearly perfect positive linear correlation ( $r= 0.991-0.987$ ) between the AVQI results obtained from the studio microphone and various smartphones. The AVQI demonstrated an acceptable level of precision in distinguishing between normal and pathological voices, with areas under the curve (AUC) ranging from 0.834 to 0.862. There were no statistically significant differences in the AUC values obtained from studio microphones compared to those obtained from smartphones. These findings indicate that the "VoiceScreen" app is a reliable and robust tool for measuring voice quality and screening for normal versus pathological voices. It has the potential to be utilized by both patients and clinicians for voice assessment purposes.

## I. INTRODUCTION

Previous studies have demonstrated the feasibility of using smartphone voice recordings, whether obtained in acoustically treated sound-proof rooms or in everyday environments, to estimate the Acoustic Voice Quality Index (AVQI) [1-2]. However, there is limited existing literature that provides data on AVQI estimation using various mobile communication applications [3-5].

The primary aim of this study was to address the following inquiries regarding the potential of the smartphone-based "VoiceScreen" application for

AVQI estimation: (1) Are the average AVQI values obtained from different smartphones consistent and comparable? (2) Does the diagnostic accuracy of AVQIs estimated by various smartphones have relevance in distinguishing between normal and pathological voices?

## II. METHODS

The study included a total of 135 adult participants, consisting of 58 men and 77 women. The average age of the participants in the study group was 42.9 years (standard deviation [SD] of 15.26). The subgroup of individuals with pathological voices comprised 86 patients, including 42 men and 44 women, with an average age of 50.8 years (SD 14.3). These patients presented with various common laryngeal diseases known to cause voice disturbances, such as benign and malignant mass lesions on the vocal folds and unilateral paralysis of the vocal fold. The subgroup of individuals with normal voices consisted of 49 carefully selected healthy volunteers, including 16 men and 33 women, with an average age of 31.69 years (SD 9.89). For AVQI estimation, the "VoiceScreen" application, which was developed as a uniform-platform-based (UPB) tool compatible with both iOS and Android operating systems was utilized. The application was installed on five different smartphones: iPhone Pro Max 13, iPhone SE (running on the iOS operating system), OnePlus 9 PRO, Samsung S22 Ultra, and Huawei P50 Pro (running on the Android operating system). The AVQI calculation and its characteristics were performed on a server, eliminating the need for the user's device to have high computational capabilities. The AVQI measures obtained from voice recordings captured with a studio microphone (AKG Perception 220) featuring a flat frequency response were compared with the AVQI results obtained using these smartphone devices.

## III. RESULTS

Table 1 presents the Pearson's correlation coefficients demonstrating almost perfect direct linear

Table 1. Correlations of AVQI scores obtained with studio microphone and different smartphones.

Microphones		iPhone SE	iPhone Pro Max 13	Huawei P50 pro	Samsung S22 Ultra	OnePlus 9 PRO
AKG Perception 220	r	0.991	0.987	0.970	0.979	0.992
	p	.001	.001	.001	.001	.001
	n	135	135	135	135	135

Abbreviations: r- Pearson's correlation coefficient; p - statistical significance

correlations between the AVQI results obtained using a studio microphone and various smartphones. The coefficients ranged from 0.991 to 0.987.

Figure 1 displays the receiver-operating characteristic (ROC) curves of AVQI derived from both studio microphone recordings and smartphone voice recordings. Upon visual inspection, it was evident that all the ROC curves exhibited a high degree of similarity and occupied a significant portion of the graph. This indicates the considerable ability of the AVQI to effectively differentiate between normal and pathological voices.

Table 2 presents the results of the receiver-operating characteristic (ROC) statistical analysis, which demonstrated a high level of precision in differentiating between normal and pathological voices using the AVQI. The analysis yielded a suggested threshold of  $AUC = 0.800$ , indicating a strong

discriminatory capability.

Table 2 presents the results of the receiver-operating characteristic (ROC) analysis, which determined the optimal cut-off values of the AVQI for differentiating between normal and pathological voices on each smartphone. All microphones used in the study surpassed the suggested threshold of  $AUC = 0.8$  and exhibited acceptable Youden-index values.

The DeLong et al. test confirmed that there were no statistically significant differences between the areas under the curve (AUCs) of the ROC curves ( $p > 0.05$ ). The greatest observed difference between the AUCs was only 0.028. These findings indicate that the diagnostic accuracy of the AVQI in differentiating between normal and pathological voices remains consistent when using voice recordings from both a studio microphone and various smartphones.

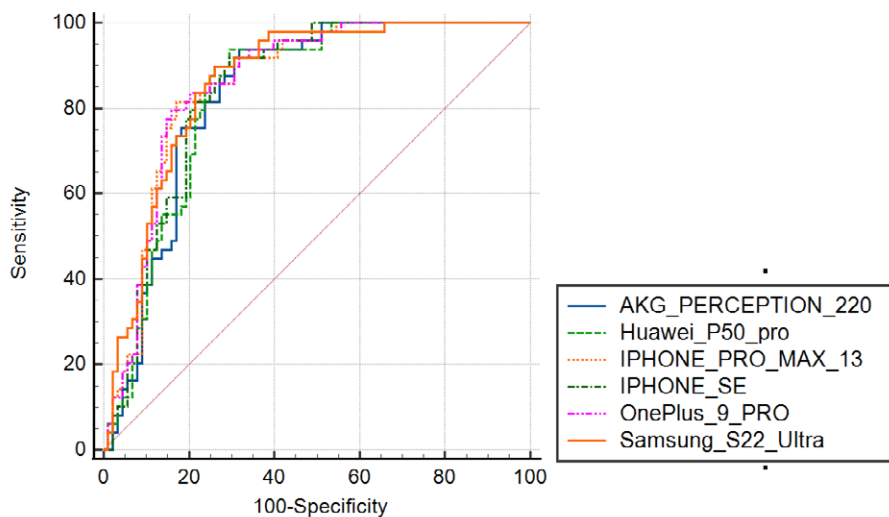


Figure 1. ROC curves illustrating the diagnostic accuracy of studio and different smartphone microphones in discriminating normal/pathological voice.

Table 2. Statistics illustrating the accuracy the AVQI differentiating normal and pathological voices recorded using studio and different smartphone microphones.

AVQI	AUC	Cut-off	Sensitivity %	Specificity %	Youden-index J
AKG Perception 220	0.834	3.27	93.88	68.18	0.62
iPhone SE	0.844	3.23	91.84	70.45	0.62
iPhone Pro Max 13	0.858	2.14	81.63	82.95	0.65
Huawei P50 pro	0.835	3.08	93.88	70.45	0.64
Samsung S22 Ultra	0.862	2.93	89.8	73.86	0.64
OnePlus 9 PRO	0.86	2.3	79.59	84.09	0.64

Abbreviations: AVQI - acoustic voice quality index, AUC - area under the curve.

#### IV. DISCUSSION

The present study employed the innovative uniform-platform-based (UPB) "VoiceScreen" application to estimate the Acoustic Voice Quality Index (AVQI) and detect voice impairments in patients with various voice disorders and healthy individuals using different smartphones.

The results of the analysis of variance (ANOVA) revealed no statistically significant differences in mean AVQI scores obtained using different smartphones ( $F=0.759$ ;  $p=0.58$ ). Additionally, the mean differences in AVQI scores ranged from 0.01 to 0.4 points when comparing AVQI values estimated with different smartphones, indicating a low level of variability. These findings align with the absolute retest difference of AVQI values proposed by Barsties and Maryn in 2013, which suggested a value of 0.54 [6,7]. Consequently, the differences in AVQI measurements between different smartphones were deemed both statistically and clinically insignificant, supporting the practical usability of the UPB "VoiceScreen" app. Furthermore, the analysis demonstrated that the AVQI exhibited a remarkable ability to distinguish between normal and pathological voices based on auditory-perceptual judgment.

These results affirm the consistent diagnostic accuracy of the AVQI in differentiating between normal and pathological voices when using voice recordings from both a studio microphone and various smartphones. This holds significant practical importance.

In summary, combining the findings from previous and current studies suggests that the performance of the UPB "VoiceScreen" app using different smartphones is reliable and yields compatible results for AVQI estimation. However, it is crucial to note that variations in recording conditions, microphones, hardware, and software may lead to differences in acoustic voice quality measurements across recording systems. Therefore, caution is advised when using the UPB "VoiceScreen" app. For voice screening purposes,

it is more reliable to perform AVQI measurements using the same device, particularly when conducting repeated measurements. Additionally, these considerations should be taken into account when comparing data of acoustic voice analysis between different voice recording systems, such as different smartphones or other mobile communication devices, and when utilizing them for diagnostic purposes or monitoring voice treatment outcomes.

#### V. CONCLUSION

The uniform-platform-based (UPB) "VoiceScreen" app proves to be a precise and reliable tool for measuring voice quality and differentiating between normal and pathological voices. It exhibits accuracy and robustness, making it suitable for voice assessment by both patients and clinicians. Moreover, the app is compatible with both iOS and Android smartphones, further enhancing its potential for widespread use.

#### ACKNOWLEDGEMENTS

**Funding:** This project has received funding from European Regional Development Fund (project No 13.1.1-LMT-K-718-05-0027) under grant agreement with the Research Council of Lithuania (LMTLT). Funded as European Union's measure in response to Covid-19 pandemic.

#### REFERENCES

- [1] C. Manfredi, J. Lebacq, G. Cantarella, Schoentgen, S. Orlandi, A. Bandini, P.H. DeJonckere. Smartphones Offer New Opportunities in Clinical Voice Research. *Journal of Voice* 2017, 31, 111.e1-111.e7,
- [2] Awan, S.N.; Shaikh, M.A.; Awan, J.A.; Abdalla, I.; Lim, K.O.; Misono, S. Smartphone Recordings are Comparable to "Gold Standard" Recordings for Acoustic Measurements of Voice. *Journal of Voice* 2023, DOI 10.1016/j.jvoice.2023.01.031.
- [3] V. Uloza, V. N. Ulozaite-Staniene, T. Petrauskas . An iOS-based VoiceScreen application: feasibility for



use in clinical settings-a pilot study. *Eur Arch Otorhinolaryngol* 2023, 280, 277-284.

[4] S. Shabnam, M. Pushpavathi, R. Gopi Sankar, K.V. Sridharan, M.S. Vasanthalakshmi. A Comprehensive Application for Grading Severity of Voice Based on Acoustic Voice Quality Index v.02.03. *Journal of Voice* 2022, DOI 10.1016/j.jvoice.2022.08.013.

[5] E.U. Grillo, J. Wolfberg, J. An Assessment of Different Praat Versions for Acoustic Measures Analyzed Automatically by VoiceEvalU8 and Manually by Two Raters. *J Voice* 2023, 37, 17-25, DOI 10.1016/j.jvoice.2020.12.003.

[6] B. Barsties, Y. Maryn. [The Acoustic Voice Quality Index. Toward expanded measurement of dysphonia severity in German subjects]. *HNO* 2012, 60, 715-720, DOI 10.1007/s00106-012-2499-9.

[7] B. Lehnert, J. Herold, M. Blaurock, C. Busch. Reliability of the Acoustic Voice Quality Index AVQI and the Acoustic Breathiness Index (ABI) when wearing CoViD-19 protective masks. *Eur Arch Otorhinolaryngol* 2022, 279, 4617-4621, DOI 10.1007/s00405-022-07417-4.

# DOMAIN ADVERSARIAL CONVOLUTIONAL NEURAL NETWORK FOR PARKINSON'S DISEASE DETECTION FROM SPEECH

E. J. Ibarra-Sulbaran<sup>1</sup>, J. D. Arias-Londoño<sup>2</sup>, M. Zañartu<sup>1</sup>, J. I. Godino-Llorente<sup>2</sup>

<sup>1</sup> Department of Electronic Engineering, Universidad Técnica Federico Santa María, Valparaíso, Chile.

<sup>2</sup> Bioengineering and Optoelectronics lab (ByO), Universidad Politécnica de Madrid, Madrid, Spain  
[emiro.ibarra@sansano.usm.cl](mailto:emiro.ibarra@sansano.usm.cl); [julian.arias@upm.es](mailto:julian.arias@upm.es); [matias.zanartu@usm.cl](mailto:matias.zanartu@usm.cl); [ignacio.godino@upm.es](mailto:ignacio.godino@upm.es)

**Abstract:** Deep learning has gained popularity in detecting Parkinson's disease (PD) from speech due to its ability to automatically extract meaningful representations from raw data. The most popular approaches are based on Convolutional Neural Network (CNN) models fed with spectrograms. However, the use of these algorithms is constrained due to the cross-dataset accuracy obtained during the validation process. Thus, in this work, we focus on studying the cross-domain effect -specifically due to different databases- for the screening of PD using a CNN-based model and two different speech corpora. To address the cross-domain challenge, we propose the use of domain adversarial (DA) training as a method to obtain discriminant and domain-invariant models. The visualization of the feature space distribution extracted by this model, using t-distributed stochastic neighbor embeddings, along with its divergence and variance by class, indicates a significant improvement in domain adaptation. These initial results provide valuable insights for further model refinement and constitute a proof of concept that domain adversarial methods offer a feasible option for creating a more generalizable speech-based PD detection model.

**Keywords:** Convolutional Neural Networks, Deep learning, Domain Adversarial, Parkinson's Disease.

## I. INTRODUCTION

Several studies have explored end-to-end deep learning techniques for screening PD directly from raw speech and time-frequency spectrograms. These techniques include CNNs [1-4], recurrent neural networks (RNNs) [5], long short-term memory (LSTM) models [6], and others [1]. Among them, CNNs have emerged as the most popular technique.

Most of the reported end-to-end deep learning methods have demonstrated high discriminative capacity in distinguishing between healthy controls (HC) and PD compared to traditional machine learning approaches. However, it is worth noting that the training and validation processes of these algorithms have been developed using a single domain, meaning a single corpus with participants sharing similar

demographics, dialectal and recording conditions.

In this line, [1, 2, 4] reveal the limitations of these models for the screening of PD when they are applied to a new dataset, resulting in a drop of accuracy of 20 absolute points. This fact highlights a significant limitation of current methods, demonstrating a noticeable degradation in their discriminative capabilities across domains. Additionally, the model relies on shortcut learning when possible, meaning it learns characteristics that differentiate between the groups but do not generalize well with respect to the underlying pathology.

In this context, we propose adding a domain adaptation step into the representation learning process, which would help to reduce the existing gap between different corpora. The goal is to ensure that the automatic screening of PD is based on features that are both discriminative and invariant to dataset changes.

In the deep learning literature, we came across the domain adversarial training method proposed in [7]. This method suggests an adversarial framework to learn domain-invariant representations. Recently, [8] proposed a speech PD classification using adversarial training to obtain speaker identity-invariant representations within a single corpus. However, they do not consider the effect of multi-dataset scenarios.

The contributions of this work are twofold: First, to analyze the robustness of an end-to-end deep learning method for PD diagnosis with respect to the shift between domains (different speech databases). Second, to study the capacity of domain adversarial training in providing more generic and reliable models for the automatic screening of PD from the speech, addressing undesired speech recording variability.

## II. MATERIALS AND METHODS

In this preliminary study, we establish a baseline model to detect PD from Mel-spectrograms by combining a CNN and a multi-layer perceptron (MLP) network, similar to those evaluated in [2-3]. We use two speech corpora to train and test the baseline model, first in a cross-domain test and then by mixing both datasets. Subsequently, the baseline model is adapted using a domain adversarial approach.

### A. Speech Corpora

The datasets used in this work were previously reported as Gita [9] and Neurovoz [10].

The Gita dataset was recorded by Clínica Noel in Medellín, Colombia. This dataset includes, among other data, diadochokinetic (DDK) tasks (i.e., repetitions of the syllable sequence /pa-ta-ka/) from 100 Colombian Spanish native speakers, with 50 HC and 50 PD patients.

The Neurovoz dataset was collected by Universidad Politécnica de Madrid in collaboration with Gregorio Marañón Hospital in Madrid, Spain. This dataset includes, among other material, DDK sequences from 86 adult speakers whose mother tongue is Castilian Spanish (44 HC and 42 PD).

Recordings for both corpora were obtained under controlled ambient conditions using a sampling rate of 44.1 kHz and 16 bits of quantization. Both datasets were recorded in compliance with the Helsinki Declaration and approved by their respective Ethics Committee.

### B. Method

The DDK speech recordings were first normalized using the maximum absolute value of amplitude. They were then segmented into 400 ms intervals overlapped 50%. Each segment was transformed into a time-frequency representation using Mel-scale spectrograms with a window size of 15 ms, a hop length of 10 ms, and 65 Mel bands. This pre-processing resulted in Mel-spectrograms of 65x41 points, which were individually normalized following a Z-score.

The baseline model consists of two modules, which we have named the feature generation network and the PD prediction network. The feature generation network receives Mel-spectrograms as input. This first module is composed of a two-dimensional convolutional layer, where each convolutional layer is followed by a batch normalization, a ReLU activation function, max-pooling (filter size: 3x3), and a dropout layer. Subsequently, the dynamic features obtained from the feature generation network are flattened to connect with the PD predictor network. This second module consists of two fully connected layers with a dropout layer in between to regularize the weights. ReLU activation is used in the first hidden layers, and a SoftMax activation function is used for classification.

For domain adversarial training, the baseline model is adapted following the Domain-Adversarial Neural Network proposed in [7]. This is accomplished by attaching a domain predictor network to the feature extractor network via a Gradient Reversal Layer (GRL). This new module contains the same architecture as the PD prediction network. The only non-standard component of the domain adversarial

architecture is the GRL, which leaves the input unchanged during forward propagation and reverses the gradient by multiplying it by a negative scalar during backpropagation [7]. The gradient reversal ensures that the feature distributions over the two domains are as indistinguishable as possible for the domain classifier, providing domain-invariant features.

Regarding training and evaluation, a stratified speaker-independent 10-fold cross-validation was used, ensuring no overlap of speakers across different folds. The hyperparameters of the baseline model were tuned with the 10-fold set of mixed data (Gita and Neurovoz) using Talos [11]. The hyperparameter search space is summarized in Table 1. The model with the best performance on the validation set for the 10 folds was selected for all experiments, including domain adversarial training (DA), where the domain predictor network parameters were set to the same values as the PD prediction network parameters.

The models were trained using the Stochastic Gradient Descent (SGD) algorithm with cross-entropy as the loss function. A learning rate schedule was used, initialized at 0.1. The PyTorch implementation of our approach is available online<sup>1</sup>.

**Table 1.** Hyperparameters search space for the baseline model

Hyperparameter	values
Training Batch size	16, 32, 64
Kernel size of conv. layer I	4, 6, 8
Kernel size of conv. layer II	5, 7, 9
Dropout rate	0.2, 0.5
Depth of convolutional layers	32, 64, 128
Units of each fully connected layer	16, 32, 64

## III. RESULTS

For mixed data training, the features extracted from the last layer of the PD prediction network for each model were labelled by class (PD and HC) and domain (Gita and Neurovoz). These features were visualized in a two-dimensional map using t-distributed stochastic neighbor embeddings (t-SNE) to study the domain adaptation effect of the baseline model in comparison to the domain adversarial network. A divergence measure was used to quantify the differences in the distribution of domain-labelled features for each class. This measure is computed using the Kullback-Leibler algorithm proposed in [12]. Additionally, the trace of the covariance matrix of the features for each class is used to quantify its variability.

### A. Cross-Domain Results

Table 2 shows the validation results obtained for the

<sup>1</sup>[https://github.com/Emiroji/Domain\\_Adversarial\\_CNN\\_Speech\\_Par\\_kinson\\_Classification](https://github.com/Emiroji/Domain_Adversarial_CNN_Speech_Par_kinson_Classification)

baseline model trained using individual datasets. The accuracy, sensitivity, specificity, and area under the ROC curve obtained for the validation sets for both Gita and Neurovoz are consistent with those reported in previous work [1-3]. We emphasize the accuracy difference between the validation and cross-domain test, which is over 30 and 20 absolute points for Gita and Neurovoz respectively. This drop in accuracy is aligned with what has been reported in the literature [1, 2, 4], confirming the mentioned limitation of end-to-end approaches trained with a limited dataset.

**Table 2.** Classification with the baseline model for each corpus. Acc: accuracy. Sens: Sensitivity. Spec: Specificity. AUC: Area under the ROC curve. Values represent the mean of 10-folds  $\pm$  standard deviation.

	Gita	Neurovoz
Acc. (%)	80.8 $\pm$ 13.4	80.1 $\pm$ 16.7
Sens. (%)	81.8	80.5
Spec. (%)	80.0	81.00
AUC	0.9	0.9
<b>Cross-Domain Acc. (%)</b>	<b>47.7 <math>\pm</math> 3.5</b>	<b>56.3 <math>\pm</math> 2.6</b>

### B. Domain Adversarial results.

Table 3 contrasts the results obtained by the baseline model and the domain adversarial network, both trained using the mixed speech corpora. The mean validation metrics decrease in comparison with experiment one (where only one dataset is used in the training process), especially for the baseline model. For example, the accuracy in the baseline model dropped by almost 8 and 4 absolute points for Gita and Neurovoz, respectively, whereas for the DA model, it was less than 5 absolute points for both validation sets, with a standard deviation slightly lower for the adversarial scheme.

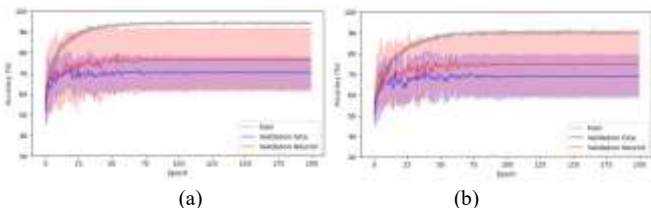
It is important to highlight that the validation metrics in Table 3 are computed based on estimations by subject. Therefore, the training and validation accuracies in terms of samples with respect to the epochs are shown in Figure 1. From these learning curves, we observe that the models reach stability before 100 epochs. As expected, the small size of the training dataset leads to overfitting of these deep learning models. These curves are consistent with those shown in [1].

The most relevant result obtained in this work is shown in Figure 2. The t-SNE representations show the features extracted by the baseline model and domain adversarial models for the fold with the best validation accuracy (the t-SNE representations for the remaining folds are available in the online GitHub repository<sup>1</sup>). The features extracted by the baseline model in the training set (Figure 2.a) report more than two classification clusters. In contrast, the domain adversarial model shows that the features of the PD

cluster for both Gita and Neurovoz share the same space, as well as for HC (see Figure 2.b). A similar trend is observed in the validation set (Figure 2.c and 2.d). However, as expected, this behavior is affected by the model's classification performance, being more evident in those folds where the model shows high accuracy.

**Table 3.** Classification with the baseline and DA models for mixed speech corpora. Acc: Accuracy. Sens: Sensitivity. Spec: Specificity. AUC: Area under the ROC curve. Values represent the mean of 10-folds  $\pm$  standard deviation.

	Baseline model		DA Model	
	Gita	Neurovoz	Gita	Neurovoz
Acc.	71.9 $\pm$ 8.8	76.7 $\pm$ 21.6	76.1 $\pm$ 11.8	80.6 $\pm$ 19.6
Sens.	70.5	80.0	76.5	80.5
Spec.	74.0	73.5	76.0	81.0
AUC	0.9 $\pm$ 0.1	0.9 $\pm$ 0.2	0.8 $\pm$ 0.2	0.9 $\pm$ 0.2



**Fig. 1.** Accuracy learning Curves during the  $k$ -fold cross-validation: (a) Baseline Model; (b) DA model. The solid line represents the mean values and the shaded regions standard deviation.

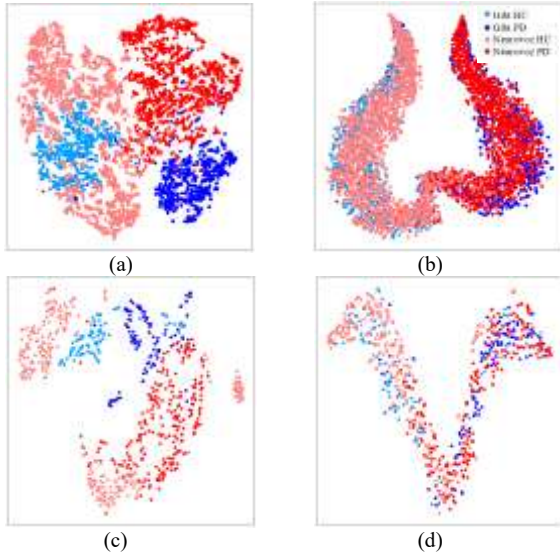
On the other hand, Table 4 shows that both the divergence and the trace of the covariance matrix between domains for each class are higher for the baseline model in contrast to its DA version. The high divergence shown in both HC and PD classes for the baseline model indicates that it presents a higher dissimilarity between the feature distributions extracted for Gita and Neurovoz. On the other hand, the results of the trace of the covariance matrix show that the baseline model exhibits a higher spread for each class.

## IV. DISCUSSION AND CONCLUSIONS

In this study, domain adaptation of an end-to-end CNN-based model for automatically PD detection using a DDK was analyzed. Although most recent work continues to compare models based solely on their accuracy in a single database, this work provides new evidence that these approaches require domain adaptation strategies to be more generalizable.

The first experiment showed that the CNN-based model learns characteristics of PD speech during internal validation. However, when tested on unseen datasets, the model failed to identify PD with sufficient accuracy. Subsequently, when both datasets were mixed during training, the features learned by the baseline model for a specific class presented a different

distribution. This behavior indicates that the model is extracting additional information relative to domain variability instead of solely obtaining PD discriminative features.



**Fig. 2.** *T-SNE of the extracted features for the training set: (a) baseline model; (b) DA model. And for the validation set: (c) baseline model; (d) DA model.*

**Table 4.** *Divergence and trace of the covariance matrix between domain distributions for each class. Values represent the mean of 10-folds  $\pm$  standard deviation.*

	Divergence		Variance	
	HC	PD	HC	PD
<b>Baseline</b>	48.8 $\pm$ 13.5	50.6 $\pm$ 9.9	33.2 $\pm$ 15.0	17.1 $\pm$ 4.5
<b>DA</b>	15.5 $\pm$ 9.3	13.5 $\pm$ 5.5	11.3 $\pm$ 2.2	9.8 $\pm$ 2.9

In contrast, domain adversarial training ensures that the model learns invariant domain features. This is evidenced by the t-SNE visualizations and by the divergence and variance metrics. These preliminary results suggest that domain adversarial training improves the generalization abilities of the network. Nevertheless, more experiments, including new speech corpora, different phonation tasks, and new architectural models, are needed.

## VI. ACKNOWLEDGEMENTS

E.J.I. thanks the support of Agencia Nacional de Investigación y Desarrollo (ANID), “Beca Doctorado Nacional 21190074”. This work was also supported by the Ministry of Economy & Competitiveness, Spain, under grants DPI2017-83405-R1 and PID2021-128469OB-I00.

## REFERENCES

[1] C. Quan, K. Ren, Z. Luo, Z. Chen, Y. Ling, “End-to-end deep learning approach for Parkinson’s disease detection from speech signals,” *Biocybern Biomed*

*Eng*, vol. 42, no. 2, pp. 556–574, 2022.

[2] J. Vásquez-Correa, J. R. Orozco-Arroyave, E. Nöth, “Convolutional neural network to model articulation impairments in patients with Parkinson’s disease,” in *Proc. Interspeech 2017*, pp. 314–318.

[3] J. C. Vásquez-Correa, C. D. Rios-Urrego, T. Arias-Vergara, M. Schuster, J. Rusz, E. Nöth, J. R. Orozco-Arroyave, “Transfer learning helps to improve the accuracy to classify patients with different speech disorders in different languages,” *Pattern Recognit. Lett.*, vol. 150, pp. 272–279, 2021.

[4] Hireš, M., Drotár, P., Pah, NN., Ngo, Q., Kumar, D., “Strengths and Limitations of Computerized PD Diagnosis from Voice”. Available at SSRN: <https://ssrn.com/abstract=4327662>, 2023.

[5] T. Fujita, Z. Luo, C. Quan, K. Mori, S. Cao, “Performance evaluation of RNN with hyperbolic secant in gate structure through application of Parkinson’s disease detection,” *Appl. Sci.*, vol. 11, no. 10, 2021.

[6] B. E. Mehmet, I. Esme, I. Ibrahim, “Parkinson’s detection based on combined CNN and LSTM using enhanced speech signals with variational mode decomposition,” *Biomed Signal Process Control*, vol. 70, p. 103006, 2021.

[7] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, V. Lempitsky, “Domain-adversarial training of neural networks,” *J. Mach Learn Res.*, vol. 17, no. 59, pp. 1–35, 2016.

[8] P. Janbakhshi, I. Kodrasi, “Supervised speech representation learning for Parkinson’s disease classification,” *Proc. ITG Conf. on Speech Communication*, July 2021.

[9] J. Orozco, J. Arias-Londoño, J. Vargas-Bonilla, M. González-Rátiva, E. Nöth, “New spanish speech corpus database for the analysis of people suffering from Parkinson’s disease,” *Proc. 9th Lang. Resources and Evaluation Conf. (LREC)*, 2014, pp. 342–347.

[10] L. Moro-Velazquez, J. Gomez-Garcia, J. Godino-Llrente, J. Villalba, J. Rusz, S. Shattuck-Hufnagel, N. Dehak, “A forced gaussians based methodology for the differential evaluation of Parkinson’s disease by means of speech processing,” *Biomed Signal Process Control*, vol. 48, pp. 205–220, 2019.

[11] Autonomio Talos [Computer software]. (2020). Retrieved from <http://github.com/autonomio/talos>. Accessed on: 26 July. 2023.

[12] F. Perez-Cruz, “Kullback-Leibler divergence estimation of continuous distributions,” 2008 IEEE Int. Symposium on Information Theory, Toronto, ON, Canada, 2008, pp. 1666-1670.

**SESSION V**  
**STUDENT COMPETITION**





# CUMULATIVE PAIR-WISE VOWEL DISTANCE (CPVD): NEW VOWEL SPACE METRICS FOR PEOPLE WITH ATYPICAL SPEECH

Tianyu Cao<sup>1</sup>, Anna Favaro<sup>1</sup>, Thomas Thebaud<sup>1</sup>, Jesús Villalba<sup>1</sup>, Piotr Żelasko<sup>2</sup>, Esther S. Oh<sup>3</sup>, Ankur Butala<sup>4</sup>, Najim Dehak<sup>1</sup>, Laureano Moro-Velázquez<sup>1</sup>

<sup>1</sup>Center for Language and Speech Processing, Johns Hopkins University, Baltimore, USA

<sup>2</sup>Meaning.team Inc., Baltimore, USA

<sup>3</sup>Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA

<sup>4</sup>Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD, USA  
{tcao7, afavaro1, tthebau1, jvillal7, ndehak3, laureano}@jhu.edu, {eoh9, ankur.butala}@jhmi.edu  
pzelasko@meaning.team

**Abstract:** Neurological disorders (NDs) display a variety of clinical manifestations and represent a challenge to global health. Few diagnostic metrics can be reliably employed to distinguish NDs without costly imaging techniques or invasive procedures. Vowel articulation from speech can be assessed in a non-invasive manner and can provide potential biomarkers of NDs. In this study, we employed a validated library called VSAmeter to evaluate the articulation of vowels from people with different NDs, i.e., Parkinson's and Alzheimer's Disease (PD, AD), and Parkinson's mimics (PDM) and compared them with a control group (CN). We also analyzed the effect of different types of speech tasks. In addition to Vowel Space Area and Vowel Articulation Index, a new metric, Cumulative Pair-wise Vowel Distance (CPVD), was explored. CPVD- $k$ , measuring the top  $k$  shortest distance pairs between vowels in the F1-F2 space, provided significant differences between PD and PDM, and the control group in two different speech tasks.

**Keywords:** Neurological Disorders, Atypical Speech, Parkinson's disease, Vowel Space Area

## I. INTRODUCTION

Neurological Disorders (NDs) are diseases of the central and peripheral nervous system, which vary in signs, symptoms, speed of onset, or progression of disease [1]. Among all NDs, Alzheimer's disease (AD) is the most common type of dementia, followed by Vascular Dementia and Lewy Body Disease [2]. Parkinson's Disease (PD) is also a common ND caused by the neurodegenerative process in the substantia nigra, affecting dopamine production [3]. Early detection of NDs is crucial for targeted interventions that may slow the progression of these conditions [4], but there are very few diagnostic tools that can be reliably employed to easily distinguish neurological disorders without using costly imaging methods or invasive procedures such as lumbar punctures to examine cerebrospinal fluid. Distinguishing PD from many conditions such atypical parkinsonian disorders or secondary parkinsonism can be challenging as these diseases share many signs and symptoms. Patients with these PD mimics (PDM) are

often misdiagnosed as having PD [5]. Consequently, more precise biomarkers are needed.

Speech and language impairments often occur in various NDs due to motor and cognitive decline [6]. Features extracted from speech and language can serve as rapid, cost-efficient, and non-invasive biomarkers of NDs. Studies have used speech formants, particularly Vowel space area (VSA)-related features, to assess PD and AD [7], [8]. VSA measures the area in the first and second formant frequency plane, in which each corner is determined by a target vowel [9], [10]. In English, the VSA is usually constructed by the Euclidean distances in the F1/F2 plane of the corner vowels /i/, /u/, and /a/, i.e., triangular VSA (tVSA), or the corner vowels /i/, /u/, /a/ and /ae/, i.e., quadrilateral VSA (qVSA) [7]. VSA has been found to be smaller in dysarthric speakers and patients with PD compared to CNs [11]. A recent study also found a smaller VSA in AD patients compared to the control group, although the results of that study are inconclusive due to a significant age difference between groups, which could motivate the VSA differences [8]. Another metric, Vowel Articulation Index (VAI), was proposed to reduce inter-speaker variability inherent to VSA. Some studies suggest that VAI is significantly reduced in PD patients [7]. These and most previous studies focused on single speech tasks and one specific ND at a time, with limited exploration of differences between NDs. In this study, we utilized a validated library, VSAmeter [10], to automatically measure and compare VSA and related features like VAI in PD, AD, and PDM patients. Speech recordings from different tasks were used, e.g., Rainbow Passage (RP) reading task (read speech) and Cookie Theft Picture (CTP) description task (spontaneous speech). A new formant-based metric called CPVD was explored. The code to reproduce our experiments is publicly available<sup>1</sup>.

<sup>1</sup><https://github.com/Neuro-Logical/VSAmeter>



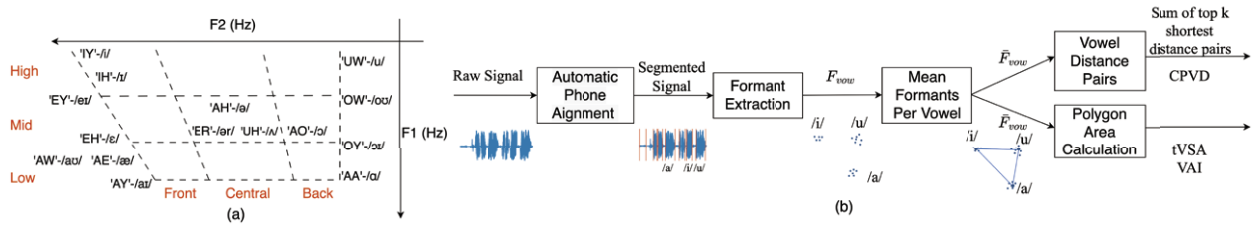


Fig. 1: A block diagram of the pipeline to obtain the metrics included in this study. (a) Fifteen vowel sounds in American English are represented as  $K/P$ , where  $K$  is the symbol employed by the Kaldi forced alignment and  $P$  is the equivalent IPA symbol. The dashed lines divide it into several areas, indicating the tongue’s position from front to back and high to low, which is the primary factor in vowel shaping. (b) The pipeline of VSAmeter.

## II. MATERIALS

NeuroLogical Signals (NLS) [6], [12] is a dataset that contains spoken responses to different tasks, e.g., object naming, picture description, or passage reading. The participants, who signed an informed consent, varied from different NDs as well as control (CN) participants. Speech signals were recorded with a headset microphone and a 24 kHz sampling rate.

TABLE I: Demographics of the study population in NLS dataset.

Category	Sample			Age	
	Total	Female	Male	Average	Range
PD	27	10	17	66.55	41-82
AD/MCI	15	3	12	70.80	57-84
PDM	14	7	7	57.29	43-74
CN	33	15	18	68.64	42-79

PD, AD/Mild Cognitive Impairment (MCI), and PDM groups are considered in this study. The participants in the PDM group were diagnosed with atypical Parkinsonian Disorders or secondary parkinsonism by the highest clinical diagnostic criteria, including: multiple system atrophy, dystonia, spinocerebellar ataxias, dementia with Lewy body in mixed pathology, corticobasal syndrome, and essential tremor. Even though cognitive and speech disorders might differ across the diseases contained in the PDM group, we group these subjects together as a neurodegenerative control group, representative of what neurologists and geriatricians can see in their daily practice in contrast to PD or AD. This group allows us to observe if the analyzed features are PD or AD specific. The AD/MCI group consists of participants with AD or Mild Cognitive Impairment due to AD by biomarker positivity (prodromal AD). Sex and age distribution for each experimental group are reported in Table I. All the participants were given unlimited time to read the RP and 1 min to describe the CTP. Transcriptions were generated using Whisper<sup>2</sup>,

<sup>2</sup>at: <https://openai.com/blog/whisper/>

and manually supervised and corrected if necessary.

## III. METHODS

The tVSA and VAI values for each participant were firstly calculated by the VSAmeter [10]. Then, one new formant-related metric, called CPVD, was explored to distinguish people with NDs from the CN group. The scheme of the feature extraction pipeline is included in Figure 1. Finally, we conducted pairwise Kruskal–Wallis tests for each metric among the four groups of participants based on the RP reading task and CTP description task, which determined whether there were statistically significant differences between the median of each group per task.

### A. tVSA and VAI

In the VSAmeter, Kaldi is employed to build a forced alignment model that can align and segment the speech recordings automatically [10]. The F1 and F2 formants are extracted by the KARMA algorithm [13] with the default settings introduced in [10], [13] for several targeted vowels. For each vowel in a transcription, we obtained the formants at 35% temporal point of the vowel segment and averaged across all the repetitions of each target vowel, as indicated in [10]. We used /i/, /u/, /a/ to calculate tVSA and VAI among the four experimental groups based on two different tasks.

### B. Cumulative Pair-wise Vowel Distance (CPVD)

We propose the CPVD metric to investigate if the distance between each pair of possible vowels in a certain language (English, in our case) in the F1/F2 plane can provide insights about vowel misarticulation in speakers with NDs. As dysarthric speakers tend to have problems articulating, measuring the individual distances between pairs of vowels can also provide insights into how close similar vowels are in the F1/F2 plane. We hypothesize that, even when there is no overall centralization or reduction of the tVSA, some people with dysarthria can have problems articulating certain vowels, which would lead to vowel spirantization, i.e., two vowels that are so close that might overlap.

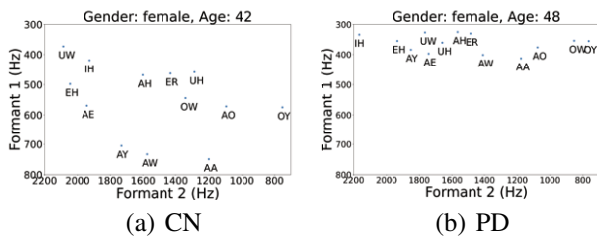


Fig. 2: Two examples of vowel sounds in F1/F2 plane generated by VSAmeter from NLS dataset.

Fifteen vowel sounds in American English, based on the International Phonetic Alphabet (IPA), were selected [14]. Their relative typical localization with respect to the F1/F2 plane is shown in Fig. 1 (a). We calculated the Euclidean pair-wise distances between the 15 vowels (all possible combinations) in the F1/F2 plane, and the top  $k$  ( $k = [5, 10, 15, 20]$ ) shortest distance pairs for each participant were summed up as a new feature named CPVD- $k$ .

Two examples of vowel sounds in F1/F2 plane are shown in Fig. 2. The participant from the PD group (right) tends to have the vowels closer to each other than the participant in the CN group.

#### IV. RESULTS

The results of pairwise Kruskal–Wallis tests for each metric among the four different groups of participants are reported based on two different tasks: the RP reading task (read speech) and CTP description task (spontaneous speech). The Kruskal–Wallis test, a rank-based non-parametric approach, is considered for testing whether there is a significant difference in medians between two groups [15] with non-normal distribution of the data. The statistical analyses of the significance of the differences between groups are shown in Fig. 3.

#### V. DISCUSSION AND CONCLUSIONS

As shown in Fig. 3, participants in PD and PDM tend to have a statistically significant smaller median of tVSA than those in the CN group during the RP (read speech) task but not in the CTP (spontaneous speech) task. These results suggest that the reduction of typical vowel space-related metrics for PD or other diseases can be task-dependent, which is consistent with the findings in [8]. In this regard, on the RP task, each participant reads the same passage so that each recording contains the same vowels, making it easier to compare differences in vowel formants across groups than in the CTP (spontaneous speech) task, in which each participant uses different words to describe the picture and, hence, vowel distribution. In a hypothetical case, some speakers might not even use some of the corner vowels making it impossible to calculate tVSA.

This could make the spontaneous speech task less suitable to compare tVSA across groups. Besides, our results suggest that people with AD tend to have a larger tVSA than those with PD, and their values are more similar to the control group. Similar trends are observed in CTP (spontaneous speech) task in Fig. 3, although no statistically significant difference between groups was found for that task and tVSA. As suggested by [8], during spontaneous and read speech tasks, the VSA is reduced in people with AD with respect to control speakers. However, we did not find statistically significant differences. Regarding VAI, it can be observed in Fig. 3 that PDM has statistically significant differences with the CN group in RP (read speech) task. However, PD shows no significant difference in median compared with the CN group. In the CTP task, the median of VAI for the PD groups is significantly smaller than that for the AD group. Except for these two cases, no statistically significant differences are observed between each pair among the four experimental groups for VAI.

CPVD-20 for the CN group is significantly larger than that for the PD group in the RP task in Fig. 3. This significant trend always holds as  $k$  decreases from 20 to 5. When  $k = 20$ , the sum of the shortest vowel distance pairs for the PD and PDM group are both significantly smaller than that for the CN group in CTP (spontaneous speech) task. When  $k$  decreases, i.e.,  $k = 15, 10, 5$ , these differences are not significant anymore. However, this metric provides significant differences between control speakers and those with motor impairment (PD and PDM) in spontaneous speech, where comparing subjects is more difficult for tVSA. This proposed metric is more effective in differentiating PD or PDM from controls when using spontaneous speech, and could be complementary to traditional VSA measures. Moreover, differences in articulation between close vowels in the F1/F2 space can be indicative of dysarthria or neurological disorders, and these are not considered by tVSA and VAI. Regarding the AD group, these always showed higher tVSA, VAI, and CPVD than PD and PDM subjects and no significant differences with the CN group. In conclusion, results suggest that tVSA, VAI, and CPVD do not provide differentiation between the AD and CN groups; moreover, these metrics are not PD-specific biomarkers as no significant differences were found between the PD and PDM groups.

In the future, additional experiments are needed to explore the impact of other factors, e.g., language or speech task length, as reduction of vowel space-related metrics in NDs was found to be task-dependent. The VSAmeter will be applied to PD datasets in Spanish to assess significant differences between PD patients

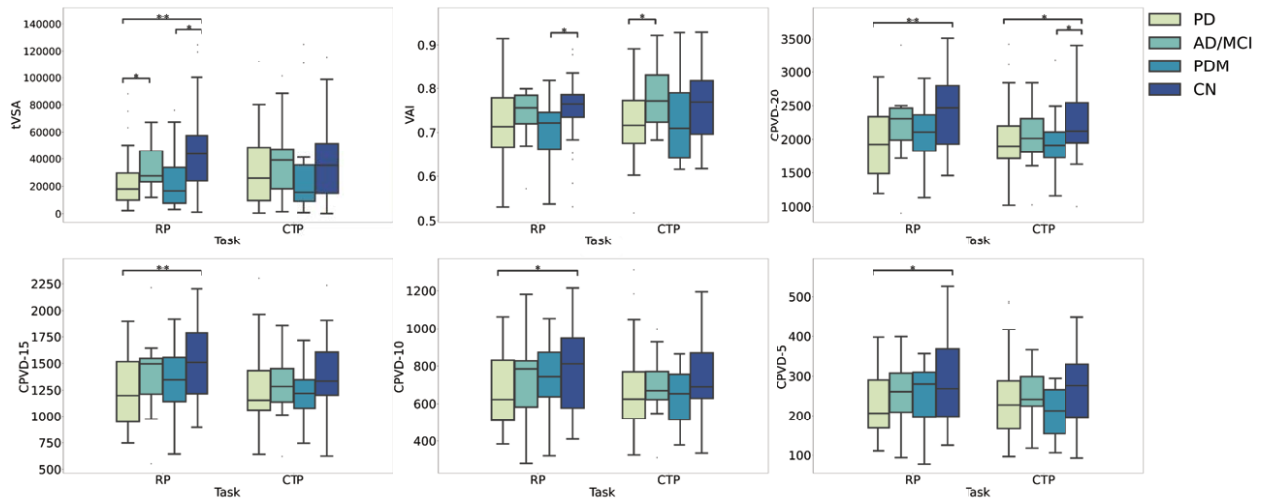


Fig. 3: Boxplots of vowel space metrics based on formant for RP and CTP tasks:  $tVSA$  (top left),  $VAI$  (top middle),  $CPVD-20$  (top right),  $CPVD-15$  (bottom left),  $CPVD-10$  (bottom middle) and  $CPVD-5$  (bottom right). Asterisks are employed to highlight statistically significant differences in the median between groups, where \* means  $0.01 < p \leq 0.05$ , and \*\* means  $0.001 < p \leq 0.01$

and CNs. New metrics and the influence of other speech tasks like shorter speech segments on VSA-related metrics will be studied. The promising VSA-related metrics measured by VSAmeter and the newly proposed  $CPVD-k$  metric will be combined with other features in future studies to distinguish different NDs. Besides, a longitudinal study is underway with a larger dataset and recordings collected at different stages of the disease. Accuracy of automatic VSA measurements and phone alignment will also be improved.

#### REFERENCES

- [1] W. H. Organization, *Neurological disorders: public health challenges*. World Health Organization, 2006.
- [2] S. Gauthier, P. Rosa-Neto, J. A. Morais, and C. Webster, "World alzheimer report 2021: Journey through the diagnosis of dementia," *Alzheimer's Disease International*, pp. 17–29, 2021.
- [3] R. F. Pfeiffer, "Non-motor symptoms in parkinson's disease," *Parkinsonism & related disorders*, vol. 22, pp. S119–S122, 2016.
- [4] R. Sitruk-Ware, B. Bonsack, R. Brinton, M. Schumacher, N. Kumar, J.-Y. Lee, V. Castelli, S. Corey, A. Coats, N. Sadanandan, *et al.*, "Progress in progestin-based therapies for neurological disorders," *Neuroscience & Biobehavioral Reviews*, vol. 122, pp. 38–65, 2021.
- [5] K. Ali and H. R. Morris, "Parkinson's disease: chameleons and mimics," *Practical neurology*, vol. 15, no. 1, pp. 14–25, 2015.
- [6] A. Favaro, C. Motley, T. Cao, M. Iglesias, A. Butala, E. S. Oh, R. D. Stevens, J. Villalba, N. Dehak, and L. Moro-Velázquez, "A multi-modal array of interpretable features to evaluate language and speech patterns in different neurological disorders," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 532–539.
- [7] S. Sapir, L. O. Ramig, J. L. Spielman, and C. Fox, "Acoustic metrics of vowel articulation in parkinson's disease: vowel space area (vsa) vs. vowel articulation index (vai)," in *International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2011.
- [8] A. Shamei, Y. Liu, and B. Gick, "Reduction of vowel space in alzheimer's disease," *JASA Express Letters*, vol. 3, no. 3, p. 035202, 2023.
- [9] S. Sandoval, V. Berisha, R. L. Utianski, J. M. Liss, and A. Spanias, "Automatic assessment of vowel space area," *The Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. EL477–EL483, 2013.
- [10] T. Cao, L. Moro-Velázquez, P. Želasko, J. Villalba, and N. Dehak, "Vsameter: Evaluation of a new open-source tool to measure vowel space area and related metrics," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 517–524.
- [11] S. Sapir, J. L. Spielman, L. O. Ramig, B. H. Story, and C. Fox, "Effects of intensive voice treatment (the lee silverman voice treatment [lsvt]) on vowel articulation in dysarthric individuals with idiopathic parkinson disease: acoustic and perceptual findings," *Journal of Speech, Language, and Hearing Research*, 2007.
- [12] A. Favaro, L. Moro-Velazquez, A. Butala, C. Motley, T. Cao, R. D. Stevens, J. Villalba, and N. Dehak, "Multilingual evaluation of interpretable biomarkers to represent language and speech patterns in parkinson's disease," *Frontiers in Neurology*, vol. 14, p. 317, 2023.
- [13] D. D. Mehta, D. Rudoy, and P. J. Wolfe, "Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1732–1746, 2012.
- [14] S. G. Lambacher, W. L. Martens, K. Kakehi, C. A. Marasinghe, and G. Molholt, "The effects of identification training on the identification and production of american english vowels by native speakers of japanese," *Applied Psycholinguistics*, vol. 26, no. 2, pp. 227–247, 2005.
- [15] E. Ostertagova, O. Ostertag, and J. Kováč, "Methodology and application of the kruskal-wallis test," in *Applied mechanics and materials*, vol. 611. Trans Tech Publ, 2014, pp. 115–120.

# EXAMINING THE DIRECTIVITY CHARACTERISTICS OF GREEK SUNG VOWELS ON FORMANT FREQUENCIES

G. Dedousis, K. Bakogiannis, A. Andreopoulou, A. Georgaki

Laboratory of Music Acoustics and Technology (LabMAT), National and Kapodistrian University of Athens  
gdedousis@music.uoa.gr, k.bakogiannis@music.uoa.gr, aandreo@music.uoa.gr, georgaki@music.uoa.gr

**Abstract:** This work presents preliminary findings on the connection between formant frequencies and directivity of the Greek singing voice, aiming to contribute to vocal production research, and to the design of simulation, auralization, and virtual reality systems with application in speech and music domains. The present study focuses on the recordings of four professional singers, two in classical music and two in Byzantine chant, recorded in a sound-treated space, singing the Greek vowels at different pitches. Directivity results are reported for each vowel in third-octave bands centered on the first three formant frequencies (F1, F2, F3) of each singer.

**Keywords:** Directivity, formants, Greek vowels, singing voice

## I. INTRODUCTION

Extensive research has been conducted on the directional characteristics of the human voice, both spoken and sang. Directivity data has been studied on the horizontal plane on both the horizontal and vertical planes or on a complete sphere. Studies comparing the directivity characteristics of the singing and speaking voice have shown evidence of variations between the two. For example, that classical singers tend to have higher directivity compared to speech [1].

Kocon & Monson [2] reported that vocal tract configuration and mouth opening change during speech which has an influence on vocal radiation. It has also been shown that mouth opening affects the directivity of the singing voice [3]. Directivity can also be affected by one's posture and head inclination, their vocal tract configurations [4] and the spectral emphasis that can be manipulated through singing techniques. One's torso and head size also have an impact on directivity, but these are fixed parameters [3]. It has also been suggested, based on a study with sustained German vowels, that directivity can be affected by the position of the sound in the mouth cavity (e.g., front vowels are more directional), pitch (e.g., higher directivity in higher pitches) and less on the type of phonation (increased in pressed versus breathy) [1].

However, factors such as the opening of one's mouth and the shape of their vocal tract are considered to be important for the frequency of the formants, with the first and second formant mainly shaping the vowel quality and the third, fourth and fifth, the voice quality ("timbre") [5]. These frequencies also affect the spectrum of the vowel [6]. Moreover, there are singing techniques that can be used to alter these frequencies, (e.g., widening one's lips or changing one's jaw opening can raise the first formant [6]). Such formant frequencies adjustments (used by singers to modify the spectral content of their voices) can affect directivity [7]. Thus, it would be interesting to study vocal directivity centered around formant frequencies.

Although, some research has been done concerning the phonetics aspects of the Greek language and, especially, the formants corresponding to it (e.g., [8]), little research has been conducted concerning the relevant formants in singing [9], with this also being the case with directivity [10]. Furthermore, studies on this topic cannot be easily compared mainly due to the lack of a common measurement protocol [8].

This study focuses on the horizontal plane directivity characteristics of the Greek sung vowels sounds (monophthongs) /a/ (α), /e/ (ε, αι), /i/ (ι, η, υ, οι, ει), /o/ (ο, ω), /u/ (ου), when investigated on the relevant formant frequencies, taking into consideration the singing style of two professional classical singers and two Byzantine chant singers, a style of singing where studies focusing on formants [11], formant tuning [12] and vocal ornamentation [13] are quite scarce.

## II. METHODOLOGY

This research is part of a larger study focused on the sound projection and directivity characteristics of various traditional Greek musical instruments and professional singers, replicating realistic performance scenarios encountered by musicians. Measurements were conducted using 29 RODE-M5, small diaphragm condenser microphones, placed symmetrically on a hemispherical thin-shell structure with a radius of 158,5cm at four elevations (+90°, +30°, 0°, -30°), which was set up in the hemi-anechoic live room at the

facilities of the Laboratory of Music Acoustics and Technology (LabMAT), NKUA. This study, because of the limited space, will only focus on the measurements on the horizontal plane, which consists of 12 microphones placed at 30° azimuthal increments. The individual impulse responses of the microphones were collected using ScanIR [14] on an M1 MacBook Pro 2020 running Matlab 2021a.

Participants were positioned at the center of the microphone array in a standing posture. The height of the configuration was adjusted using elevation probes to align with each singer's mouth position. A plumb and laser beams were used to maintain proper alignment throughout the measurements, while creating a more natural singing experience, allowing for small body and head movements related to technique and vocal projection. These movements have been shown not to have any perceptual impact on the collected directivity data [10].

Prior to measurements, all input signals were level calibrated using pink noise (78 dBA), generated by a Brüel & Kjær omnidirectional loudspeaker (OmniPower SoundSource Type 4292-L) placed at the singer's position, ensuring consistent levels across the array microphones with a tolerance of  $\pm 0.5$  dB. The pink noise signals were analyzed in third-octave bands, and calibration levels were obtained to equalize the RMS levels in each band. Data acquisition was performed using two Yamaha TF1 digital mixers (interconnected via DANTE), utilizing their built-in preamplifiers, and recorded on an i5 laptop running Cubase 11.

Once aligned with the microphone array, participants (2 male professional classical singers and 2 male Byzantine chant singers) were asked to intone each of the five vowel sounds /a/ ( $\alpha$ ), /e/ ( $\epsilon$ ,  $\alpha$ ), /i/ ( $\iota$ ,  $\eta$ ,  $\upsilon$ ,  $\omicron$ ,  $\epsilon$ ), /o/ ( $\omicron$ ,  $\omega$ ), /u/ ( $\upsilon$ ) twice on pitches A2, E3 and C#4 for about two seconds each. The audio recordings per participant and microphone were deconvolved with the microphone responses to minimize the impact of the measurement setup on the analyzed data and level calibrated per third-octave band, to obtain omnidirectional responses. In order to suppress the impact of noise introduced in the data by frequency bands with insufficient energy, the signal-to-noise level was calculated and a noise floor threshold was derived suppressing any data within 3dB of its level [10]. Formant analysis was done using Fast Track [15], carefully adjusting the parameters according to the participants.

### III. RESULTS

For each vowel and pitch, the horizontal plane directivity data was calculated on third-octave bands centered around the corresponding F1, F2, and F3 frequencies of each singer. These frequencies were

calculated from the mean values of the nine front microphones (front, front-left, front-right on the three elevation angles: +30°, 0°, -30°).

The directivity of all vowels at the F1 frequency exhibit a more omnidirectional-like pattern due to the low frequency range of F1 (around 250Hz to 1000Hz), while F2 and F3 frequency bands exhibit varying directional characteristics. For brevity, "Fig. 1" depicts the difference in directivity between the F1 and F2, and the F1 and F3 frequency bands for the four measured participants: Classical Male 1 (CM1), Classical Male 2 (CM2), Byzantine Male 1 (BM1), Byzantine Male 2 (BM2).

As can be seen, the projection of vowel /a/ in the F2 and F3 formant regions is less omni than in F1. However, that is not always the case, as CM2 seems to have greater dispersion in the F2 region (1044,5Hz) at C#4, with differences ranging from 2,23dB (side) to 4,85dB (front and back). CM1 has greater dispersion at E3 and C#4 in the F2 region (1005Hz and 1064Hz, respectively), only on the sides and back. BM1 appears to have greater projection in the front, across all three pitches, in the F2 region (1065Hz, 1070Hz and 1105Hz respectively), but the difference is very small (0,68dB to 1,99dB).

Vowel /e/ seems to exhibit more directional properties in the F2 and F3 regions, with the F1 region being the one with the most dispersion. There is a slight indication that CM2 might exhibit greater dispersion at E3 in the F2 region (1505Hz), but the difference is very small (-0,35dB to 2,83dB in 30°). Similarly, BM2 seems to have slightly greater dispersion in the F2 region (1626,5Hz) at the front and sides, at A2.

For the vowel /i/, all singers but BM2, appear to have significantly greater dispersion in the F2 and F3 regions, to all directions, when singing C#4. In addition, CM1 has greater dispersion in the F2 and F3 regions (1852,2Hz, 2702Hz, respectively) at A2, across all directions. BM1 seems to have greater dispersion in F2 and F3 regions, at A2 and E3, on the front and sides, but to a lesser extent.

Moving on to the /o/ vowel, the data shows that the F2 and F3 regions appear to have less dispersion than the F1, to most of the singers. However, BM2 appears to have greater dispersion in the F2 region (755,5Hz), at A2, to all directions. Similarly, there is slight indication that CM1 also exhibits the same behavior in F2 region, albeit to a lesser extent.

Finally, vowel /u/, exhibits similar behavior to vowel /i/, i.e., CM1, CM2 and BM1 appear to have greater dispersion in the F2 region (847,5Hz) at C#4 towards all directions, while CM1 and CM2 have similar results in their F3 region (2538,5Hz, 2565,5Hz respectively). Additionally, there is a slight indication that, for CM1, CM2, and BM1, the F2 region exhibits greater dispersion than F1 at E3, although to a lesser extent.

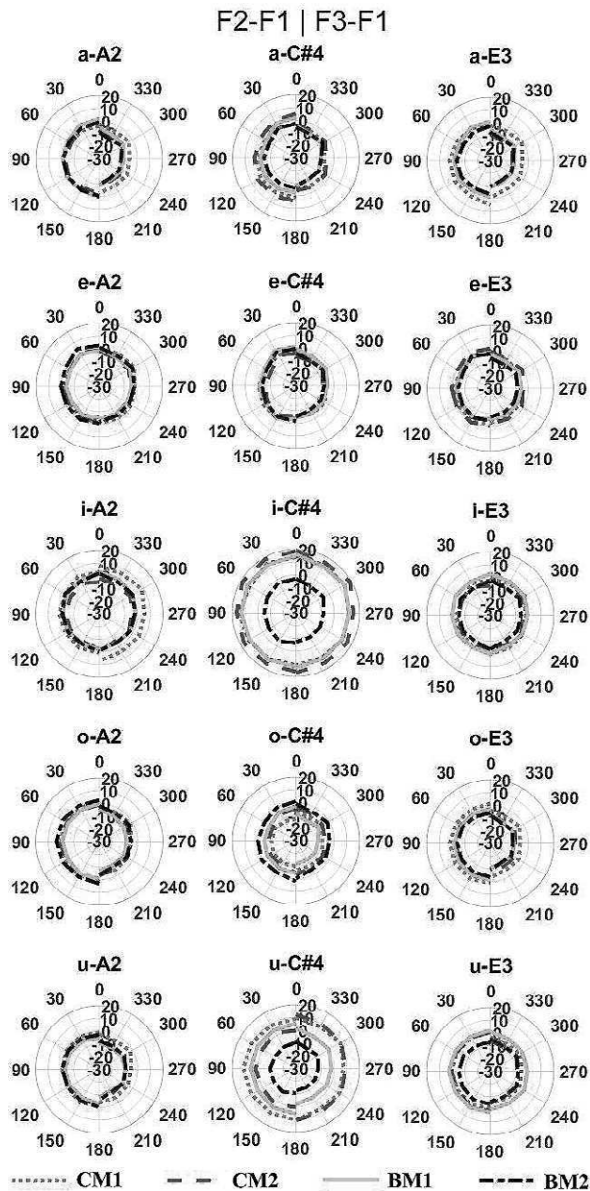


Fig.1 Directivity plots for the four singers, on the horizontal plane (5 vowels, 3 pitches). The left side of each plot shows the directivity of the 2<sup>nd</sup> formant in relation to the 1<sup>st</sup> (F2-F1), and the right side the 3<sup>rd</sup> formant in relation to the 1<sup>st</sup> (F3-F1).

#### IV. DISCUSSION

The reported results are in line with the relevant literature concerning the left/right symmetric projection of the singing voice [10], [16], [17]. Directivity patterns can change according to pitch and center frequency they are measured [1], [18]. However, it is difficult to infer a safe conclusion concerning the type of singing (classical, Byzantine chant) and the way it might affect directionality, as no specific pattern seems to emerge, although there is some indication that classical singers tend to widen their projection at the F3 region (ranging

from 2219,5Hz to 2719Hz), in relation to the F1 at C#4 of the Greek vowels /i/ and /u/. Furthermore, BM1 and BM2, although both Byzantine chant singers, have different values for the vowel /i/ in the F2 and F3 region at C#4 and the vowel /u/ in the F2 region at C#4 and E3, where BM1 appears to follow the classical singers' pattern. A larger sample size of Byzantine chant singers could provide more insight into this difference.

It has also been reported that, on a scale from the most to the least directional, vowels appear in the following order /a/, /e/, /i/, /o/, /u/ [18], [19], which seems to correspond to the opening of one's mouth [3], [19]. Our data shows that the Greek /u/ and /i/ vowels are less directional, especially at C#4 in the F3 region, while the /e/ appears to be the most directional in relation to F1. This finding can partly be attributed to the different mouth opening of the singers, as it is suggested that the larger the opening (such as in the case of /a/) the narrower the direction [3], and, partly, to the way the Greek vowels may be pronounced in relation to other languages researched, which also has implication on the formant frequencies (especially the 1<sup>st</sup> and the 2<sup>nd</sup>).

One limitation of the current study is the small number of the participants which prohibits overall statistics. Although many studies in the relevant literature have been carried out with similar sample size, a larger sample could provide more generalized results. Another shortcoming is that the formant frequencies are calculated as a mean of nine microphones rather than one microphone placed closely in front of the mouth of the singer. Although it has been suggested that the recording distance can affect the formant frequencies, this seems to mostly affect the weaker formants and the resulting deviation could be found in the range of same speaker variations [20].

#### V. CONCLUSION

This study considered the directivity patterns of Greek sustained vowels in third-octave bands centered on the first three formant frequencies (F1, F2, F3), of two professional classical and two Byzantine chant singers respectively. The results extended the relevant literature, given the very limited published research on formant and directivity analysis on the singing voice in the Standard Modern Greek language and Byzantine chant. Future work will focus on centering the directivity at the frequencies of the fourth and fifth formants (F4, F5), while offering more insight into the directivity of the singer's formant region. Additionally, the analysis will expand to include directions beyond the horizontal plane and more singers of various training levels and singing genres. Our aim is to find possible connections between the Greek language, formant analysis, singing genres and training, with directivity and vocal projection.



## REFERENCES

- [1] M. Brandner, M. Frank, and A. Sontacchi, "Horizontal and Vertical Voice Directivity Characteristics of Sung Vowels in Classical Singing," *Acoustics*, vol. 4, no. 4, Art. no. 4, Dec. 2022, doi: 10.3390/acoustics4040051.
- [2] P. Kocon and B. B. Monson, "Horizontal directivity patterns differ between vowels extracted from running speech," *J. Acoust. Soc. Am.*, vol. 144, no. 1, pp. EL7–EL12, Jul. 2018, doi: 10.1121/1.5044508.
- [3] M. Brandner, R. Blandin, M. Frank, and A. Sontacchi, "A pilot study on the influence of mouth configuration and torso on singing voice directivity," *J. Acoust. Soc. Am.*, vol. 148, no. 3, pp. 1169–1180, Sep. 2020, doi: 10.1121/10.0001736.
- [4] R. Blandin and M. Brandner, "Influence of the vocal tract on voice directivity," in *PROCEEDINGS of the 23rd International Congress on Acoustics*, Aachen, Germany, Sep. 2019, pp. 1795–1801.
- [5] B. H. Story, "The vocal tract in singing," in *The Oxford handbook of singing*, G. F. Welch, D. M. Howard, and J. Nix, Eds., Oxford University Press, 2016, pp. 131–211.
- [6] J. Sundberg, "The Singing Voice," in *The Oxford Handbook of Voice Perception*, S. Frühholz and P. Belin, Eds., 1st ed. in Oxford Library of Psychology, vol. 1. Oxford University Press, 2019, pp. 117–142. doi: 10.1093/oxfordhb/9780198743187.013.6.
- [7] D. Cabrera, P. J. Davis, and A. Connolly, "Long-Term Horizontal Vocal Directivity of Opera Singers: Effects of Singing Projection and Acoustic Environment," *J. Voice*, vol. 25, no. 6, pp. e291–e303, Nov. 2011, doi: 10.1016/j.jvoice.2010.03.001.
- [8] A. Arvaniti, "Greek Phonetics, The State of the Art," *J. Greek Linguist.*, vol. 8, no. 1, pp. 97–208, 2007, doi: 10.1075/jgl.8.08arv.
- [9] S. Kalozakis, A. Georgaki, and G. Kouroupetroglou, "FORMANT TUNING IN CRETAN RIZITIKO SINGING," in *Models and Analysis of Vocal Emissions for Biomedical Applications: 12th International Workshop*, C. Manfredi, Ed., Firenze, Italy: Firenze University Press, 2021, pp. 127–130.
- [10] K. Bakogiannis, G. Dedousis, Y. Malafis, and A. Andreopoulou, "On the spherical directivity and formant analysis of the singing voice; a case study of professional singers in Greek Classical and Byzantine music," presented at the Audio Engineering Society Convention 153, Audio Engineering Society, Oct. 2022. Accessed: Jun. 21, 2023. [Online]. Available: <https://www.aes.org/e-lib/browse.cfm?elib=21960>
- [11] D. Delviniotis, "Analysis of Byzantine music - Greek orthodox chant- by using signal processing techniques," National and Kapodistrian University of Athens, Greece, 2002. [Online]. Available: <http://dx.doi.org/10.12681/eadd/13220>
- [12] G. Chrysochoidis, G. Kouroupetroglou, D. Delviniotis, and S. Theodoridis, "Formant tuning in Byzantine chant," in *Proceedings of the Int. Conference Sound and Music Computer*, Stockholm, Sweden, 2013, pp. 217–223.
- [13] D. S. Delviniotis and S. Theodoridis, "On Exploring Vocal Ornamentation in Byzantine Chant," *J. Voice*, vol. 33, no. 2, p. 256.e17-256.e34, Mar. 2019, doi: 10.1016/j.jvoice.2017.10.016.
- [14] J. Vanasse, A. Genovese, and A. Roginska, "Multichannel Impulse Response Measurements in MATLAB: An Update on ScanIR," presented at the Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio, Audio Engineering Society, Mar. 2019. Accessed: Jun. 22, 2023. [Online]. Available: <https://www.aes.org/e-lib/browse.cfm?elib=20416>
- [15] S. Barreda, "Fast Track: fast (nearly) automatic formant-tracking using Praat," *Linguist. Vanguard*, vol. 7, no. 1, Jan. 2021, doi: 10.1515/lingvan-2020-0051.
- [16] B. Katz and C. d'Alessandro, "Directivity measurements of the singing voice," presented at the 19th International Congress on Acoustics (ICA 2007), Madrid, Spain, Sep. 2007, p. 6. Accessed: Jun. 07, 2023. [Online]. Available: <https://hal.science/hal-01712590>
- [17] B. B. Boren and A. Roginska, "Sound radiation of trained vocalizers," *Proc. Meet. Acoust.*, vol. 19, no. 1, p. 035025, Jun. 2013, doi: 10.1121/1.4800053.
- [18] A. H. Marshall and J. Meyer, "The Directivity and Auditory Impressions of Singers," *Acta Acust. United Acust.*, vol. 58, no. 3, pp. 130–140, Aug. 1985.
- [19] C. Pörschmann and J. M. Arend, "Investigating phoneme-dependencies of spherical voice directivity patterns," *J. Acoust. Soc. Am.*, vol. 149, no. 6, pp. 4553–4564, Jun. 2021, doi: 10.1121/10.0005401.
- [20] E. B. Brixen and S. Christensen, "Influence of Recording Distance and Direction on the Analysis of Voice Formants - Initial Considerations," presented at the AES 131st Convention, New York, NY, USA: Audio Engineering Society, Oct. 2011. Accessed: Jun. 07, 2023. [Online]. Available: <https://www.aes.org/e-lib/online/browse.cfm?elib=16021>

# AI TECHNIQUES APPLIED TO ACOUSTICAL FEATURES OF PARALYTIC DYSPHONIA VERSUS DYSPHONIA DUE TO BENIGN VOCAL FOLD MASSES

Federico Calà<sup>1</sup>, Lorenzo Frassinetti<sup>1,2</sup>, Giovanna Cantarella<sup>3\*</sup>,  
Ludovica Battilocchi<sup>3</sup>, Giulia Buccichini<sup>3</sup>, Antonio Lanata<sup>1\*</sup>, Claudia Manfredi<sup>1\*</sup>

<sup>1</sup> Department of Information Engineering, Università degli Studi di Firenze, Firenze, Italy

<sup>2</sup> Department of Information Engineering, Università degli Studi di Pisa, Pisa, Italy

<sup>3</sup> IRCSS Ca' Granda Foundation, Ospedale Maggiore Policlinico Milano, Milano, Italy

[federico.cala@unifi.it](mailto:federico.cala@unifi.it), [lorenzo.frassinetti@unifi.it](mailto:lorenzo.frassinetti@unifi.it), [giovanna.cantarella@policlinico.mi.it](mailto:giovanna.cantarella@policlinico.mi.it),  
[ludovicabattilocchi@gmail.com](mailto:ludovicabattilocchi@gmail.com), [giulia.buccichini@unimi.it](mailto:giulia.buccichini@unimi.it), [antonio.lanata@unifi.it](mailto:antonio.lanata@unifi.it), [claudia.manfredi@unifi.it](mailto:claudia.manfredi@unifi.it)

\*joint project coordinators

**Abstract:** Automatic assessment of voice disorders can support otolaryngologists' diagnosis. In this study, features extracted from the sustained vowel /a/ and the Italian word /aiuole/ were considered separately and concurrently in three machine learning experiments. The dataset is made of 55 male and 100 female subjects. The aim was to distinguish between subjects diagnosed with benign lesions and unilateral vocal fold paralysis. The best classification performance was obtained by merging feature sets of both /a/ and /aiuole/, with up to 84% and 88% accuracy for female and male cohorts, respectively. Such results are also confirmed by statistical analysis, with significant differences in several parameters.

**Keywords:** voice disorders, machine learning, statistical analysis, BioVoice

## I. INTRODUCTION

Dysphonia is characterized by higher irregular pitch, lower vocal quality and loudness with respect to normophonic subjects, related to gender and age [1]. It represents one of the most evident markers of voice pathologies and disorders. The etiopathogenesis includes tissue infections and surface irritations, mechanical stress (e.g., vocal nodules and polyps), tissue changes (e.g., tumors or cysts), neuromuscular anomalies (e.g., bi- and unilateral paralysis, spasmodic dysphonia, Parkinson's disease) and cognitive impairment (e.g., Alzheimer's disease) [2]. Clinical diagnosis is typically based on direct visualization of the vocal folds by laryngoscopy [3]. Although this method represents the gold standard, several alternatives were proposed to take into account the lack of high-resolution endoscopy in decentralized ambulatories, inter-rater variability, physicians' experience, and the need to be physically present in

hospitals, which can be demanding for the elderly and severely ill patients [4]. Acoustical analysis and the application of artificial intelligence techniques can provide powerful tools to support diagnosis, monitor vocal properties after treatments, and early detect voice pathology symptoms. As pointed out in [1], the choice of the input utterance plays a relevant role in feature extraction and depends on the type of voice pathology. Most studies rely on the sustained vowel /a/ due to open vocal tract, stable tongue and jaw position [5], relative independence from language and dialects, and intonation [1]. However, sustained vowels might not fully reflect voice characteristics and do not take into account some aspects of speech that other tasks may reveal. Therefore, other utterances have been proposed, e.g., running speech [4, 6] or enumeration tasks [7]. In this paper, acoustical features are extracted both from the sustained vowel /a/ and a standardized constantly voiced Italian word /aiuole/. The aim was that of finding which vocal task performs better in distinguishing patients affected by benign lesions (BL) (nodules, polyps, cysts) and unilateral vocal fold paralysis (UVFP) and to understand whether the concurrent use of acoustical parameters from both utterance types can lead to an improved outcome. The approach proposed here was already explored in [8] but, to the authors' knowledge, it was never applied to Italian pathological voices. Therefore, it could represent a useful procedure to be used in the future also for other cases.

## II. METHODS

Adult patients (55 males, M, 100 females, F) were recruited for this study. They were diagnosed with BL (27 M, 43 F), including nodules, polyps, and cysts, and UVFP (28 M, 57 F). The voice samples were recorded using a dynamic microphone (C1000S, AKG, Wien,



Austria) at a fixed distance of 5 cm from the patient's mouth during the production of a sustained /a/ and the word /aiuole/. After manual segmentation, audio files were processed using the BioVoice [9] open-source software for feature extraction. Among them, T0(F0 min) and T0(F0 max) parameters, which respectively represent the time instant where the minimum and maximum of the fundamental frequency occur, were normalized with respect to the total duration of the utterance. This allowed obtaining a reliable estimation of the timing where the minimum or maximum of F0 occurs, i.e., at the beginning, in the middle, or at the end of each recording.

BioVoice acoustical features were used to perform a machine-learning experiment to evaluate if supervised classifiers can distinguish between BL and UVFP.

Several supervised binary classifiers were implemented and validated using a k-fold cross-validation framework (k=10) [10]. The following models were considered: Support Vector Machine (SVM, linear and with Gaussian kernel), Ensemble models (AdaBoost and RobustBoost), and k-NN (The MathWorks, Inc., Natick, MS, USA). Bayesian hyperparameter optimization was used, choosing the model with the highest Accuracy (ACC) value. The number of iterations was set to 200. Each predictor was standardized during each cross-validation step for the training and validation sets, according to the current training statistics avoiding possible data leakage. Furthermore, in order to reduce the number of predictors, the following feature selection methods were investigated: Pearson correlation, ReliefF, LASSO, and mrMR. More precisely, before ReliefF, LASSO and mrMR, the highly correlated predictors were removed from the training and validation set, retaining only those with an absolute Pearson correlation coefficient (PCC) < 0.8 in the training set [11]. For all the validated models, the True Positive Rate (TPR) for each class was also considered, hereinafter denoted as TPR1 for the BF class and TPR2 for the UVFP one.

Machine-learning models were developed considering M and F observations separately. Moreover, models were trained using three different sets of features:

- 1) Features extracted from the sustained vowel /a/.
- 2) Features extracted from the Italian word /aiuole/
- 3) Merge of the two sets of features 1) and 2).

Finally, a statistical analysis was performed, both for the F and the M cohort, considering all the features extracted by BioVoice from /a/ and /aiuole/. A Mann-Whitney test was used, with a level of significance  $\alpha=0.05$ . These tests were performed to evaluate if statistical differences exist between BL and UFVP.

### III. RESULTS

Table 1 shows the results of the F dataset. In addition to the mean and standard deviation of classifiers' performance, the number of features (NF) used for each model is also reported.

*Table 1. Cross-validation results for the F Dataset. NF=number of features. Mean  $\mu$  and standard deviation  $\sigma$ , obtained considering the performance from each validation folder are reported.*

Feature set	F Dataset				
	ACC (%)	TPR1 (%)	TPR2 (%)	NF	Model
	$\mu\pm\sigma$	$\mu\pm\sigma$	$\mu\pm\sigma$		
/ a /	80 $\pm$ 3	70 $\pm$ 7	88 $\pm$ 4	20	RobustBoost
/ aiuole/	74 $\pm$ 4	67 $\pm$ 8	80 $\pm$ 7	19	AdaBoost
/a+/aiuole/	84 $\pm$ 5	76 $\pm$ 9	90 $\pm$ 3	39	SVM

Table 2 reports the cross-validation results for the M Dataset.

*Table 2. Cross-validation results for the M Dataset. NF=number of features. Mean  $\mu$  and standard deviation  $\sigma$ , obtained considering the performance from each validation folder are reported.*

Feature set	M Dataset				
	ACC (%)	TPR1 (%)	TPR2 (%)	NF	Model
	$\mu\pm\sigma$	$\mu\pm\sigma$	$\mu\pm\sigma$		
/ a /	79 $\pm$ 4	71 $\pm$ 10	80 $\pm$ 6	16	AdaBoost
/ aiuole /	81 $\pm$ 5	85 $\pm$ 7	79 $\pm$ 10	14	RobustBoost
/a+/aiuole/	88 $\pm$ 5	90 $\pm$ 6	92 $\pm$ 5	37	AdaBoost

Tables 5 and 6 report only the significant differences for the F group for /a/ and /aiuole/, respectively. Analogously, Table 7 and 8 summarizes significant differences for the M cohort. In Tables 5-8, directions of effect between class BL and UVFP are reported. Specifically, BL $\uparrow$  UVFP $\downarrow$  denotes a feature with a higher median value for the BL class than the UVFP one.

*Table 3. Statistical results for the F dataset: acoustical features obtained from /a/.*

Feature	p-value	Direction (Median)
F2 Std	0.004	BL $\downarrow$ UVFP $\uparrow$
F2 max	0.02	BL $\downarrow$ UVFP $\uparrow$
F1 max	0.02	BL $\downarrow$ UVFP $\uparrow$
F2 mean	0.04	BL $\downarrow$ UVFP $\uparrow$
% voiced	0.045	BL $\downarrow$ UVFP $\uparrow$

Table 4. Statistical results for the F dataset: acoustical features obtained from /aiuole/

Feature	p-value	Direction (Median)
% voiced	e-04	BL↓ UVFP↑
signalDuration	0.002	BL↑ UVFP↓
F0 median	0.003	BL↓ UVFP↑
F0 mean	0.004	BL↓ UVFP↑
T0(F0 max)	0.005	BL↑ UVFP↓
T0(F0 min)	0.01	BL↑ UVFP↓
voicedDuration	0.02	BL↓ UVFP↑
F1max	0.02	BL↓ UVFP↑
F1 std	0.02	BL↓ UVFP↑

Table 5. Statistical results for the M dataset: acoustical features obtained from /a/.

Feature	p-value	Direction (Median)
F1 max	e-04	BL↓ UVFP↑
F2 std	e-04	BL↓ UVFP↑
F1 std	0.001	BL↓ UVFP↑
voicedDuration	0.001	BL↑ UVFP↓
signalDuration	0.004	BL↑ UVFP↓
Jitter	0.004	BL↓ UVFP↑
F0 std	0.009	BL↓ UVFP↑
F1 min	0.01	BL↓ UVFP↑
NNE	0.01	BL↓ UVFP↑
F1 mean	0.01	BL↓ UVFP↑
F1 median	0.03	BL↓ UVFP↑
F0 min	0.03	BL↑ UVFP↓
F2 mean	0.03	BL↓ UVFP↑

Table 6. Statistical results for the M dataset: acoustical features obtained from /aiuole/.

Feature	p-value	Direction (Median)
T0(F0 min)	e-05	BL↑ UVFP↓
signalDuration	e-04	BL↑ UVFP↓
% voiced	e-04	BL↓ UVFP↑
F0 max	0.007	BL↓ UVFP↑
F1 mean	0.008	BL↓ UVFP↑
F0 std	0.008	BL↓ UVFP↑
voicedDuration	0.02	BL↓ UVFP↑
F1max	0.03	BL↓ UVFP↑
F1 median	0.03	BL↓ UVFP↑
Jitter	0.04	BL↓ UVFP↑
F2 mean	0.04	BL↓ UVFP↑

Figure 1 shows the boxplots for the T0(F0 min) parameter for F and M cohorts.

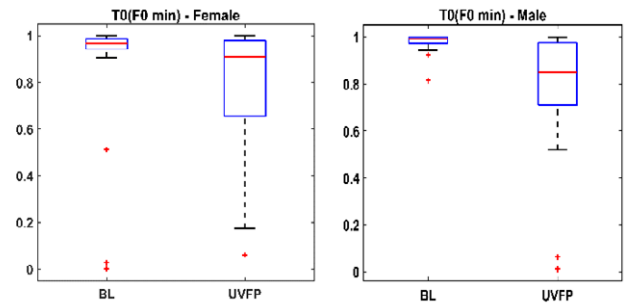


Fig. 1: Boxplots of T0(F0 min) for F (left) and M (right) datasets.

#### IV. DISCUSSION

This work proposes a voice pathology binary detection system based on machine learning techniques and acoustical parameters extracted from the sustained vowel /a/ and a constantly voiced Italian word /aiuole/. It was investigated whether the vocal task can influence the classification performance and if the combination of their features may lead to improvements [8]. Tables 1 and 2 show that models based on features extracted from a single utterance type present a similar global accuracy, in line with [6]. Table 1 shows that for the F group, the UVFP class is characterized by a higher TPR when considering both vocal tasks. This outcome can be of interest in the clinical practice since a model that can reduce the number of false positives for this pathology might spare patients undergoing long and expensive diagnostic procedures (e.g., MRI, CT). On the contrary, vocal task influences TPR direction for the M cohort: a higher TPR for the BL class can be noted in the case of /aiuole/, whereas a higher TPR for UVFP occurs with /a/. This might suggest that UFVP can be better detected with a sustained phonation rather than an articulation task.

Moreover, the concurrent use of /a/ and /aiuole/ acoustical features led to models' performance improvements. Though there is an increase of 10% for the M cohort, this result should be taken with caution as these classifiers were trained and validated with a smaller number of observations with respect to the F group. Nevertheless, an improvement can be seen for the F cohort as well. In the case of classifiers belonging to the ensemble family (AdaBoost, RobustBoost), the *predictorImportance.m* (MATLAB 2020b, The MathWorks, Inc., MS, USA) function was applied in order to evaluate the relevance of acoustical features in separating data into two classes, and to investigate whether these metrics were significantly different. In the F cohort, the most relevant parameter for /a/ was % voiced (i.e., the percentage of voiced parts over the total audio duration). Table 3 shows that such a metric is also statistically significant, although the lowest p-

value does not characterize it. Similarly, in the /aiuole/ classification experiment, the most important predictor was T0(F0 min), which also appears in Table 4. It is interesting to notice that parameters associated with vocal instability or noise (e.g., jitter, Normalized Noise Energy) were neither relevant for models nor for statistical analysis. A similar result was found in [12] when assessing the severity of spasmodic dysphonia probably because no control group is included and comparisons are made only between pathological classes. Therefore, it seems that such parameters should not be used to differentiate these classes.

In the M cohort, the most important parameter for /a/ was F2std (i.e., the standard deviation of the second formant), which interestingly appears as the most significant feature as well. In the /aiuole/ experiment, the most relevant parameter was F0std (standard deviation of F0), and Table 6 confirms it as a statistically significant measure. Furthermore, when both feature sets are used, the most relevant parameter matches the most significant one in Table 6, i.e., T0(F0 min). It is also important to notice that, in this group, jitter seems to be statistically different between the two pathological classes in both vocal tasks. However, since the M dataset is smaller than the F one, such result should be validated on a larger dataset.

Finally, it is noteworthy that T0(F0 min) is significant for /aiuole/ in both groups (Table 4 and 6). Moreover, it presents a distribution with higher variance for UVFP with respect to BL and the same direction of effect regardless of gender: a lower value characterizes UVFP with respect to BL (Fig. 1).

In future works, a larger dataset will be considered, and the proposed approach will consider separately nodules, polyps, and cysts, comparing pre- and post-treatment vocal quality, as well as other voice disorders.

## V. CONCLUSION

Machine learning applied to voice pathology assessment is a powerful tool that may support clinical diagnosis. In this paper, binary classifiers showed high accuracy up to 88% in distinguishing patients with BL and UVFP when merging feature sets extracted from /a/ and /aiuole/ vocal tasks. Moreover, statistical analysis showed significant differences between the two considered categories of pathologies for both utterances.

## REFERENCES

- [1] J. A. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, "On the design of Automatic Voice Condition Analysis Systems. part I: Review of concepts and an insight to the state of the art," *Biomed Signal Process Control*, vol. 51, pp. 181–199, 2019.
- [2] J. Mekyska et al., "Robust and complex approach of pathological speech signal analysis," *Neurocomputing*, vol. 167, pp. 94–111, 2015.
- [3] P. Harar et al., "Towards robust voice pathology detection," *Neural Comput Appl*, vol. 32, no. 20, pp. 15747–15757, 2018.
- [4] Z. Ali, G. Muhammad, and M. F. Alhamid, "An automatic health monitoring system for patients suffering from voice complications in Smart Cities," *IEEE Access*, vol. 5, pp. 3900–3908, 2017.
- [5] I. Hidalgo-De la Guía, E. Garayzábal-Heinze, and P. Gómez-Vilda, "Voice characteristics in Smith–Magenis Syndrome: An acoustic study of laryngeal biomechanics," *Languages*, vol. 5, no. 3, p. 31, 2020.
- [6] J. I. Godino-Llorente, R. Fraile, N. Sáenz-Lechón, V. Osma-Ruiz, and P. Gómez-Vilda, "Automatic detection of voice impairments from text-dependent running speech," *Biomed Signal Process Control*, vol. 4, no. 3, pp. 176–182, 2009.
- [7] Z. Ali et al., "Voice pathology detection based on the modified voice contour and SVM," *BICA*, vol. 15, pp. 10–18, 2016.
- [8] L. Frassinetti et al., "Analysis of vocal patterns as a diagnostic tool in patients with genetic syndromes," in: *Proceedings 12th International Workshop Models and Analysis of Vocal Emissions for Biomedical Applications*, 14-16 December, 2021, Firenze, Italy, pp. 83-86 Firenze University Press Ed.
- [9] M. S. Morelli, S. Orlandi, and C. Manfredi, "BioVoice: A multipurpose tool for voice analysis," *Biomed Signal Processing Control*, vol. 64, p. 102302, 2021.
- [10] T. Hastie, J. Friedman, and R. Tibshirani, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2017.
- [11] L., Frassinetti, C., Manfredi, B., Olmi, and A. Lanatà, "A Generalized Linear Model for an ECG-based Neonatal Seizure Detector," in: *Proceedings 43rd Annual International Conference of the IEEE EMBC November, 2021*, pp. 471-474, IEEE.
- [12] F. Calà et al., "Machine learning assessment of spasmodic dysphonia based on acoustical and perceptual parameters," *Bioengineering*, vol. 10, no. 4, p. 426, 2023.

[1] J. A. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, "On the design of Automatic Voice Condition Analysis Systems. part I: Review of

# PERFORMANCE EVALUATION OF 3D NEURAL NETWORKS APPLIED TO HIGH-SPEED VIDEOS FOR GLOTTIS SEGMENTATION IN DIFFICULT CASES

A. A. Dadras, P. Aichinger

Medical University of Vienna, Department of Otorhinolaryngology, Division of Phoniatics-Logopedics,  
Speech and Hearing Science Lab, Vienna, Austria

[armin.dadras@meduniwien.ac.at](mailto:armin.dadras@meduniwien.ac.at), [philipp.aichinger@meduniwien.ac.at](mailto:philipp.aichinger@meduniwien.ac.at)

**Abstract:** This paper reports the performance of 2D and 3D deep learning models for glottis segmentation in high-speed videoendoscopy (HSV). Using a public dataset for training and the BAGLS and in-house datasets for evaluation, we assess the model's capabilities. Both models exhibit satisfactory accuracy on BAGLS, achieving intersection over union (IoU) scores of 0.81 and 0.84, respectively. However, the 3D model outperforms the 2D model on the in-house dataset with an IoU score of 0.76 compared to 0.63. By augmenting the data with various perturbations, we observe different behaviors for the two architectures. The 3D model excels in handling rotations and grid dropout, while the 2D model performs better in handling textural changes. We discuss the models' generalization capabilities, suggesting that the 3D model's rotation invariance contributes to its superior performance on unseen data. We emphasize the importance of comprehensive evaluations to uncover model behavior in challenging scenarios and identify potential issues in real-world applications.

**Keywords:** High Speed Videoendoscopy, Deep Learning, Computer Vision, Performance Characterization, Segmentation

## I. INTRODUCTION

Data-driven analysis of High-speed videoendoscopy (HSV) is opening new fields of research regarding phonatory mechanisms and diagnostic measures. The resulting videos offer with framerates of up to 5000 frames per second precise recordings of glottal movements during speech. To analyze this large amount of information further automated processing needs to be comprehensible for a human reader. Existing methods rely on the delineation of the glottis and subsequent surface measurement for every frame. These numeric features, e.g. the GAW or PVG have been exploited successfully in the past for numerous classification tasks [1].

To obtain reliable results it is crucial to ensure that the initial segmentation of the glottis is sufficiently accurate. To segment the images semi-automatic or automatic techniques such as region growing [2] or

active contour models [3] have been applied in the past. They sometimes still rely on human input and show limited performance. In recent years so-called deep learning techniques have proven to be superior when large amounts of data are available. These perform quantitatively well but sometimes fail completely on difficult cases. We try to advance this line of research by evaluating the performance of these automatic methods with and without the temporal component on challenging cases.

## II. METHODS

### A. Data

For training we used the public dataset from Trier University [4]. It consists of 130 unique Videos with 100 images per Video. We used the same train/test/validation split of 100/15/15 as in the original paper.

For evaluation we use a subset of the BAGLS dataset which contains manually segmented videos [5] and an inhouse dataset. The BAGLS dataset contains 59.250 single annotated frames and 19.200 annotated consecutive frames.

We created our inhouse dataset using 150 videos that were recorded at the medical university Vienna [6] using an HRES Endocam 5562 from Richard Wolf with a resolution of 256x256, RGB and 4000 fps. Each video has 8192 frames and is presegmented using the GAT software [7]. The segmentations are controlled and corrected by a human reader if necessary. For evaluation we randomly sample 4050 sets of 12 consecutive frames.

To make use of the temporal dimension of the data we define a timeframe within the video and treat it like a 3D object as input for the network. We choose a frame size of 12 as suggested in [5] because it approximately captures one cycle of the glottis.

### B. Architecture

For our experiments, we trained a 2D and a 3D variant of the commonly used U-net architecture [8]. As preprocessing the images are converted into grayscale,

scaled to 256x256 pixels, and normalized between 0 and 1.

We follow the results from [9] where it has been shown that a reduced parameter size is better suited for glottis segmentation. Thereby we reduced the size of the channels, especially in the bottleneck (see Table 1). Our 2D Unet has 1.4M parameters, whereas the standard Unet has 7.7M parameters. The 3D model we used had 3M parameters. In comparison the original 3D Unet [10] has 19M parameters.

Table 1: Differences in training setup between architectures

3D Unet	Channels: [16, 32, 64, 128, 32] Batch Size: 8 Gradient Clipping: yes (1)
2D Unet	Channels: [32, 64, 128, 256, 64] Batch Size: 32 Gradient Clipping: no

### C. Training

As an activation function we use a sigmoid for the 2D Unet. We use the arctangent and gradient clipping to the value of 1 for the 3D Unet to avoid local minima. As initialization technique ‘Kaiming’ is applied.

We use the DICE score as loss [11] (see E.) and train our networks on a A100 GPU for 100 Epochs. Our code is implemented using the Pytorch library.

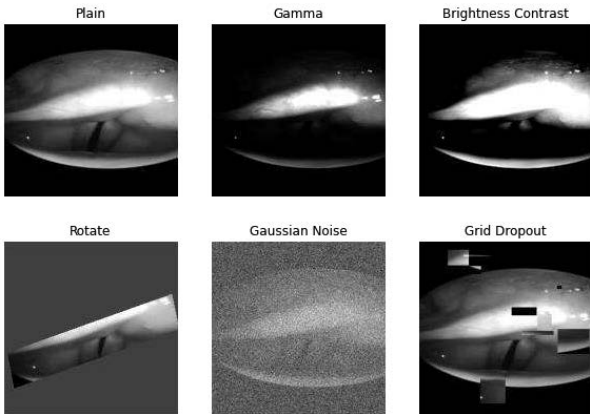


Fig. 1: Data augmentations on BAGLS with amp=2.5

For training we make use of five light data augmentation techniques with the standard values of the Volumentation library [12] (see Table 2). They are applied randomly with probability p during training.

Table 2: Augmentation parameters, p only applies during training, amp = 1 during training

Augmentation	Values	p
Rotation	x_limit = (-10, 10 * amp), y_limit = (-5, 5 * amp), z_limit = (-5, 5 * amp)	.75

Brightness/Contrast	brightness_limit=0.2 * amp contrast_limit=0.2 * amp	.75
Gamma	gamma_limit=(int(80 / amp), 120 * amp)	.75
Gaussian Noise	var_limit=(0.10 * amp, 0.50 * amp)	.5
Grid Dropout	blocks=int(amp*8)	0

### D. Evaluation

During evaluation we test data augmentations separately with p=1 and increasingly amplify them by a factor **amp**. During training amp is equal to 1. In addition, we implement our own version of grid dropout for evaluation, where cutouts of random size in the range of 9x50x50 pixel are placed randomly within the 3D object. The amp factor changes the number of blocks.

We run the segmentation with the 3D net once for every frame, such that we only evaluate the frame in the middle of the window (i.e., position 6 with a frame size of 12) and compare the performance to segmenting the whole window once. To test the usage of the temporal component we also replace single frames of the window randomly with noise.

### E. Error measures

For evaluation we use intersection over union (IoU) instead of DICE, as it is widespread practice in related papers with the given segmentation task. These metrics are based on the ratio of True/False (T/F) Positives/Negatives (P/N):

$$IoU = \frac{TP}{TP+FP+FN} \quad (1)$$

$$DICE = \frac{2TP}{2TP+FP+FN} \quad (2)$$

## III. RESULTS

We evaluated our 2D and 3D models using BAGLS and our inhouse experiment on different data augmentations with increasing amp. Moreover, we tested the 3D capabilities of our 3D network by replacing single frames with noise and comparing segmenting the videos in chunks vs. one frame at a time with a sliding window. Our experiments show no significant difference (less than 0.01 IoU) between segmenting single frames with the 3D net or the whole window. This means computational cost of segmenting can be reduced by a factor of the window size (here 12) without loss of accuracy in comparison to the sliding window approach. The replacement of random frames with noise also did not hurt performance by more than 0.01 IoU. Left out

frames were segmented correctly using neighboring frames.

Table 3: Evaluation IoU±variance among frames without augmentations

Dataset/ Architecture	Trier Validation	BAGLS	Inhouse
3D Unet	0.84±0.014	0.80±0.02	0.76±0.035
2D Unet	0.81±0.021	0.79±0.018	0.63±0.021

Both networks performed well on the validation set from trier. The 3D net reached an IoU of 0.84 while the 2D net also showed a satisfactory performance with 0.81. The difference vanishes on the BAGLS data. Despite using data from different sources for training both perform well (0.8/0.79 IoU). In the case of the inhouse dataset the 3D still performs with an IoU of 0.76, whereas the 2D net drops to an IoU of 0.63. According to [13] performances below 0.74 are considered unreliable for further processing.

Table 4: Mean Evaluation IoU±variance among frames with amplified augmentations

Dataset/ Architecture	BAGLS	Inhouse
3D Unet	0.75±0.024	0.72±0.037
2D Unet	0.72±0.02	0.60±0.019

The data augmentations result in different behavior for both architectures. While the Unet shows more stability in gamma correction, brightness/contrast changes and gaussian noise, the 3D Unet was clearly superior when rotations and grid dropout were applied (see Fig. 2). Small irregularities, where the performance increases non-monotonically in more difficult settings, are due to the randomized augmentation parameters.

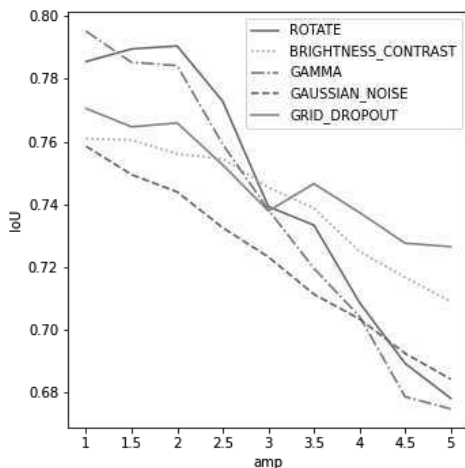


Fig. 2 Evaluation results of 3D net on inhouse set

Depending on the used evaluation set, the order in which augmentations decrease accuracy changes slightly. Grid dropout causes significantly more difficulties with the

inhouse dataset (see Fig. 2) for the 3D net. Since the accuracy already differs strongly without augmentations a comparison with the 2D net on the inhouse set is not feasible.

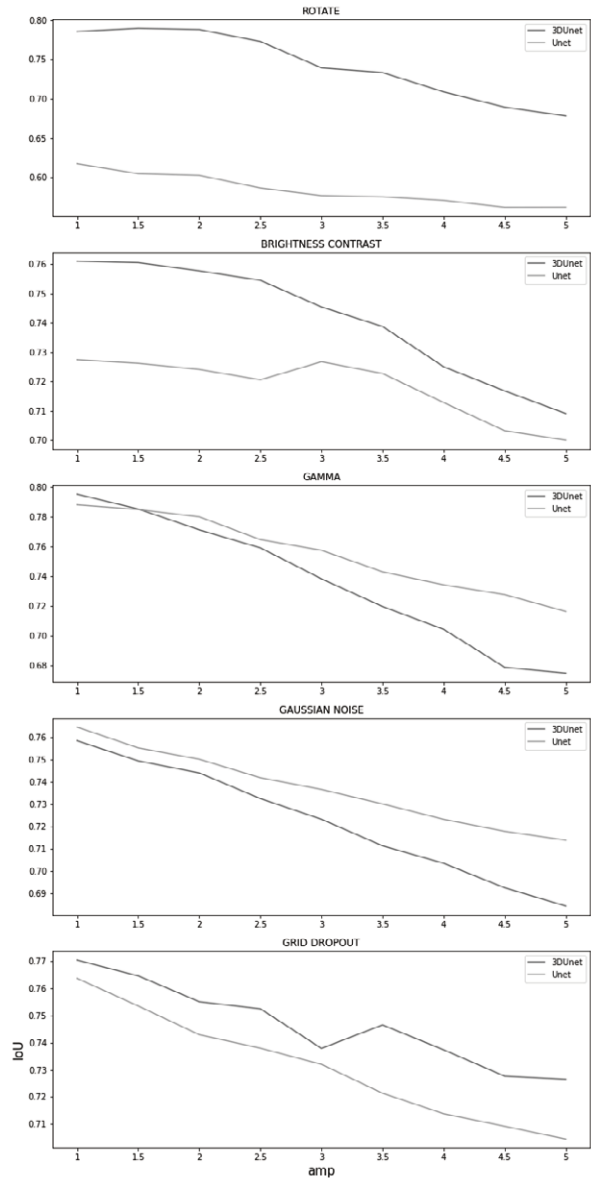


Fig. 3 Evaluation results using data augmentations

#### IV. DISCUSSION

Our results show that 3D features may be more stable among different datasets. We hypothesize that the rotation invariance of the 2D neural network is inferior to the 3D variant since the augmentations are also performed in 3D space. This could also mean that the 2D network is more vulnerable to changes due to camera positions and angles. However textural changes, simulated using noise, brightness contrast and gamma, are easier for the 2D features to handle.

The increased severity of grid dropout may be due to the view of our inhouse set. While we decided for the native camera view, the BAGLS set is cropped to a region of interest. Therefore, random crops can include a wider variety of visual appearances in the inhouse set.

Our networks reached their maximum accuracy after 25 epochs for the 3D model and 34 epochs for the 2D model. Since we wanted to keep the training setting similar, we did not do more exhaustive hyper parameter optimization. More advanced techniques like self-supervised learning or attention modules for example may furthermore affect the performance of the networks differently.

#### V. CONCLUSION

We investigated the effect of different perturbations of the input data on 3D and 2D networks. Our result shows the utility of 3D features on unseen data and highlights the importance of more exhaustive evaluation tasks. The performance may not differ too strongly on the validation set and BAGLS, but our 3D model generalizes better on unseen data than the 2D model. Stronger data augmentations shed light on the behavior in challenging cases which may help to identify unforeseen problems in real settings

#### REFERENCES

- [1] Aichinger, P., Roesner, I., Schneider-Stickler, B., Bigenzahn, W., Feichter, F., Fuchs, A. K., ... & Kubin, G. (2013, December). Spectral analysis of laryngeal high-speed videos: case studies on diplophonic and euphonic phonation. In *Proceedings of the 8th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications* (pp. 81-84).
- [2] Demeyer, J., Dubuisson, T., Gosselin, B., & Remacle, M. (2009, May). Glottis segmentation with a high-speed glottography: a fully automatic method. In *3rd Adv. Voice Funct. Assess. Int. Workshop*
- [3] Karakozoglou, S. Z., Henrich, N., d'Alessandro, C., & Stylianou, Y. (2012). Automatic glottal segmentation using local-based active contours and application to glottovibrography. *Speech Communication*, 54(5), 641-654.
- [4] Fehling, M. K., Grosch, F., Schuster, M. E., Schick, B., & Lohscheller, J. (2020). Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep Convolutional LSTM Network. *Plos one*, 15(2), e0227791.
- [5] Kruse, E., Döllinger, M., Schützenberger, A., & Kist, A. M. (2023). GlottisNetV2: Temporal Glottal Midline Detection using Deep Convolutional Neural Networks. *IEEE Journal of Translational Engineering in Health and Medicine*.
- [6] Aichinger, P., Roesner, I., Leonhard, M., Denk-Linnert, D. M., Bigenzahn, W., & Schneider-Stickler, B. (2016, May). A database of laryngeal high-speed videos with simultaneous high-quality audio recordings of pathological and non-pathological voices. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 767-770).
- [7] Kist, A. M., Gómez, P., Dubrovskiy, D., Schlegel, P., Kunduk, M., Echternach, M., ... & Döllinger, M. (2021). A deep learning enhanced novel software tool for laryngeal dynamics analysis. *Journal of Speech, Language, and Hearing Research*, 64(6), 1889-1903.
- [8] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18* (pp. 234-241). Springer International Publishing.
- [9] Kist, A. M., Breininger, K., Dörrich, M., Dürr, S., Schützenberger, A., & Semmler, M. (2022). A single latent channel is sufficient for biomedical glottis segmentation. *Scientific Reports*, 12(1), 14292.
- [10] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19* (pp. 424-432). Springer International Publishing.
- [11] Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3), 297-302.
- [12] Solovyev, R., Kalinin, A. A., & Gabruseva, T. (2022). 3D convolutional neural networks for stalled brain capillary detection. *Computers in biology and medicine*, 141, 105089.
- [13] Groh, R., Dürr, S., Schützenberger, A., Semmler, M., & Kist, A. M. (2022). Long-term performance assessment of fully automatic biomedical glottis segmentation at the point of care. *Plos one*, 17(9), e0266989.



# HERMESPEECH RECORDER: A NEW OPEN-SOURCE WEB PLATFORM TO RECORD SPEECH TO THE CLOUD

Jang-Woo Park<sup>1</sup>, Maximilian Zinkus<sup>1</sup>, Jim Huang<sup>1</sup>, Ankur Butala<sup>2</sup>, Jayne Zhang<sup>2</sup>, Lora Clawson<sup>2</sup>, Sarah Cust<sup>3</sup>, Victoria Chovaz<sup>4</sup>, Najim Dehak<sup>5</sup>, Helin Wang<sup>5</sup>, Laureano Moro-Velazquez<sup>5</sup>

<sup>1</sup>Department of Computer Science, Johns Hopkins University, US

<sup>2</sup>Department of Neurology, Johns Hopkins University School of Medicine, US

<sup>3</sup>Department of Physical Medicine and Rehabilitation, Johns Hopkins University School of Medicine, US

<sup>4</sup>School of Nursing, Johns Hopkins University, US

<sup>5</sup>Center for Language and Speech Processing, Johns Hopkins University, US

{jpark278, mzinkus, jhuan185, ndehak3, hwang258, laureano}@jhu.edu,

{ankur.butala, jz, lclawson, scust1, vchovaz1}@jhmi.edu

**Abstract:** HermeSpeech Recorder is a user-friendly and open-source platform designed to record speech from a large cohort of participants. The platform allows users to easily initiate, pause, and stop recordings of their assigned scripts or speech tasks. The completed speech recordings are encrypted and stored in a cloud service, complying with HIPAA regulations for data privacy, if needed. The resulting files are uploaded asynchronously to enable recordings in areas with low connectivity. The platform also provides features for administrators to easily create new participants, load and manage many scripts at once and assign tasks to participants, simplifying the management of large cohorts, and making it a streamlined solution for remote speech recording in a HIPAA compliant and efficient manner. Finally, we describe our experience recording a group of participants with atypical speech using this open-source tool.

**Keywords:** speech recording, spoken language understanding, open-source

## I. INTRODUCTION

Speech recording platforms are fundamental tools for collecting voice samples that can be used in speech and language technologies, including speech and speaker recognition, spoken language understanding (SLU), or text-to-speech. These become essential when we want to collect data that is difficult to crawl from publicly available resources, such as pathological and atypical speech. Existing offline speech recording platforms, while helpful, require in-person recording [1], whereas platforms that allow for remote recordings can be more flexible and accessible, as they can be used in-person to record under controlled conditions, and at home. Some current web platforms [2] are not compliant with the requirements of the Health Insurance Portability and Accountability Act (HIPAA), are not adapted to provide intent annotations, or are not very flexible when it comes to predefining a large list of participants and

assigned speech tasks, which might be required for clinical studies.

To address these challenges, we present HermeSpeech Recorder,<sup>1</sup> a new open-source speech recording web platform designed to record speech remotely. HermeSpeech provides the participant (speaker) with a certain speech task, i.e., text to read aloud, and associates the resulting recordings with the read text and intent annotations, if any. Our platform leverages encrypted cloud storage and a secure transport design while offering a user-friendly interface that allows for easy setup and management of recording sessions. In contrast to real-time audio streaming platforms, the recordings are generated locally within the speaker's browser and then uploaded to the cloud to avoid network bandwidth limitations that could affect streaming audio quality.

In this paper, we describe the platform, which has the potential to expand the pool of available speech data for research and development of speech technologies. Our ultimate goal is to facilitate access to a speech recording tool for underresourced scenarios, such as atypical and pathological speech applications, enabling more inclusive spoken language technologies. Then, we summarize our experience using the platform with a group of participants with atypical speech.

## II. TOOL DESCRIPTION

When HermeSpeech is deployed to a web server, it provides participants with a *login screen* to access a *task dashboard* using a unique token previously assigned (Fig. 1). Once users gain access, they can record the assigned set of text scripts. The platform also allows recording unscripted speech or monologues

<sup>1</sup><https://github.com/Neuro-Logical/HermeSpeechRecorder>



(elicited tasks). Finally, those recordings and associated metadata are uploaded to a cloud service.

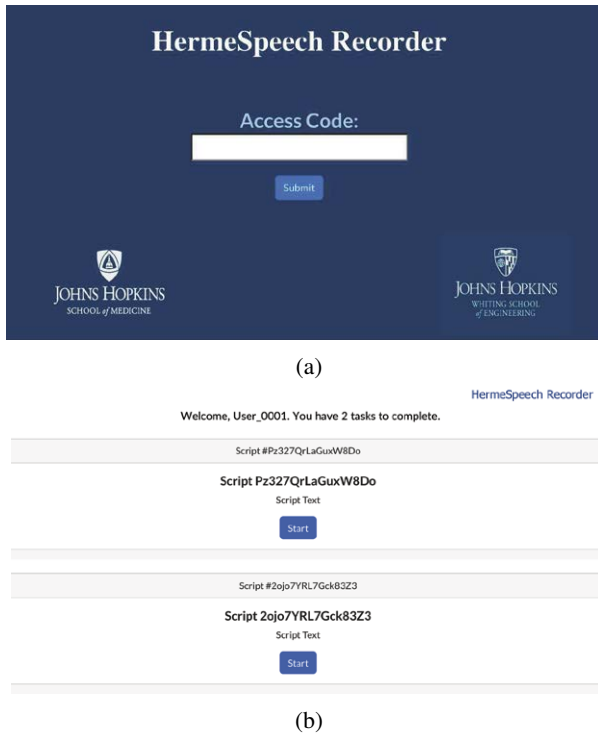


Fig. 1: Login interface of HermeSpeech Recorder (a) and script dashboard listing two scripts to be completed by the participant (b). Background images can be easily changed.

### A. User Interface

The user interface of the platform is designed to be user-friendly and efficient, using React<sup>2</sup> and various open-source libraries for frontend components. Once logged in, participants are presented with a *task dashboard* that provides an overview of their assigned and completed tasks, which are the text scripts to record. The dashboard provides a clear and organized view of the tasks, allowing participants to track their progress easily. Participants are assigned scripts to record using the *admin management* console detailed in Section II-C.

### B. Script Recording

Within each task or script, the user interface offers a *recording module* that allows participants to control the recording process. This *recording module* includes intuitive controls, such as buttons for starting, stopping, and repeating the recording. The interface also provides options for participants to review and listen to

<sup>2</sup><https://react.dev/>

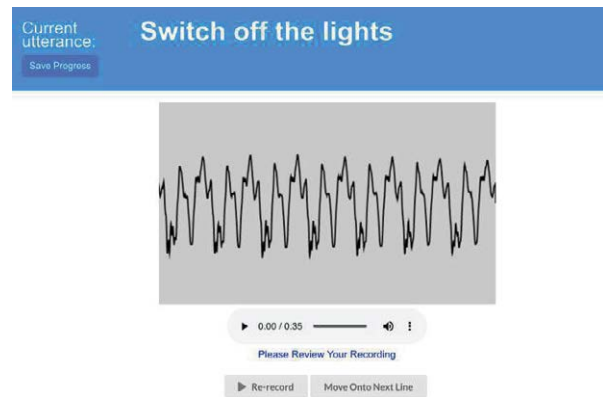


Fig. 2: Recording module interface. The upper part of the screen contains the text to be read (Switch off the lights, in this case), the central includes the speech waveform to indicate the participant that there is signal and it is being recorded, and the bottom part allows the participant to listen to the last recording, re-record it if something went wrong, or move on to the next utterance.

their recordings before submission. This review feature allows participants to playback their recordings to ensure they contain the required information. This helps participants to verify the quality and word errors of their recorded speech segments and re-record sentences when needed. The different controls of the *recording module* are included in Fig. 2.

To ensure that the speech segments are not cut at the beginning or end, the platform provides pre- and post-buffer features. These features allow platform administrators to set a predetermined amount of time that is recorded before the sentence to be read or speech task appears on screen and after clicking the stop button, ensuring that the full utterance is captured. This means that when participants initiate the recording, the text to be read is slightly deferred (pre-buffer). Similarly, when participants stop the recording, the platform continues to record for a specified duration (post-buffer). This avoids that the participant starts speaking before the platform is recording, or stops recording in the last word of the utterance, which are both common errors in some participants.

After a text script is recorded, the platform uploads the recordings and a descriptor file to a cloud service. The descriptor file contains the name of the recorded files, the participant who recorded them, the text that was read in each file, intents annotations, if the script had them associated, and other extra fields, depending on the project's needs.

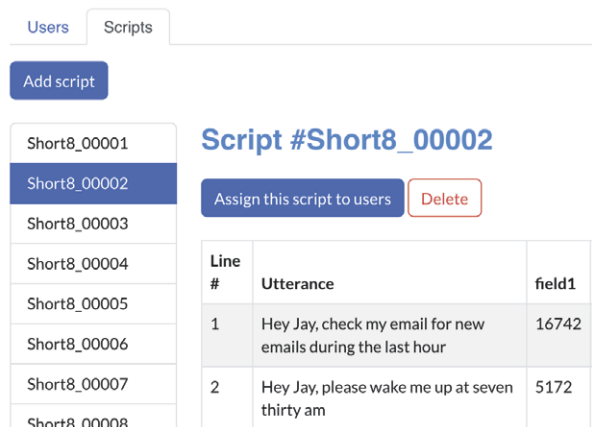


Fig. 3: Admin management interface including a list of possible scripts and their text. The interface allows loading new scripts, assigning them to users, and modifying them.

### C. Admin management

The admin management side of the platform is also designed with intuitive and streamlined interfaces, providing means for managing and creating scripts, assigning scripts to participants, and creating participant profiles (Fig. 3).

One key feature of the admin console is the ability to easily populate the script database by uploading a CSV file with script information, such as text to be read and the intent of the text for SLU-related projects. The admin console automatically processes the data in the CSV file, which must have a specific format, creating scripts based on the provided information. Similarly, administrators can create participant profiles by uploading a CSV file, simplifying the process of adding multiple participants to the platform at once. This data management feature saves time and effort for administrators, ensuring that scripts and participants are quickly and accurately added to the system.

### III. DATA STORAGE AND MANAGEMENT

The platform utilizes MySQL and the Sequelize Object Relational Mapper (ORM) as the foundation for the database infrastructure. In the initial version of HermeSpeech Recorder, we include the instructions to securely encrypt and save the recordings into Microsoft Azure, a cloud computing platform that is HIPAA compliant, for health data privacy. Adaptation to other cloud services such as AWS will be implemented in future releases. The use of MySQL and Sequelize ORM in conjunction with Microsoft Azure allows for efficient data management, retrieval, and analysis. Sequelize

ORM simplifies the interaction between the platform and the MySQL database.

HermeSpeech uses Internet-standard Transport Layer Security (TLS) for all web-based interactions. This protects participant’s Protected Health Information (PHI) between their web browser and our servers. Data is not persisted within the browser, mitigating PHI risk on the participant’s side. Within our servers, data is stored encrypted at rest, and is transported to backing long-term storage using an API-authenticated TLS communication, after which it is also stored encrypted at rest. Credentials for API communication are managed at the Azure system level, protecting them from compromise in the course of regular application use. Administrators are individually identified by their login credentials. Further, all involved systems are configured according to cybersecurity hardening best practices. We have engineered the HermeSpeech platform to comply with all relevant tenets of HIPAA regarding PHI data protection, portability, and accountability.

### IV. RECORDING WITH HERMESPEECH

We are using HermeSpeech in the collection of a new corpus of atypical speech. All the speakers are being recorded at the Johns Hopkins Medical Institutions, where the Institutional Review Board approved the data collection. All participants signed informed consent. The goal of the corpus is to provide the scientific community and developers with a new corpus of atypical speech including annotated transcriptions and intent. SLU and Automatic speech recognition (ASR) systems are not always trained to have a good performance with atypical speech, because this type of data is not commonly available. The collection of new corpora including atypical speech and intent labels will facilitate to train and adapt new SLU systems, which can be crucial for the interaction between atypical speakers and speech assistants. Table I includes the demographics, word error rate (WER) and character error rate (CER) of the first 18 atypical speakers recorded with HermeSpeech. We automatically transcribed all the utterances with Whisper<sup>3</sup>. The average number of utterances recorded per speaker is 239. The speakers read sentences that overlap on intent with the Fluent Speech Commands (FSC) dataset [4], which will allow us to perform comparative analyses and use models pre-trained with FSC for ASR and SLU. In our case, all the sentences started with the wakeword "Hey Jay", in contrast to FSC where there is no wakeword. The table also includes the average WER and CER of the FSC corpus for comparison, also transcribed with Whisper.

<sup>3</sup><https://github.com/openai/whisper>

TABLE I: ASR results on the collected data and Fluent Speech Command dataset with Whisper [3]. Here, we report word error rates (%) and character error rates (%).

Dataset	Speaker	Diagnosis	Number of audios	Sex	Age	WER	CER	
Atypical speech	0005	Episodic Ataxia	256	Male	65	28.7	17.3	
	0006	Episodic Ataxia	228	Male	39	1.75	1.13	
	0007	Idiopathic Cervical Dystonia	281	Male	40	0.36	0.14	
	0008	Acute Ischemic Stroke	208	Female	80	48.86	36.70	
	0009	Cerebral and Focal Dystonia	554	Male	41	37.18	23.45	
	0010	Wernicke-Korsakoff syndrome	168	Female	72	36.91	23.07	
	0011	Episodic Ataxia	160	Male	75	22.85	14.22	
	0012	Parkinson's Disease	208	Male	53	1.29	0.75	
	0013	Parkinson's Disease	208	Female	80	27.09	16.10	
	0014	Sensorineural Hearing Loss	208	Male	80	0.50	0.43	
	0015	Parkinson's Disease	150	Male	60	42.86	25.15	
	0017	Control Subject	285	Female	75	23.91	13.92	
	0018	Multiple System Atrophy	206	Female	65	15.96	10.13	
	0019	Spinocerebellar ataxia	285	Male	50	61.00	42.82	
	0020	Spinocerebellar ataxia	181	Female	56	4.30	2.03	
	0021	Amyotrophic Lateral Sclerosis	289	Male	75	6.62	3.74	
	0022	Parkinson's Disease	317	Male	68	4.06	1.76	
	0023	Stiff Person Syndrome	218	Female	58	35.88	22.18	
		all	-	4306	-	-	23.14	14.80
	FSC	-	-	-	-	-	6.19	3.70

The comparison between FSC and our recordings evidence the differences of WER and CER between atypical speech and the FSC speakers. The recorded speakers, with a variety of etiologies and speech disfluency severity ranging from mild to severe, have an absolute 17% higher WER than the FSC speakers, even when the latter contain native and non-native English speakers [4].

During the collection of this first part of the cohort, we observed that some participants said the wakeword right before clicking on "Start recording" if the text to be read was already present on the screen. In other cases some participants tended to click on the button to continue to the next line before finishing the current utterance. This led to incomplete recordings. Therefore, we empirically set up a pre-buffer time of 1 s and a post-buffer time of 2 s. This means that the sentence was not shown on screen until 1 s after the platform started recording, and the recording did not stop until 2 s after the participants clicked on "Move Onto Next Line".

Although most atypical speakers can get tired of speaking sooner, in comparison to typical speakers, we observed that recording more than 230 utterances in a single session of less than 1 h was feasible and did not cause any significant fatigue to most of them.

This ongoing corpus will be made available after completion, and it will include speech recordings, transcriptions, intent annotations, type of speech disfluency, severity, and other annotations and rating scales made by speech pathologists and neurologists.

## V. CONCLUSIONS

In conclusion, HermeSpeech Recorder is an open-source comprehensive web solution that facilitates the recording of speech segments from participants remotely. The platform provides means of capturing and managing speech recordings for research and development purposes.

We aim to provide this platform to researchers and institutions seeking a streamlined and efficient solution for medium- and large-scale semi-automated speech recording practices, while we are committed to continuously improving and customizing the platform. We plan to expand the platform by adding more cloud and data storage options, as well as incorporating additional buffer and blob storage architecture parameters and features. We are dedicated to ongoing development and enhancement of the platform to ensure it remains a valuable open-source option for researchers and institutions alike.

## REFERENCES

- [1] P. Boersma and V. Van Heuven, "Speak and unspeak with praat," *Glott International*, vol. 5, no. 9/10, pp. 341–347, 2001.
- [2] R. e. a. Ardila, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.
- [3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.
- [4] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," *Proc. Interspeech 2019*, pp. 814–818, 2019.

# OPERATIC SINGING MULTI-SENSOR RECORDING PROTOTYPE: A PILOT EVALUATION STUDY

E. Angelakis<sup>1</sup>, K. Bakogiannis<sup>1</sup>, A. Andreopoulou<sup>1</sup>, M. Habela<sup>2</sup>, A. Georgaki<sup>1</sup>

<sup>1</sup> National and Kapodistrian University/Department of Music Studies, Laboratory of Music Acoustics and Technology, Athens, Greece

<sup>2</sup> Haute Ecole Spécialisée de Suisse Occidentale/Haute Ecole de Musique de Genève, Geneva, Switzerland  
angelakisv@music.uoa.gr, kbakogiannis@music.uoa.gr, aandreo@music.uoa.gr, marcin.habela@hesge.ch,  
georgaki@music.uoa.gr

**Abstract:** The Operatic singing music genre is characterized by the utilization of the vocal mechanism in some of its most demanding and complex kinetic manifestations. The study presented here serves as a) an introduction to a newly designed multi-sensor recording prototype for operatic singing, b) an account of its use in a research project, as well as c) a preliminary test for the prototype's evaluation. The proposed prototype employs sensors that record acoustic and electroglottographic data, breathing kinetic actions, and data regarding pertinent postural and body movement behavior. It was recently utilized for the recording of an experiment with 28 operatic singers. Captured data for three of the participants was used for the present pilot study, and short videos of these three singers were given to an expert vocal trainer for 'vocal technique problems' empiric evaluation via a questionnaire. A subsequent examination of the recorded multi-sensor data resulted to the successful detection of objective evidence to support the 'vocal technique problems' reported by the expert.

**Keywords:** operatic singing, sensors, EGG, posture, breathing

## I. INTRODUCTION

Operatic singing is an art form that spans more than 400 years. It is an empiric trait delivered from singer to singer through the ages, one that demands great precision, control, and artistry. Although scientific research on the subject has progressed much during the last few decades, and numerous studies have been conducted [1], there is still much to be uncovered regarding the details of the mystifying functions of the singing voice, and operatic singing in particular. Singing, in any genre, is essentially the result of neuromuscular processes which involve voluntary and involuntary control of both external and internal mechanisms of the human body [2].

It thus stands to reason that an examination and tracing of such biomechanical functions would be essential in order to delve deeper into the 'secrets' of this extremely rigorous art. Voice research "concern

with functionality is increasing" since 2010 [1], and multi-sensor research [3,4] and software/hardware applications for the singing voice -such as VoceVista Video Pro (Sygyt Software, Bochum, Germany)- are expanding sectors. Practical implementations of the above, concerning the art form of singing, can be found in vocal pedagogy [5], but have also been used for the recording of rare vocal music genres [6].

This paper discusses the design of an operatic voice multi-sensor recording prototype and its pilot evaluation study. The prototype features the possibility to record with 6 sensors (using commercial software), a control interface programmed in Max/MSP for reading and recording data from the skeletal tracking camera, and MATLAB code for synchronizing and manipulating all the recorded data. The prototype comprises the following sensors:

- ✓ Condenser microphone – Behringer ECM8000
- ✓ Electroglottograph (EGG) - Glottal Enterprises EG2-PCX2 (recording both vocal fold degree of contact and Vertical Laryngeal Position -VLP)
- ✓ Two distinct Respiratory Effort Transducers – Biopac SS5LB
- ✓ Time-of-Flight (ToF) Skeletal tracking camera – Azure Kinect Microsoft
- ✓ HD video – iPhone 13 Pro.

The sensors were selected taking into consideration the requirements for high portability, low invasiveness, and representation of a relatively large number of pertinent-to-singing kinetic functions. The innovation of the proposed prototype lies in its ability to generate 12 sensor data streams, seven of which calculated by its gesture-following functionality. This enables the acquisition of synchronized, quantifiable data regarding singing-related biometrics. The present pilot study was conducted with the aim of investigating whether the combined use of these sensors yields data that is relevant for the comprehensive evaluation required for the final prototype assessment.

## II. METHODS

*Multi-sensor Recording Protocol:* The proposed prototype was used to record an experimental part of a

larger project led by the Haute Ecole de Musique de Genève. The recordings took place in three venues. A total of 28 singers were recorded, all of them graduate or post-graduate students and some young professionals.

Prior to the measurements, participants were instructed to arrive for the study vocally warmed-up and in good vocal condition. They were also asked to memorize an Italian aria of their choice, a song in their native language, and the first two phrases of the Aria Antica 'Caro mio ben'. All participants were requested to sign informed consents, as well as to complete a demographics and vocal health questionnaire.

The measurement protocol for each participant began with a calibration process during which all sensors were manually adjusted. The recording phase commenced with data synchronization events, followed by various vocal exercises, the 'Caro mio ben' phrase, the Italian aria, the song, and concluded with ending synchronization events.

*Data Collection:* Microphone (voice) and Electroglottograph (EGG and Laryngeal Tracking signals) data were recorded on 3 mono channels using a Steinberg UR44-C external sound card, and Cubase 12 at a 48 kHz sampling rate. The thoracic and abdominal breath monitoring transducers were connected onto the specialized 'Biopac MP35 Four Channel Data Acquisition System' and their data were recorded at a 25 kHz frame rate, using BSL4 Pro Software (Biopac Systems, Inc., Santa Barbara, CA). Both devices (MP35 and UR44-C) were connected to the same laptop pc through USB-C and Hi-speed USB ports respectively. Breathing transducer data streams were exported as .wav sound files.

A second laptop was used to read and record the skeletal tracking data from the Kinect Azure DK camera. A Max/MSP patch was developed to record the coordinates and orientation of 15 body 'joints' (head, right eye, left eye, right ear, left ear, nose, head centre, neck, thorax, right shoulder, left shoulder, navel, pelvis, right hip, left hip). These data were automatically exported along with their corresponding timestamps and sound level values into a .txt file with a sample rate of 60 Hz. Video recordings of all experiment trials were made in 1080p 30 fps video using an iPhone 13 Pro.

Each measurement started and was concluded with a short synchronization sequence, which consisted of three hand claps, followed by three small 'cough-like' glottal attack sounds, produced simultaneously with an abdominal muscle inward activation, and a small, sharp downward head bend. This latter event was selected as it provides information recorded by all sensors (microphone, EGG, breathing transducers, skeletal tracking) and can thus be used for data synchronization verification. A large pause of about 10 seconds was introduced between the first and second clap, during which participants were asked to stand, in complete

silence, in what they considered their personal optimal upright posture. Skeletal tracking data from this pause was used to set each user's reference posture.

*Data Processing:* The collected data were resampled to 44.1 kHz, synchronized, and clipped automatically in MATLAB. Synchronization was achieved by automatic alignment of the audio streams recorded in all data packages (DAW, BSL4, Kinect Patcher, Video). MATLAB was also used to process the skeletal tracking data and output 7 distinct data streams of specific movements of the singer's body. These movements were selected upon consultation with internationally acclaimed singing teachers as the most appropriate for this study. They were clearly visible movements that could either impact vocal production, or be good indicators of a technical, habitual, or physiological 'issue' that can impede the optimal function of the voice's kinetic mechanisms. These selected movements were: 1) body posture, 2) up-down head bend, 3) left-right head turn, 4) parallel front-back head movement, 5a) right shoulder up-down, 5b) left shoulder up-down, 6) shoulder front-back (kyphosis-backward stretch).

*Evaluation:* The preliminary evaluation of the multi-sensor prototype relied on the objective and subjective assessment of a subset of the collected data (three participants), selected using the following criteria: a) same gender, b) recorded at the same venue c) different level of expertise (advanced: singer01, novice: singer02, and young professional: singer03). The videos of these three singers were exported synced with the audio from the measurement microphone, and the following excerpts were selected for subsequent evaluation: two scale exercises, the 'Caro mio ben' phrases, and the most demanding part of their selected aria. Their approximate duration ranged between 2'20'' and 2'40''. These excerpts were used for the subjective assessment section of the preliminary evaluation, which aimed towards a two-fold objective. First and foremost, to demonstrate as to whether 'vocal technique problems' (VTPs) and their indications (as reported by an expert) had been recorded and were discernable within the research data. The second objective was to establish a methodology for a large-scale study with many expert judges. In the above scope, the data was sent to an expert assessor for subsequent evaluation along with a questionnaire and detailed instructions. The selected expert judge is an internationally acclaimed operatic singer with a 23-year singing career, 18 years of teaching experience, and a comprehensive understanding of the physiological mechanisms pertinent to vocal production. The questionnaire which they were asked to follow consisted of 9 questions regarding mainly a) the VTPs they perceived, b) the indications that led to the detection of each VTP, and c) suggestions on muscular systems each participant should work on.

### III. RESULTS

*A. Expert judge report and examination of multi-sensor data:* Questionnaire answers from the expert who judged the three participants' videos were examined and the reported VTPs are listed below, sorted by sensor data stream where they have been recorded, or where evidence for these 'problems' can be detected.

- ✓ *EKG signal:* insufficient glottal adduction, arytenoid cartilages strain, increased air pressure,
- ✓ *VLP signal:* low position of larynx, downwards pressure of larynx, lack in laryngeal control of movement,
- ✓ *Kinect spinal posture in conjunction with abdominal breathing sensor:* Insufficient body support, air control,
- ✓ *Kinect shoulder forward/backward bend/stretch in conjunction with thoracic breathing sensor:* thoracic spinal region tension,
- ✓ *Uncategorized (no pertinent recorded data):* tongue lower part stiffness.

The expert was also requested to provide indications that led to the report of VTPs. These indications are listed here and can mostly be detected in an audio spectrographic analysis, EGG signal analysis, and supported by data from the breathing sensors and postural data.

*Indications:* instability in vibrato & intonation and voice quality change, unstable intonation and vibrato, breathy sound, unstable dynamics, sound distortion, growling sound, short duration of high note, face observation, posture observation.

#### *B. Data analysis:*

In order to provide a few characteristic examples of quantifiable indications for reported 'vocal technique problems', a selection of recorded data analysis for participant singer01 is presented in this section. Reported issues for this singer that can be observed here are: "Larynx position without adequate control", "insufficient air pressure control through body support", "intense subglottal air pressure", "vibrato frequency decrease", "breathy sound", "muscular tension" in the thoracic spinal region.

*Example #1:* Evidence of a reported VTP is illustrated in Fig. 1 from an analysis of the VLP and audio data streams. It concerns a sustained note of singer01 on a G3 note (196 Hz) and the word "ben" (first phrase of 'Caro mio ben' Aria), which was indicated as problematic by the expert judge. The expert's report for lowering of the larynx is apparent on the onset of this and the following syllable, while lack of its control could be attested by the constant variation of the VLP during the sustained vowel/tone, which is connected also to the reported vibrato instability.

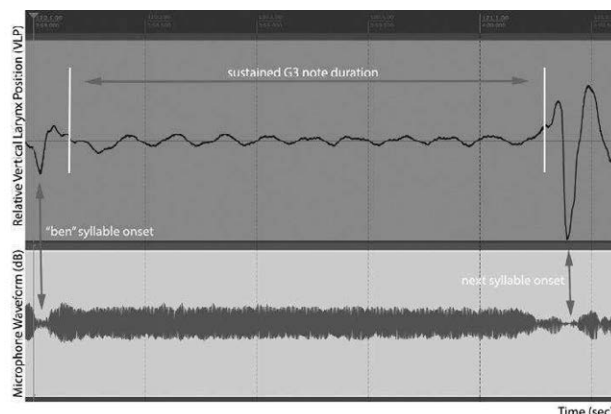


Figure 1. Example #1 depiction of VLP (top) and microphone (bottom) waveforms for reported G3 note.

*Example #2:* Vibrato rate for singer01 has been measured in VoceVista, as seen in Fig. 2, to occasionally decrease from an already relatively low 6 Hz [7] to values between 4.5-5.0 Hz in sustained notes, thus confirming the expert's report for "vibrato frequency decrease". Analysis also interestingly revealed a high vibrato extent. While vibrato extent in operatic singers has been found to range "between  $\pm 34$  and  $\pm 123$  cent" and "the mean across tones and singers amounted to  $\pm 71$  cent." [8], singer01 was measured to have an extent range of about  $\pm 49$  to  $\pm 200$  cent, with an astonishing extent rise during the note in question (among others) up to a range of  $\pm 159$  to  $\pm 282$  cent, which should greatly add to the audible vibrato fluctuation effect.

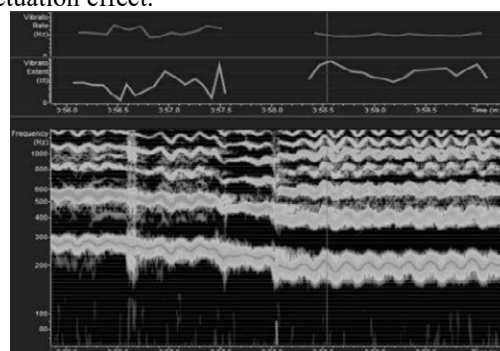


Figure 2. Example #2 vibrato rate in Hz (top), vibrato extent in cents (middle), and spectrogram (bottom) for the first 4-note except of 'Caro mio ben' aria, with the last note being the note used for Example #1.

*Example #3:* Reported muscular tension in the thoracic part of the spine cannot easily be measured with non-invasive techniques, such as the ones deployed in the present prototype. However, recorded data (as shown in Fig. 3 for the whole first phrase of 'Caro mio ben') demonstrate that singer01 tended to employ a combination of abdominal and thoracic breathing that seem to commence simultaneously but follow distinct recession slopes.



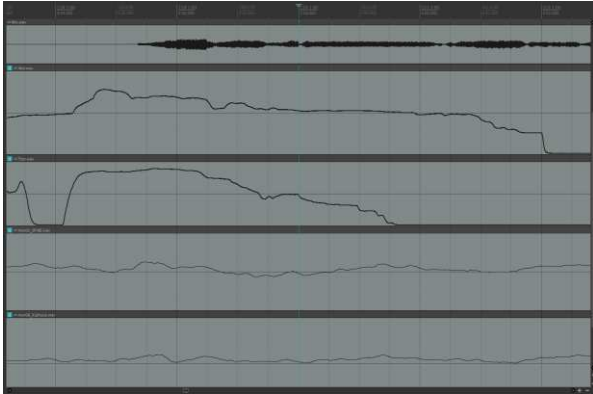


Figure 3. Example #3 selected waveforms (top to bottom): 1) microphone in Hz, 2) abdominal circumference relative variation, 3) thoracic circumference relative variation, 4) spinal posture relative variation from lordotic (negative values) to kyphotic (positive values), 5) shoulders bend relative variation from backward (negative values) to forward (positive values). Depiction pertains to the first 8-note except of 'Caro mio ben' aria.

More specifically, thoracic volume appears to collapse rapidly, while abdominal muscle activation seems to persist a while longer. Consequently, the second part of the singer's phrase is sung with what the expert described as "insufficient air pressure control through body support". This lack of proper muscular activation, and especially the chest cavity collapse, is coupled with a tendency for a kyphosis-like spinal and shoulder forward bend (which can be seen in lines 4 and 5 of Fig. 3). This conjunction of elements that are apparent in the data streams, illustrate a condition in which the singers' breathing muscles are contracted and tension starts to build up, as the body has no resources with which to control the subglottal pressure. The above effect could be what the expert judge noted as "muscular tension" in the thoracic spinal region.

*Example #4:* Contact Quotient (CQ) appeared to be over 0,60 for the most part of the trial, when computed with VoceVista using a hybrid method, where contact instant computed by the derivative of EGG signal, opening instant computed using a threshold set at 0,43 as indicated by Herbst [9]. For comparison with a previous study on CQ in pressed phonation [10], CQ was also calculated using a criterion method with a threshold set at 0,35 and was found to range between 0.61 and 0.73 for the specified note, values higher than previously reported even for pressed phonation [10]. This seems to be in accordance with the expert judge's mention of increased subglottal pressure.

*Example #5:* Vocal sound 'breathiness' characteristic in singing has been shown to be predictable from the audio and the EGG data, using computations, such as the Multi-Dimensional Voice Profile [3], or the (currently expanded/revised) multiple regression model CDH [11].

#### IV. DISCUSSION

A comprehensive inspection of the recorded data revealed evidence of expert-reported VTPs in most cases. Such information could be obtained either from singular sensor output, or through combinatory analysis of two or more synchronized data-streams. The sole case of reported VTP that was not recordable in the prototype data was "tongue lower part stiffness". Similarly, the sole 'indication' not quantified was "facial muscle activation", which was nevertheless recorded in video. Moreover, in response to a questionnaire question, the expert judge provided a list of suggested vocal exercises targeting various muscular systems, most of which can be monitored by the proposed prototype. This positions the prototype as a promising candidate to evolve into an assistive tool for vocal pedagogy.

*Limitations and Future work:* The present study served as a pilot study and, therefore, was conducted using a limited number of participants and judges. A large-scale study is already in progress, using a revisited analysis methodology, more participants and experts, and participant questionnaire analysis. Finally, there is the possibility of replacing the breathing sensors with more readily available options, as well as the potential evolution of the synchronization method and code.

*Acknowledgements:* The authors would like to thank Biopac Systems Inc. and David Fedorko, as well as the expert judge and participants to the study.

#### REFERENCES

- [1] P.M. Pestana, S. Vaz-Freitas, M.C. Manso, Trends in Singing Voice Research: An Innovative Approach, *J. Voice*. 33 (2019) 263–268. <https://doi.org/10.1016/j.jvoice.2017.12.003>.
- [2] P.J. Davis, S.P. Zhang, A. Winkworth, R. Bandler, Neural control of vocalization: Respiratory and emotional influences, *J. Voice*. 10 (1996) 23–38.
- [3] M. Aaen, J. McGlashan, K.T. Thu, C. Sadolin, Assessing and Quantifying Air Added to the Voice by Means of Laryngostroboscopic Imaging, EGG, and Acoustics in Vocally Trained Subjects, *J. Voice*. 35 (2021) 326.e1-326.e11.
- [4] S. Ternström, S. D'Amario, A. Selamtzis, Effects of the Lung Volume on the Electroglottographic Waveform in Trained Female Singers, *J. Voice*. 34 (2020) 485-e1.
- [5] F.M.B. Lã, M.B. Fiuza, Real-Time Visual Feedback in Singing Pedagogy: Current Trends and Future Directions, *Appl. Sci.* 12 (2022).
- [6] P. Chawah, S.K. Al Kork, et al., An educational platform to capture, visualize and analyze rare singing, *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*. (2014) 2128–2129.
- [7] J. Sundberg, Acoustic and psychoacoustic aspects of vocal vibrato, 1994.
- [8] E. Prame, Vibrato extent and intonation in professional Western lyric singing, *J. Acoust. Soc. Am.* 102 (1997) 616–621.
- [9] C.T. Herbst, Electroglottography – An Update, *J. Voice*. 34 (2020) 503–526.
- [10] K.G. Ong Tan, Contact Quotient of Female Singers Singing Four Pitches for Five Vowels in Normal and Pressed Phonations, *J. Voice*. 31 (2017) 645.e15-645.e22.
- [11] E. Angelakis, N. Kotsani, A. Georgaki, Towards a singing voice multi-sensor analysis tool: System design, and assessment based on vocal breathiness, *Sensors*. 21 (2021) 1–25.

# ELECTRODERMAL ACTIVITY AND ACOUSTICAL ANALYSIS IN A WORDS/NON-WORDS READING TASK

Federico Calà<sup>1</sup>, Lorenzo Frassinetti<sup>1,2</sup>, Pietro Tarchi<sup>1</sup>, Valentina Guarguagli<sup>1</sup>, Claudia Manfredi<sup>1</sup>, Antonio Lanata<sup>1</sup>,

<sup>1</sup> Department of Information Engineering, Università degli Studi di Firenze, Firenze, Italy

<sup>2</sup> Department of Information Engineering, Università degli Studi di Pisa, Pisa, Italy

[federico.cala@unifi.it](mailto:federico.cala@unifi.it), [lorenzo.frassinetti@unifi.it](mailto:lorenzo.frassinetti@unifi.it), [pietro.tarchi@unifi.it](mailto:pietro.tarchi@unifi.it), [valentina.guarguagli@stud.unifi.it](mailto:valentina.guarguagli@stud.unifi.it), [claudia.manfredi@unifi.it](mailto:claudia.manfredi@unifi.it), [antonio.lanata@unifi.it](mailto:antonio.lanata@unifi.it)

**Abstract:** Non-words are meaningless terms used in clinical research to detect and assess speech and language disorders. As an alternative to standardized scoring methods, this work proposes an innovative approach based on electrodermal activity (EDA) and speech, to evaluate possible physiological changes during a word/non-word reading task. A group of forty-five healthy volunteers were involved into experimental sessions. After preprocessing and male/female cohort separation, fifty features were extracted for each group. The aim was to investigate differences in the sympathetic nervous system activation between 4-syllables non-words, 4-syllables words and baseline conditions. Moreover, possible relationships between EDA and speech parameters for both groups were investigated. Results show that EDA can discriminate between tasks. For non-words, significant correlations were found between EDA signal complexity metrics, fundamental frequency and voice intensity.

**Keywords:** Words/non-words, sympathetic nervous system, electrodermal activity, speech analysis.

## I. INTRODUCTION

Speech and language disorders (SLD) concern up to 12% of the global population [1]. The American Speech-Language-Hearing Association defines language disorders as an impaired comprehension or use of spoken, written, or other symbol systems. It may involve and undermine language syntax, semantics and communication [1]. Such conditions can be caused by neurodegenerative diseases or traumatic events, and they usually appear during childhood. Early symptoms of SLDs are represented by a delay in speaking abilities, at around 2 years of age, with a significantly reduced expressive vocabulary with respect to typically developing children, without other concurrent developmental delays or sensory disorders [2]. These subjects present a higher risk of developing a form of

SLD, which can later negatively affect verbal communication and educational progress. Several methods exist for the screening, detection and monitoring of SLDs, such as questionnaires or test batteries [1, 3, 4]. Non-words repetition (NWR) and word/non-word aloud reading (WNWAR) represent standardized techniques based on meaningless terms, specifically designed to respect phonotactic characteristics and prosodic features of real words. As compared to other language measures, these tasks proved to be a relevant clinical marker in SLD research. WNWAR was successfully used to track changes and monitor improvements in fluency and pronunciation correctness in children with SLD before and after rehabilitation sessions [5, 6]. These evaluations are qualitative: clinicians listen to non-words utterances and note errors, stops, omissions and give scores. A useful integration is given by acoustical analysis, that allows an objective quantification of speech production capabilities and detect discomfort in participants' voice and speech in various experimental conditions [7, 8]. Another physiological signal that may be considered to quantify the activation of the sympathetic nervous system (SNS) is the electrodermal activity (EDA). Walsh and Ulser [9] have used its features to assess the awareness of stuttering in preschool children considering the relationship between SNS and speech production, whereas Marzi et al. [10] discovered a significant relationship between electrodermal activity and arousal in an affective word reading task. However, to the best of authors' knowledge, no studies investigated possible relationships between EDA and speech production. The aim of this study is to detect whether a NWAR task elicits any alteration in physiological responses and understand if relationships between them exist. The adopted paradigm is the same as in clinical and logopedic procedures. To evaluate and validate its robustness and accuracy, an exploratory analysis is performed on healthy subjects aiming at modelling their physiological elicited responses.



## II. MATERIALS AND METHODS

### A. Experimental set-up

Forty-five healthy volunteers (18 male M, 27 female F, mean age =  $22,4 \pm 2,1$  years) were recruited for this study. A DSI-24 (Wearable Sensing, San Diego, CA, USA) dedicated module was used to acquire EDA recordings. A Shure SM58 (Shure Inc., Niles, IL, USA) microphone was used for voice recordings. Sampling frequency was set at 300Hz for EDA and at 44.1kHz for voice recordings. Five Italian words (W) and five non-words (NW) of two, three and four syllables each were selected from the test battery developed by Dispaldro et al. [11], for a total of 30 stimuli. They were presented in two separate blocks to each participant: W first and then NW. The order of appearance of words and non-words within trials was randomised [12]. The entire dataset was not used to reduce the negative impact that instrumentation and experiment duration might have on fatigue and stress. Instead of repeating listened NW [4], subjects read stimuli from a monitor and then uttered them without time limits or correction from experimenters [5,6]. In this way, the role of phonological short-term memory, activated in the NWR task, was reduced. A motion artifact and three testing utterances were used to synchronize all devices. After a 180s baseline period, the experiment started and it was divided into three sections:

I - subjects completed a counting number task, read a standardized sentence and uttered three sustained cardinal vowels (/a/, /i/, /u/), following the guidelines of the Società Italiana di Fonologia e Logopedia [13].

II - subjects read 15 W.

III - subjects read 15 NW.

An optical sensor-based triggering system was used to track sections and transitions between stimuli.

In this work, only the results concerning electrodermal activity and speech, related to II and III will be presented.

### B. EDA Analysis

EDA signals were preprocessed with a low-pass FIR filter of order 500, with cut-off frequency 5 Hz and z-scored. The cvxEDA algorithm [14] was then applied for artifact removal, tonic (skin conductance level, SCL) and phasic (skin conductance response, SCR) component separation. SCL accounts for the slow-varying, spontaneous changes in the baseline of SC signal, and is related to the general psychophysiological state of the subject [15]. SCRs represent fast varying SC changes, directly evoked by stimuli. Features extraction was based on [16]: PS MSymp (SC mean spectral power in the 0.045-0.25Hz band), SCL Mean, SCL Std, SCR Peak (max peak

amplitude in SCR), SCR N (number of SCR peaks), SCR AmpSum (SCR peaks amplitude sum), Lat2 (latency between stimulus and peak). Moreover, Phasic and Tonic Sample Entropy (SE), and CompEDA [17] were included. These parameters were calculated both considering the whole duration of each task and single stimuli.

### C. Acoustical Analysis

Triggering signal and MATLAB 2020b (The MathWorks Inc., Natick, MS, USA) function detectSpeech.m were concurrently used to automatically segment audio files. Each utterance was processed with the open-source software BioVoice [18] giving 34 acoustical parameters concerning both frequency domain, such as fundamental frequency F0, formants and jitter, and time domain, such as voiced duration, pause duration, voiced/unvoiced percentage. According to [9] six more-parameters were computed: amplitude envelope (AmpEnv), root-mean square (RMS) and zero crossing rate (ZCR), spectral roll-off, spectral bandwidth and spectral centroid. All 40 parameters were computed separately for F and M groups to account for physiological differences between genders.

### D. Statistical Analysis

The first aim of this work was to evaluate whether EDA features only can discriminate between baseline, W and NW tasks. A Kruskal-Wallis test was performed on EDA metrics computed on the whole duration of the three tasks, both for tonic and phasic component. Pairwise differences were investigated by applying Bonferroni post-hoc correction. The remaining EDA features have been evaluated between W and NW tasks by means of a Mann-Whitney test. Level of significance was set at 0.05 for both tests.

Pearson correlation analysis between acoustical and EDA features was performed to understand whether SNS activation, expressed by EDA features, induces concurrent vocal properties alterations. Overall, 430 comparisons were investigated between each EDA and acoustical feature. The level of significance was 0.01. Correlations were calculated separately for the F and M cohorts.

## III. RESULTS

Table 2 shows the results of statistical analysis for EDA features computed for baseline, W and NW. Table 3 shows statistical results related to EDA features in the WNWAR tasks.

Table 2 – Statistical results for EDA analysis, considering features available for all the conditions. (\*), (#) and (+) denote significant differences between baseline and W, baseline and

NW, W and NW, respectively. The Bonferroni post-hoc correction was applied (Kruskal-Wallis test, level of significance 0.05). Median and interquartile range (iqr) values are reported.

EDA feature	Baseline Median (iqr)	Task	
		W Median (iqr)	NW Median (iqr)
<i>PS MSymp</i> <sup>**</sup>	0.89 (0.16)	0.48 (0.18)	0.96 (0.45)
<i>SCL Mean</i> <sup>**</sup>	-0.44 (0.22)	0.21 (0.26)	0.72 (0.33)
<i>SCL Std</i> <sup>**</sup>	0.90 (0.15)	0.58 (0.21)	0.24 (0.22)
<i>CompEDA</i> <sup>+</sup>	0.14 (0.15)	0.10 (0.12)	0.29 (0.16)
<i>SE Phasic</i> <sup>**</sup>	0.04 (0.04)	0.08 (0.07)	0.10 (0.09)
<i>SE Tonic</i> <sup>**</sup>	0.006(0.005)	0.009(0.006)	0.019(0.014)

Table 3 – Statistical results for EDA analysis, considering features available only for the W and NW tasks. (+) denotes a statistical difference between W and NW (Mann-Whitney test, level of significance 0.05). Median and interquartile range (iqr) values are reported.

EDA feature	Task	
	W Median (iqr)	NW Median (iqr)
<i>SCR peak</i> <sup>+</sup>	1.13 (1.02)	0.76 (0.82)
<i>SCR N</i> <sup>+</sup>	84 (29)	56 (21)
<i>SCR AmpSum</i> <sup>+</sup>	28.2 (29.0)	14.7 (11.3)
<i>Lat2</i>	45.1 (59.8)	37.3 (94.2)

Concerning the correlation analysis between acoustical and EDA features, Table 4 reports Pearson’s  $\rho$  coefficient and the p-value for the W and the NW tasks, limited to four syllables stimuli. Only the significant correlations in common between M and F are reported.

Table 4 – Correlation results between EDA and acoustical parameters. Level of significance 0.01.

Correlation	W			
	F		M	
	$\rho$	p-value	$\rho$	p-value
<i>SCR N – F0<sub>median</sub></i>	0.24	0.003	-0.37	e-0.4
<i>SE Tonic – ZCR</i>	-0.24	0.004	-0.27	0.008
Correlation	NW			
	F		M	
	$\rho$	p-value	$\rho$	p-value
<i>Comp EDA – RMS</i>	-0.23	0.006	0.37	e-04
<i>SCR N – F0<sub>median</sub></i>	0.22	0.007	-0.36	e-04
<i>Comp EDA – AmpEnv</i>	-0.24	0.004	0.29	0.004

Figure 1 shows the scatter plot for CompEDA and RMS for M (observations are denoted by ‘x’) and F (‘o’) groups, as well as their respective regression lines (solid and dashed for M and F cohorts, respectively).

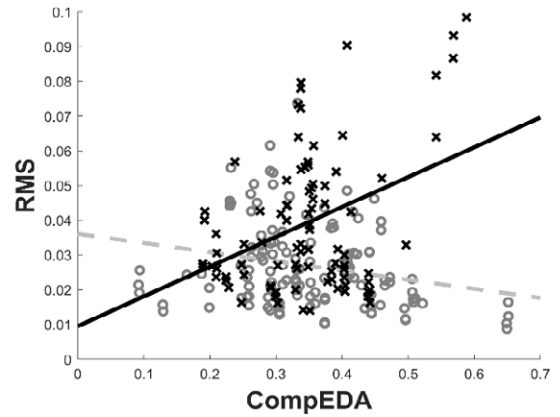


Figure 1: Scatter plot between CompEDA and RMS for M (x) and F (o) groups.

#### IV. DISCUSSION

This work reports a preliminary analysis of physiological signals acquired during W and NW tasks. The aim is assessing possible physiological changes during NW reading that can be detected by EDA and speech analysis.

Firstly, it was evaluated if EDA features are able to discriminate between tasks and baseline. EDA features related to the tonic component (e.g., SCL Mean, SCL Std) allowed a pairwise discrimination between the three tasks. Significant differences were obtained as well on EDA phasic component features. For instance, SE phasic or SCR peak allowed discriminating between W and NW tasks. Furthermore, the NW task has the highest entropy values for the phasic component (SE Phasic) as compared to the W task and the baseline (Table 2).

The M cohort presents a negative correlation on F0 median ( $\rho=-0.37$ , p-value e-04), while the F cohort shows a positive correlation for the same parameter ( $\rho=0.24$ , p-value 0.003). An alteration of F0 may be associated with discomfort, stress or cognitive workload [9, 19]. Giddens et al. [19] highlighted that both increases and decreases of F0 can be detected in response to such conditions. Table 4 shows that the correlation of F0 median with EDA feature was found significant both for the W and NW tasks: thus, this shift may be simply induced by experimental conditions regardless of the proposed tasks.

Table 4 shows a relevant difference between W and NW tasks. Specifically, while EDA signal complexity increases during NW reading, which is expressed by CompEDA, opposite effects on RMS and AmpEnv acoustical features appear, in line with [19]. These parameters represent good approximations of vocal intensity, therefore it seems that, when uttering NW, M tend to increase their loudness, whereas F decrease it (Figure 1).

To assess if the considered features, as well as the administered tasks, could be sensitive to SLDs this approach must be applied on a larger dataset and on pathological subjects. Also, future work should be devoted to account for various inter-physiological factors such as higher sensitivity to stress or stress unresponsiveness to this task. Nevertheless, results suggest that EDA and speech features can be used as an objective alternative to scoring methods in their assessment. Future work will extend analyses to words with different syllable length and will consider EEG analysis for brain activity differences during the proposed tasks, correlating its features with speech and EDA ones, as well as investigating possible causality relationships among them.

#### V. CONCLUSION

This paper proposes a first exploratory analysis of physiological signals acquired on healthy subjects during a W and NW reading test. Preliminary results suggest that EDA features allow discriminating SNS activity during three tasks. Moreover, significant relationships between EDA complexity measures, F0 and voice intensity were observed. These outcomes open the way to the application of the proposed approach to pathological subjects, to better characterize SLDs and cognitive delays.

#### REFERENCES

- [1] I. F. Wallace *et al.*, “Screening for speech and language delay in children 5 years old and younger: A systematic review,” *Pediatrics*, vol. 136, no. 2, 2015.
- [2] R. W. Cheung, C. Hartley, and P. Monaghan, “Multiple mechanisms of word learning in late-talking children: A longitudinal study,” *J Speech Lang Hear*, vol. 65, no. 8, pp. 2978–2995, 2022.
- [3] G. Sartori and R. Job, *DDE-2: Giunti O.S. Organizzazioni Speciali: Batteria per La Valutazione Della Dislessia e Della Disortografia Evolutiva-2: Protocollo Di Registrazione*. Firenze: Giunti O.S., Organizzazioni Speciali, 2007.
- [4] M. Dispaldro, B. Benelli, S. Marcolini, and G. Stella, “Real-word repetition as a predictor of grammatical competence in Italian children with typical language development,” *IJLCD*, vol. 44, no. 6, pp. 941–961, 2009.
- [5] A. Battisti *et al.*, “Effects of a short and intensive transcranial direct current stimulation treatment in children and adolescents with developmental dyslexia: A crossover clinical trial,” *Front Psychol*, vol. 13, 2022.
- [6] F. Corallo *et al.*, “Improvement of self-esteem in children with specific learning disorders after donkey-assisted therapy,” *Children*, vol. 10, no. 3, p. 425, 2023.
- [7] A. Hall *et al.*, “Acoustic analysis of surgeons’ voices to assess change in the stress response during surgical in situ simulation,” *BMJ Simul*, vol. 7, no. 6, pp. 471–477, 2021.
- [8] Arushi, R. Dillon, and A. N. Teoh, “Real-time stress detection model and voice analysis: An Integrated VR-based game for Training public speaking skills,” *2021 IEEE Conference on Games (CoG)*, 2021.
- [9] B. Walsh and E. Usler, “Physiological correlates of fluent and stuttered speech production in preschool children who stutter,” *J Speech, Lang Hear*, vol. 62, no. 12, pp. 4309–4323, 2019.
- [10] C. Marzi, A. Greco, E. P. Scilingo, and N. Vanello, “Towards a model of arousal change after affective word pronunciation based on electrodermal activity and speech analysis,” *Biomed Signal Process Control*, vol. 67, p. 102517, 2021.
- [11] M. Dispaldro, L. B. Leonard, and P. Deevy, “Real-word and nonword repetition in Italian-speaking children with specific language impairment: A study of diagnostic accuracy,” *J Speech Lang Hear*, vol. 56, no. 1, pp. 323–336, 2013.
- [12] N. Ahufinger *et al.*, “Consistency of a nonword repetition task to discriminate children with and without developmental language disorder in Catalan–Spanish and European portuguese speaking children,” *Children*, vol. 8, no. 2, p. 85, 2021.
- [13] Maccarini, A. R. and Lucchini, E., ‘La valutazione soggettiva ed oggettiva della disfonia. Il Protocollo SIFEL, Relazione Ufficiale al XXXVI Congresso Nazionale della Società Italiana di Foniatria e Logopedia’, *Acta Phoniatica Latina*, vol. 24, no. 1-2, pp. 13-42, 2002.
- [14] A. Greco, G. Valenza, A. Lanata, E. Scilingo, and L. Citi, “CVXEDA: A convex optimization approach to electrodermal activity processing,” *IEEE Trans Biomed Eng*, pp. 1–1, 2016.
- [15] A. Greco *et al.*, “Acute stress state classification based on Electrodermal Activity modeling,” *IEEE Trans Affect Comput*, vol. 14, no. 1, pp. 788–799, 2023.
- [16] A. Baldini *et al.*, “Subjective fear in virtual reality: A linear mixed-effects analysis of skin conductance,” *IEEE Trans Affect Comput*, vol. 13, no. 4, pp. 2047–2057, 2022.
- [17] M. Nardelli, A. Greco, L. Sebastiani, and E. P. Scilingo, “ComEDA: A new tool for stress assessment based on electrodermal activity,” *Comput Biol Med*, vol. 150, p. 106144, 2022.
- [18] M. S. Morelli, S. Orlandi, and C. Manfredi, “BioVoice: A multipurpose tool for voice analysis,” *Biomed Signal Processing Control*, vol. 64, p. 102302, 2021.
- [19] C. L. Giddens, K. W. Barron, J. Byrd-Craven, K. F. Clark, and A. S. Winter, “Vocal indices of stress: A Review,” *J Voice*, vol. 27, no. 3, 2013.

**ROUND TABLES, LABORATORY AND  
LECTURE**



## **ROUND TABLE I: ACOUSTIC AND PHYSIOLOGICAL ASPECTS OF SINGING**

Moderator: Johan Sundberg.

Panelists: Silvia Capobianco, Nathalie Henrich Bernardoni, Malte Kob.

Acoustic and physiological aspects of singing is relevant from several points of view.

First, singing develops for esthetical purposes, such as in *Lieder Abend*, as well as for practical purposes, such as in kulning, for calling cattle in the forest singing. Hence singing is a manifestation of human culture deserving scientific analysis aiming at deepening the understanding of mankind.

Second, professional singers can use their voices extensively without detrimental effects on the voice organ. This implies that they have developed a vocal technique that contains the principle of vocal economy. This makes professional singing highly relevant also to voice care, education and training.

Third, singers learn how to orthogonalize the three phonatory dimensions vocal loudness, pitch and phonation type. For non-singers, by contrast, these dimensions are typically interdependent; increase of vocal loudness is typically associated with increase of both pitch and phonation type. This makes research on singers' voice function particularly worthwhile.

In this session three aspects of vocal art in singing will be elucidated. Silvia Capobianco will report on her and her associates' study of Acoustical features of Early Music Singing. Since the recent Early Music (EM) revival, a subset of singers has begun to specialize in a style of singing that is perceptually different from the more "mainstream" Romantic Operatic (RO) singing style. The aim of this contribution is to describe EM singing in terms of acoustic analysis of voice and breathing patterns.

In a first study, 10 professional singers (5 F; 5M) versed in both the EM and the RO repertoire were enrolled in the study. Each singer recorded the first 10 bars of the famous Aria "Amarilli Mia Bella" (Giulio Caccini, 1602) a cappella, in RO and EM styles, in random order. Vibrato features and the Singer's Formant power were analyzed using the software Biovoice. Vibrato in EM singing was characterized by a higher rate, a smaller extent, and less regular cycle-cycle period duration compared to RO singing. As in previous studies, RO singing presented a more prominent Singer's Formant, as indicated by a smaller QR.

In a second study, a novel portable device that simultaneously monitors vocal activity and breathing patterns without interfering with natural singing was developed, combining a miniature accelerometer to measure vocal doses from skin vibrations on the neck and two respiratory inductive plethysmography (RIP) bands to estimate the breathing pattern by measuring changes in the thoracoabdominal cross-sectional area. The device was tested on 13 professional EM singers and 14 untrained individuals during the execution of singing tasks. EM singers demonstrated a higher asynchronous motion between ribcage and abdomen during singing and a reduced use of the thoracic compartment in favor of a more compliant abdominal compartment. Comparing vocal doses with breathing patterns, it was possible to build graphs where "efficiency regions" identified strategies applied only by singers.

Acoustical and breathing patterns analysis was used to characterize EM singing, highlighting peculiar features in comparison with Romantic Operatic singing. Given the acoustical distinctions between EM and RO styles, future scientific and musicological studies should consider distinguishing between the two styles rather than using a singular term for Western Classical singing.

In the second talk Nathalie Henrich Bernardoni will present Insights Into Mixing in Classical and Modern Singing. Among the world of singing-voice registers, the notion of «mixed voice» is one of the most puzzling. Both classical and non-classical singers use this particular register to avoid vocal breaks during *passaggio*. In this presentation we will explore the physiological and acoustical characteristics of mixing in singing on two main databases: a database composed of professional lyrical singers and a database composed of modern pop-rock singers. We will discuss the main behaviours shared by all singers and the specificities of each singing style.

In the third talk, Studies of Singers' Use of Velopharyngeal Opening I will overview of a series of investigations I have had the privilege of carrying out together with colleagues from different fields, phoniatrician Miriam Havel, singing teacher Brian Gill, singer Jessica Lee, voice researcher and singer Filipa MB Lã, and acoustician Svante Granqvist.

Advantages and disadvantages of singing with a velopharyngeal (VP) opening has long been a theme of controversies between teachers of singing and between voice experts. Some argue that the VPO should be “patently closed” in vowel production, while others report voice improvements after instructing students to sing vowels with a narrow VP opening. I will present a family of studies, which I have had the privilege of carrying out with friends and colleagues from different fields, and which seem to shed light on effects of such an opening.

Together with Birch and associates I examined the VP port in 17 professionally performing opera singers by means of a nasofiberscope. For the vowels /u, a/, but rarely for the vowel /i/, we found VP openings of different sizes in many of these singers. A listening test revealed that a narrow VP port did not necessarily result in a nasalized vowel quality. In a follow-up study we analyzed the effect of connecting a 10 cm long quasi-nasal-tube resonator to a 20 cm long-quasi-vocal tract tube. We found, as expected, that the resonance frequency of the shorter tube created a dip in the transfer function of the quasi-vocal tract tube.

In an experiment with Gill and associates we asked singers to sing a vowel sequence at different pitches with the VP port (i) closed, (ii) slightly open and (iii) wide open. We measured the audio signal and also the oral and nasal airflow signals as picked up by a Glottal Enterprises flow mask. The latter signals allowed us to verify that the singers managed to produce reliable examples of the three conditions. The audio signal was analyzed in terms of long-term-average spectra (LTAS). The results showed that, as an average across participants, a narrow VP opening enhanced the level of the 2 – 4 kHz range of the LTAS by about 5 dB relative to the overall LTAS level. In other words, a narrow VP opening tended to change the spectrum balance in favor of the spectrum partials in the frequency range of the singers’ formant cluster.

Recent experiments with Filipa MB Lã and Svante Granqvist support the assumption that an open VP port can reduce the risk for voice breaks caused by source-filter interaction. Nine student choir singers sang glide tones on an intended neutral vowel while pressing against the mouth the end of a 70 cm long tube. They did this under three conditions: (i) with the far end of the tube open, (ii) with the far end of the tube open but while nasalizing the vowel and (iii) with a piece of cotton wool in the far end, thus attenuating the lowest resonances of the compound vocal tract&tube resonator. Under condition (i) a great number of voice breaks were observed, but the number of breaks was almost halved in (ii), when the participants nasalized the vowel. Likewise, it was almost halved under condition (iii) when a piece of cotton wool attenuated the lowest resonances. Thus, the risk of voice breaks was reduced when the lowest resonances were attenuated.

Together with Miriam Havel and associates I have experimentally measured effects of a VP opening on the sound transfer of the vocal tract. We used 3-D models of vocal tracts coupled to 3-D models of nasal tracts via coupling tubes of different sizes. Increasing the cross-sectional area of the coupling tubes left the high frequency range basically unaffected but systematically attenuated the low frequency range, like the piece of cotton wool did in the 70 cm long tube. This supports the assumption that a VP opening can reduce the risk for voice breaks caused by source-filter interaction. However, it also shows that it can increase the levels in the high frequency range, an effect that otherwise typically requires an increase of vocal effort.

Taken together, these studies suggest that a narrow VP opening, habitually used by many professional opera singers, can reduce the risk for voice breaks due to source-filter interaction and at the same time automatically enhance the singers’ formant cluster without requiring increase of vocal effort.

**Author:** Malte Kob

**Abstract:** The interaction of voice source and voice filter play a major role for voice timbre. On one hand the source characteristics is altered by the transfer through vocal and nasal tracts which challenges the estimation of source properties using acoustic methods. On the other hand the voice source is the end part of the tracts and provides a time-variant boundary condition to these resonators, ranging from open to closed. The mouth opening finally provides another boundary condition which balances the energy between internal resonance support and projection to the listeners. Singers can modulate all boundaries to achieve specific timbres. This part of the round table discusses the potential use of pedagogic and physical methods to teach and understand these interactions.

## **ROUND TABLE II: FOCUSING ON VOICE ONSET: A CRITICAL MOMENT OF PHONATION. FROM BASIC SCIENCE TO THERAPY**

Moderator: P. H. DeJonckere.

Panelists: Johan Sundberg, Giovanna Cantarella, Malte Kob, Philipp Aichinger.

### **Outline:**

- Introduction and presentation of the topic
- Typology – imaging – biophysics
- Acoustics / EGG
- Glottal and vocal tract impedance in the prephonatory and phonatory status
- Modelling
- Heath and deviant/dyskinetic modalities of voice onset in (artistic) speech & song
- Clinical aspects, phonotraumatic tissue reactions and treatment options

The onset of vocal fold vibration is a complex transient event, in which the forces at play progressively adjust until a stationary state is reached. The acting forces are lung pressure, intraglottal pressure, myoelastic tension of the vocal fold oscillator generating the glottal impedance, and inertance of the supraglottal vocal tract.

Three categories of vocal onsets are generally recognized: soft (or “coordinated”), hard, and breathy (or “aspirate”). In normal subjects, the most frequently observed type of voice onset in spontaneous speech is the soft onset, and it may be considered as the “physiological” onset, with a few oscillations (possibly a single one) preceding the first glottal closure (‘collision-free oscillations’). Singers sometimes use the term ‘articulations’, e.g. differentiating: (1) staccato, that is short tones separated by short voiceless segments; (2) voiceless aspirated bilabial plosive followed by a vowel and (3) unaspirated bilabial plosive followed by a vowel. Accurate synchronization of glottal adduction and building up of subglottal pressure is essential and fails in pathology.

Different types of phonation onset have different acoustic characteristics, as well as typical airflow patterns. According to Sundberg, in a breathy onset the transglottal airflow waveform results in a voice source with strongly dominating fundamental, whereas the higher overtones arrive with a slight delay. In hard glottal attacks, this delay is typically minimized or eliminated.

The combined physical, physiological, imaging and acoustic parameters measured simultaneously in vivo provide a detailed qualitative and quantitative insight into the complex mechanisms of vocal onset and make possible a comprehensive understanding of the intraglottal mechanical events and fluid dynamics, particularly turbulence, at the precise moment when oscillation starts. This is essential for vocal modelling and voice synthesis, as well as for automated speech recognition and substitution voicing.

Also, the specific acoustic and airflow patterns characterizing the different types of onset, are suited for differentiation by machine learning and practical applications in pedagogy of artistic voice (singing / acting) as well as in early detection of voice diseases.

Relevance of specifically assessing the mechanism of voice onset is supported by the clinical experience: Habitual, dyskinetic respiratory and laryngeal behaviors may lead to voice complaints and phonotraumatic tissue reactions requiring therapeutic approaches.



**Title:** Advances and Frontiers in the Analysis and Synthesis of Pathological Voices: Voice Onsets and Offsets

**Author:** Philipp Aichinger.

**Affiliation:** Speech and Hearing Science Lab, Div. Phoniatrics-Logopedics, Dept. Otorhinolaryngology, Medical University of Vienna.

**Abstract:** Voice onsets have a profound influence on voice quality. Classical features of the voice onset include the voice onset time and phonation threshold pressure. However, in the past, the analysis of midvowel segments often took priority over the analysis of the voice onset. In this presentation, we will revisit onsets and offsets in diplophonic voice, voice onsets in simulated high-speed videos of vocal fold vibration, as well as recent advances in deep learning-based speech synthesis.

For analyzing and modeling onsets and offsets in diplophonia, a hidden Markov model (HMM) was used [1]. It relies on modeling the onset and offset probabilities of glottal oscillators over time. In the analysis-by-synthesis of glottal area waveforms of sustained phonation, it was shown that the model fidelity achieved by an HMM-based system exceeded the fidelity achieved by a deep autoencoder and was approximately on par with the so-called 'WaveGlow' approach. As an added value, the HMM has a larger explanatory power than the other two approaches.

The HMM-based model was also used to model the frequency of occurrence of diplophonation in audio recordings of German standard text readings [2]. A feature termed 'diplophonia rate (%)' enabled the distinction between frequently diplophonic speakers, rarely diplophonic speakers, and non-diplophonic speakers.

For simulating voice onsets in high-speed videos of vocal folds, a kinematic model was combined with computer graphics [3]. The most important control parameters for modeling voice onset are the posterior glottal opening (mm), enabling abduction and adduction of the vocal folds, as well as the vibratory amplitude (mm), enabling the start and stop of the vibration. For an example of a breathy voice onset, temporally smooth adduction and amplitude transition were combined. The initial posterior glottal width was 4 mm and reduced to 0 mm within 25 ms, approximately. Simultaneously, the amplitude was increased from 0 mm to 0.7 mm. For modeling an example of a pressed voice onset, glottal width was 0 prior to the fade-in of the amplitude, which is faster in the pressed example than in the breathy example.

In conclusion, recent advances in the modeling of voice onset have improved our understanding of voice function. Deep learning-based synthesis is astonishingly realistic but lacks physiological explanation. For clinical assessment and the control of treatment, features of voice onset and offset may contain relevant information. For example, voice disorders influencing the voice onset include vocal fold paralysis, laryngitis, and functional dysphonia. Detailed analyses of the vocal fold kinematics and acoustics at voice onset and offset most likely require the simultaneous use of high-speed videolaryngoscopy and audio recordings, which should be considered to become the standard.

## References

- [1] 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, doi: 10.1109/TASLP.2021.3053387.
- [2] P. Aichinger and J. Schoentgen, 'Detection of Diplophonation in Audio Recordings of German Standard Text Readings,' *J. Voice*, vol. 33, no. 6, pp. 949.e1–949.e10, 2019, doi: 10.1016/j.jvoice.2018.06.009.
- [3] P. Aichinger, S. Kumar, H. Lehoux, and J. Švec, 'Simulated Laryngeal High-Speed Videos for the Study of Normal and Dysphonic Vocal Fold Vibration,' *J. Speech, Lang. Hear. Res.*, vol. 65, no. 7, pp. 2431–2445, 2022, doi: 10.1044/2022\_JSLHR-21-00673.
- [4] T. L. Shiba and D. K. Chhetri, 'Dynamics of Phonatory Posturing at Phonation Onset,' *Laryngoscope*, vol. 126, no. 8, pp. 1837–1843, 2016, doi: 10.1002/lary.25816.

**Author:** Malte Kob

**Abstract:** The conditions of respiratory organs, larynx and vocal articulators determine the source characteristics as well as the timbre and radiation of the voice. Whereas the glottal status determines pitch, source spectrum, and the voice quality (from pressed to breathy), the articulatory organs adjust the resonators of the vocal and nasal tracts for a desired voice timbre (formants) and projection (mouth opening). The pre-phonatory configuration implements the intended voice characteristics by adjustment of laryngeal and articulatory muscles, cartilages and ligaments, shortly before phonation starts. In this part of the round table methods for potential assessment of the physiology before, during and after voice onset are discussed.

**LECTURE:  
THE « VIRTUAL MUSEUM OF PHONIASTRICS »  
A GUIDED TOUR BY THE CURATOR-IN-CHIEF  
P. H. DEJONCKERE**

As a multifaceted discipline, Phoniatics has a rich and diverse history. This worthwhile past constitutes a heritage that we need to safeguard and to organize, in order to make it better known, available for more in depth investigation and for referral.

These are the main aims of the “Virtual Museum of Phoniatics”, that has been named, as a respectful tribute, after Prof. Dr. Antoinette Am Zehnhoff-Dinnesen, former President of the UEP.

Like for most medical disciplines, the history of Phoniatics involves aspects related both to basic sciences and to clinical practice, but to societal and cultural issues as well, with prominent figures, writings, instruments, techniques and significant milestones marking the progress in understanding and managing disease and cure. All these items receive an appropriate place within one of the seven galleries of the Museum:

Persons (not including persons who are still alive)

Instruments and devices (for physiology / diagnosis / assessment / treatment)

Congresses (programs / proceedings...)

Historical books

Historical articles

Articles on history

Historical videos / films

Nowadays, digitization makes possible a true ‘virtual’ museum, designed to be freely accessible at any time within the UEP-website.

Obviously, the current collection is but a starting point, and it requires being complemented and upgraded by additional material and documentation over time.

The presentation will consist in a guided tour, giving a general overview of the current collection and illustrating more in detail some noteworthy items.



# **LABORATORY: A MULTIMODAL SYSTEM FOR THE CHARACTERIZATION OF PHYSIOLOGICAL DYNAMICS DURING SIFEL PROTOCOL AND WORDS/NON-WORDS READING TASKS**

Organizers: Lorenzo Frassinetti, Federico Calà, Pietro Tarchi, Valentina Guarguagli, Claudia Manfredi, Antonio Lanatà.

In the last years, the possibility to integrate more sources of information has allowed a better characterization and a deeper comprehension of the neurophysiological dynamics involved in several cognitive and decisional processes (Lahat et al., 2015, Verma and Tiwary 2014, Ha et al., 2015). This multimodal approach was applied in several Event Related Potential (ERP) tasks, and it has successfully highlighted significant correlations between different types of biosignals which might underline relationships between physiological systems as well (Muhammad et al. 2021, Nweke et al. 2019).

During the experimental session, a multimodal system will be shown and applied to volunteer subjects. A number of biosignals will be recorded in two separate experiments: a subset of the SIFEL Italian protocol (Lucchini et al. 2002) and a protocol for the assessment of dyslexia (Job and Tressoldi 2007), based on Words and Non-Words reading (Dispaldro et al. 2013).

In detail, during the experience the following devices and their integration/synchronization will be presented:

- Electroencephalogram DSI-24 (dry-EEG) (Wearable Sensing, San Diego, CA, USA)
- Electrocardiogram Shimmer Sensing 3 (Shimmer Research Ltd, Dublin, Ireland)
- Galvanic Skin Response (GSR) device (Wearable Sensing, San Diego, CA, USA)
- Microphone SHURE
- Trigger HUB (Wearable Sensing, San Diego, CA, USA) and photodiode for the device synchronization
- Display on a monitor for task presentation.

Volunteers from MAVEBA 2023 will perform the following reading tasks:

- From the SIFEL protocol (Lucchini et al. 2002):
  - number listing task (from 1 to 10)
  - the italian sentence “io amo la aiuole della mamma”
  - vowels /a/, /i/, /u/ sustained for three seconds
- From the Italian Words-Non/Words task for dyslexia assessment proposed by Dispaldro et al. 2013:
  - 15 Words
  - 15 Non-Words

The aim of these experiments was to evaluate and demonstrate that, by combining different sources of information, it is possible to obtain a deeper insight into physiological dynamics regarding speech processing aspects not yet investigated, allowing a better understanding of several voice and speech disorders. Such systems may also be used to quantify both the disease progression and the effects of rehabilitation program (e.g. logopedics).

Although the experiment was designed for Italian speakers, at MAVEBA 2023 participation is open/encouraged and extended to all the attendees. This could represent an interesting point of discussion regarding the evaluation of signals recorded from non-Italian speakers or possibly bilingual speakers.



## INDEX OF AUTHORS

- Aichinger P. 87, 107  
Andreopoulou A. 47, 79, 95  
Angelakis E. 95  
Arias-Londoño J. D. 69
- Bakogiannis K. 79, 95  
Baracca G. 43  
Bastiani L. 29, 43  
Battilocchi L. 83  
Berrettini S. 29  
Blažauskas T. 65  
Bruschini L. 29  
Buccichini G. 83  
Butala A. 75, 91
- Calà F. 83, 99, 111  
Cantarella G. 83, 107  
Cao T. 75  
Capobianco S. 29, 43, 105  
Chovaz V. 91  
Clawson L. 91  
Cust S. 91
- Dadras A. 87  
Damaševičius R. 65  
Dedousis G. 79  
Dehak N. 75, 91  
DeJonckere P. H. 33, 107, 109  
Drioli C. 19
- Evdokimova V.V. 37  
Evgrafova K.V. 61
- Favaro A. 75  
Foresti G. L. 19  
Frassinetti L. 83, 99, 111  
Frère J. 15
- Georgaki A. 47, 79, 95  
Gerber S. 15  
Godino-Llorente J. I. 69  
Guarguagli V. 99, 111
- Habela M. 95  
Hagmüller M. 23  
Henrich Bernardoni N. 15, 105  
Huang J. 85
- Ibarra-Sulbaran E.J. 69
- Kob M. 105, 107
- Lanata A. 83, 99, 111  
Lebacqz J. 33  
Linke J. 23  
Loevenbruck H. 15  
Lohrmann S. 23
- Manfredi C. 83, 99, 111  
Maskeliūnas R. 65  
Maximova M. R. 37  
Moro-Velazquez L. 75, 91
- Nacci A. 29, 43
- Oh E.S. 75
- Park J. 91  
Paroni A. 15  
Pokorny F. 23  
Pribuišis K. 65  
Pützer M. 51
- Reyes-Galaviz O. F. 57  
Reyes-Garcia C. A. 57
- Schuppler B. 23  
Shvaley N.V. 61  
Simoni F. 29  
Sokolova N.S. 61  
Sundberg J. 105, 107
- Tarchi P. 99, 111  
Thebaud T. 75
- Uloza V. 65  
Ulozaitė-Stanienė N. 65
- Valencia-Hernandez I. A. 57  
Villalba J. 75
- Wang H. 91  
Wokurek W. 51
- Zañartu M. 69  
Zelasko P. 75  
Zhang J. 91  
Zinkus M. 91





ISSN 2704-601X (print)  
ISSN 2704-5846 (online)  
ISBN 979-12-215-0145-2 (Print)  
ISBN 979-12-215-0146-9 (PDF)  
ISBN 979-12-215-0147-6 (XML)  
DOI 10.36253/979-12-215-0146-9

[www.fupress.com](http://www.fupress.com)