UNIVERSITÀ
DEGLI STUDI
FIRENZE
DINFO
DIPARTIMENTO DI
INGEGNERIA
DELL'INFORMAZIONE

14th
INTERNATIONAL
WORKSHOP

MODELS AND ANALYSIS OF VOCAL EMISSIONS FOR BIOMEDICAL APPLICATIONS

December 16-17, 2025
Firenze, Italy

PROCEEDINGS

FIRENZE
UNIVERSITY
PRESS

PROCEEDINGS E REPORT

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

139

MODELS AND ANALYSIS OF VOCAL EMISSIONS FOR BIOMEDICAL APPLICATIONS

14TH INTERNATIONAL WORKSHOP

December, 16-17, 2025
Firenze, Italy

Edited by
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi

Firenze University Press
2025

Models and Analysis of Vocal Emissions for Biomedical Applications : 14th International Workshop, December, 16-17, 2025 / edited by Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi. - Firenze : Firenze University Press, 2025.
(Proceedings e report ; 139)

https://books.fupress.com/isbn/9791221508215

ISSN 2704-601X (print)
ISSN 2704-5846 (online)
ISBN 979-12-215-0820-8 (Print)
ISBN 979-12-215-0821-5 (PDF)
ISBN 979-12-215-0822-2 (XML)
DOI 10.36253/979-12-215-0821-5

Cover: designed by CdC, Firenze, Italy.

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

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

MAVEBA 2025

Firenze, Italy

The MAVEBA 2025 Workshop is sponsored by:

Electronics – Section Bioelectronics | An Open Access Journal from MDPI
https://www.mdpi.com/journal/electronics/sections/bioelectronics

Logopedics Phoniatrics Vocology (LPV, Taylor & Francis)
https://www.tandfonline.com/journals/ilog20

La Voce Artistica
https://www.voceartistica.it/

Fondazione CR Firenze
https://fondazionecrfirenze.it

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

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

CONTENTS

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

SPECIAL SESSION II – HISTORICAL ASPECTS OF VOICE RECORDINGS AND ANALYSIS (organized by P.H. DEJONCKERE)

SPECIAL SESSION III – UPDATES ON VOICE RANGE PROFILE MEASUREMENTS (organized by M.KOB and G. BARACCA)

SESSION IV – VOICE EMOTIONS AND PERSONALITY

MAVEBA
2025
Firenze, Italy

FOREWORD

This book of Proceedings includes the contributions presented at the 14th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications – MAVEBA 2025, held in Firenze from 16 to 17 December, 2025. The previous edition of MAVEBA, held in September 2023 was celebrated on site in a nice Florentine summer, that allowed participants enjoy outdoor visits. This 14th edition comes back to the traditional winter time, which compensates for the colder climate with the warmth of the Christmas atmosphere. We are both happy and proud that we are together again: a demonstration of strength and continuity of this small scientific community that kept us in touch and in friendship despite the difficulties for a quarter of a century.Indeed, the series of MAVEBA International Workshops started in 1999 and, overcoming pandemics, 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 and can benefit from new and powerful scientific tools such as Artificial Intelligence. The emphasis is still translational research, the link with the "real" complex world of the human being. This 14th 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, phoniatricians, 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 2025 are divided into three Special Sessions and other four Sessions:

SPECIAL SESSIONS
I – THE ROLE OF AI IN THE FIELD OF PHONIATRICS AND LARYNGOLOGY: PRESENT AND FUTURE (organized by G. CANTARELLA)
II – HISTORICAL ASPECTS OF VOICE RECORDINGS AND ANALYSIS (organized by P.H. DEJONCKERE)
III – UPDATES ON VOICE RANGE PROFILE MEASUREMENTS (organized by M.KOB AND G. BARACCA)

SESSION I – ANALYSIS OF PATHOLOGICAL VOICE
SESSION II – SINGING VOICE ANALYSIS
SESSION III – VOICE ANALYSIS AND SYNTHESIS
SESSION IV – VOICE EMOTIONS AND PERSONALITY

As for the past edition, MAVEBA 2025 includes a Student Competition, which selects the best paper, both as far as contents and the student presentation and communication skills. A prize of 200 CHF is offered by Electronics Journal (MDPI, Section Bioelectronics). Moreover, the winner will be granted by a certificate and one year of free online access offered by Logopedics Phoniatrics Vocology journal (LPV), an international, scientific, peer-reviewed, open access journal. We are very grateful to both journals for the generous offer that allows young scholars to see their talent recognized.

ACKNOWLEDGEMENTS
We greatly acknowledge PhD. student Eng. Federico Calà, PhD. student Eng Pietro Tarchi, Eng. Valentina Guarguagli, who managed and constantly updated the website, collaborated in reviewing the Proceedings and in solving the daily difficulties with patience and professionalism. Last but not least, we would like to thank the anonymous reviewers for their dedication and constructive criticism on the papers that are collected in this volume in their corrected version. Thanks also to the Fondazione Cassa di Risparmio di Firenze for the economic contribution, 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 us for many years in this adventure. But above all we thank all the participants who, with their presence, wanted to be next to us 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
Antonio Lanatà
Lorenzo Frassineti
MAVEBA 2025 Chairs

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

SESSION IA
ANALYSIS OF PATHOLOGICAL VOICE

VOICE MAPPING IN CLINICAL PRACTICE: TRACKING OBJECTIVE CHANGES AFTER INJECTION LARYNGOPLASTY

S. Capobianco[1], G. Björck [2], F. Forli [1,3], L. Bruschini [1], A. Nacci [1], S. Ternström [4]

[1] ENT, Audiology and Phoniatrics Unit, Pisa University Hospital, Pisa, Italy
[2] Division of Ear, Nose and Throat, Department of Otorhinolaryngology, Phoniatric Section, Karolinska University Hospital, CLINTEC, Karolinska Institutet, Stockholm, Sweden
[3] Implant Section, Karolinska Institutet, Stockholm, Sweden
[4] Division of Speech, Music and Hearing, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden

silvia.capobianco@phd.unipi.it; gunnar.bjorck@regionstockholm.se; francesca.forli@unipi.it; luca.bruschini@unipi.it; a.nacci@med.unipi.it; stern@kth.se

Abstract: ***Objective:*** **to explore the use of voice mapping for assessing changes in phonatory function following injection laryngoplasty in patients with unilateral vocal fold paralysis (UVFP).** ***Materials and methods:*** **Two patient cohorts were analyzed. Cohort 1 (*N*=8) received in-office injections of hyaluronic acid or calcium hydroxylapatite, with voice recordings acquired before and immediately after treatment. Cohort 2 (*N*=4) underwent autologous fat injection under general anesthesia, with follow-ups at 1 and 3 months. All patients completed standard speech tasks with simultaneous acquisition of acoustic and electroglottographic (EGG) signals. Voice maps were computed using the FonaDyn system. Perceptual GRBAS ratings were provided by three blinded expert raters.** ***Results:*** **Voice mapping was feasible in all patients and revealed consistent treatment effects. Across both cohorts, the cycle-rate sample entropy (CSE) decreased, while the normalized peak dEGG (Q_Δ) and the Index of Contacting (I_c) both increased, indicating improved phonatory stability and vocal fold contact. Perceptual ratings showed corresponding reductions in breathiness and overall dysphonia.** ***Conclusions*****: The voice map representation clearly visualized and quantified phonatory changes post-treatment for UVFP, with potential applications in clinical monitoring and outcome evaluation.**

Keywords: **vocal fold paralysis; injection laryngoplasty; voice mapping; electroglottography; vocal fold contact**

I. INTRODUCTION

Unilateral vocal fold paralysis (UVFP) is a relatively common cause of persistent dysphonia. Its annual incidence is estimated at 3.0-3.5 per 100,000 in the general population [1], rising to 0.42% among patients who have undergone surgery under general anesthesia [2]. It usually results from recurrent laryngeal nerve injury after thyroid, thoracic, or cervical spine surgery, but may also occur in case of malignancies, trauma, or be idiopathic [3]. The impaired vocal fold shows reduced mobility and varying resting positions, leading in some cases to glottic insufficiency with a breathy, weak, and effortful voice [3].

Traditional assessment of UVFP typically combines perceptual, acoustic, aerodynamic, endoscopic, and self-reported measures, though many are designed for general dysphonia and may miss UVFP-specific deficits [1]. Perceptual tools like the GRBAS scale are widely used but inherently subjective and limited by inter-rater variability [1]. Acoustic parameters, including jitter, shimmer, harmonics-to-noise ratio, and cepstral peak prominence, are typically derived from sustained vowel phonation at a comfortable pitch and loudness. This approach suffers from intrinsic limitations, as many acoustic features are non-linearly dependent on both sound pressure level (SPL) and fundamental frequency (f_o), making comparisons across conditions and time points problematic [4]. Moreover, in pathological voices such as those affected by UVFP, compensatory strategies and unstable glottal behaviors may further confound the interpretation of traditional acoustic measures, especially if phonation is limited to a single point in the SPL–f_o space. As a result, these assessments often fail to capture the multidimensional and dynamic nature of phonatory function and may provide little insight into the physiological effects of treatment [4,5].

Voice mapping is a fairly new framework for visualizing and analyzing voice quality across a continuous range of SPL and f_o. Unlike traditional acoustic analyses that rely on elicitation of isolated vowels, voice mapping captures phonatory behavior over some relevant range, such as habitual speech. This can reveal stable and unstable regions of voice production, abrupt transitions, and patterns of irregularity across a variety of acoustic and EGG-derived parameters [4].

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

Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

This approach has demonstrated a particular strength in capturing the variability and fine structure of phonation in professional voice users and healthy individuals [6-8] but it has been applied only occasionally in patients affected by voice pathologies [5]. Importantly, voice maps often reveal abrupt and regime-shifting changes in both acoustic and EGG-derived parameters that correspond to the onset of vocal fold contact, a physiologically critical event in vocal fold biomechanics (Fig.1). For instance, the normalized peak dEGG (Q_Δ) rises sharply as soon as contact begins, while the quotient of contact by integration (Q_{ci}), the index of contacting I_c (= $Q_{ci} \times \log_{10}(Q_\Delta)$) and the cycle-rate sample entropy (CSE) of the EGG signal display steep gradients across the contact boundary, indicating a sudden transition from disordered to more stable vibration patterns [9,10]. These effects occur over narrow SPL bands, highlighting the importance of analyzing phonation across the full f_o-SPL space rather than at isolated points. This makes voice mapping a potentially valuable tool for tracking physiologically meaningful changes in glottic behavior, particularly relevant in conditions such as UVFP, where restoration of vocal fold contact is a primary therapeutic goal.

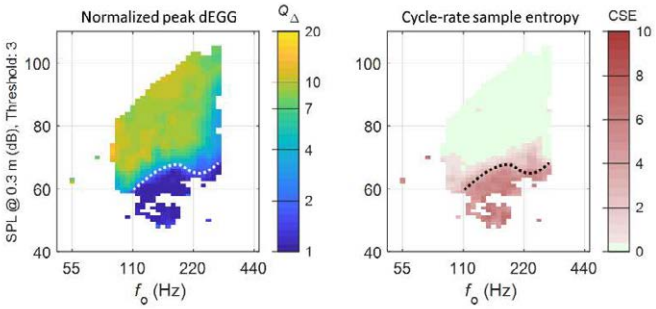


Fig.1*: Maps of the metrics Q_Δ (left) and CSE (right), illustrating EGG wave-shape variability across the voice range of a professional baritone singing in modal voice. Darker shades indicate softer contacting (Q_Δ=1 means no contact) and higher entropy (less stable contact patterns). The superimposed dashed line represents the onset of vocal fold contacting threshold, based on the Q_Δ metric. As soon as contact is established, VG vibration becomes more stable and CSE decreases.*

Despite its potential, voice mapping has rarely been applied to the study of pathological voices, and standardized protocols for its clinical implementation remain lacking. In particular, no studies to date have systematically investigated how voice maps evolve in response to therapeutic interventions aimed at restoring glottal contact. Given the abrupt acoustic and electroglottographic transitions observed at the onset of vocal fold contact in healthy voices, this method may offer unique insights into the effectiveness of medialization procedures in UVFP. In this study, we applied the voice mapping technique to a cohort of patients undergoing injection laryngoplasty for UVFP, with the aim of evaluating whether changes in voice maps reflect increased glottic contact and correlate with perceptual and instrumental clinical outcomes.

II. METHODS

Study design and participants

This observational study assessed the immediate and longer-term effects of injection laryngoplasty in patients affected by UVFP across two independent cohorts.

Cohort 1 included eight patients (3 F, 5 M; mean age 64.4 years) treated at Karolinska University Hospital (Stockholm, Sweden) with in-office injection laryngoplasty under local anesthesia (6 hyaluronic acid, 2 calcium hydroxylapatite). Six had received prior injections. Etiologies included lung disease (n=3, namely lung transplant, pleural mesothelioma and lung cancer), thyroidectomy (n=1), multiple sclerosis (n=1), vagus schwannoma (n=1), cerebellar infarction (n=1), and idiopathic paralysis (n=1).

Cohort 2 comprised four female patients (mean age: 51.5 years) undergoing first-time injection laryngoplasty with autologous fat under general anesthesia at Pisa University Hospital (Pisa, Italy). Etiologies were iatrogenic (n=2) due to thyroidectomy (n=1) and thoracic surgery (n=1), or idiopathic (n=2).

Data acquisition was performed at different time points in the two cohorts: in cohort 1, patients were recorded before (T0) and immediately after (T1a) the injection procedure; in cohort 2, recordings were collected before treatment (T0) and at one (T1b) and three months (T2b) post-intervention.

Given the differences in injection materials, surgical settings, and follow-up timelines, this preliminary study considers the two cohorts not as directly comparable, but as providing complementary insights into voice mapping after laryngoplasty.

Speech tasks and data collection

All participants performed the same standardized vocal tasks at each assessment time point: reading of a short phonetically-balanced text ("*Nordanvinden och solen*" in Swedish or "*Il Deserto*" in Italian), counting from 1 to 10, and reciting the days of the week. Each session lasted approximately 2 minutes in total. Recordings were conducted with a fixed 30 cm mouth-to-microphone distance. Audio was captured using a Line Audio OM1 microphone and Focusrite Scarlett 2i2 interface; EGG signals were recorded via the Laryngograph A100. The sample rate was 44.1 kHz and the sample width was 24 bits.

Voice mapping and outcome measures

Voice maps visualize phonatory variation across the SPL-f_o plane. They were computed using the program FonaDyn [9]. The acoustic metrics evaluated were cepstral peak prominence (CPP) and spectrum balance

(SB). The EGG metrics were Q_{Δ} , Q_{ci} , I_c , and EGG Cycle-rate Sample Entropy (CSE), being proportional to the disorder or irregularity of vocal fold contacting. A phonation detection threshold based on the autocorrelation of the acoustic signal was set to 0.8, indicating adequate periodicity and signal quality.

Additional assessments included maximum phonation time (MPT, in seconds) and perceptual evaluation using the GRBAS scale, rated by three independent, blinded experts.

Ethical approval

The study was approved by the Swedish Ethical Review Authority (Dnr 2024-00457-01) and the Ethics Committee of Tuscany Region – Northwest Area (26216_FORLI). All participants gave written informed consent. The study was conducted in accordance with the Declaration of Helsinki and Good Clinical Practice.

III. RESULTS

Immediate effect of injection laryngoplasty (Cohort 1)

Perceptual evaluation: Voice quality was assessed in both cohorts using the GRBAS scale, independently rated by three expert phoniatricians or speech-language pathologists blinded to condition and timepoint. Inter-rater agreement ranged from moderate to substantial (intra-class correlation coefficient, ICC=0.60–0.80). Immediately after injection, perceptual ratings showed a consistent reduction in breathiness (B, 5/8 pts) and asthenia (A, 4/8 pts) across the cohort. Conversely, increases in roughness (R, 4/8 pts) and strain (S, 5/8 pts) were observed in most patient. Global grade of dysphonia (G) scores improved in 5 patients, supporting an overall positive shift in perceived voice quality.

Voice mapping: Voice maps computed from simultaneous audio and EGG recordings using the FonaDyn software exhibited consistent trends immediately following injection laryngoplasty. Among the extracted parameters, spectrum balance (SB) decreased in 7 out of 8 patients across the speech range, within regions of overlap between pre- and post-injection recordings, indicating a reduction in high-frequency spectral components. EGG-based measures showed systematic changes: Q_{Δ} increased in 6 patients, suggesting stronger and more abrupt vocal fold contact; the Index of Contacting (I_c) increased in 6 patients; and cycle sample entropy (CSE), which reflects cycle-to-cycle variability in glottal contact, decreased in 6 out of 8 patients, indicating more stable phonatory patterns (Fig. 2). To quantify these trends, Wilcoxon signed-rank tests were applied to pre/post differences across all patients. The analysis revealed statistically significant changes for several metrics, particularly those associated with increased glottal contact (CSE: $p = 0.059$; Q_{Δ}: $p = 0.059$; I_c: $p = 0.059$; SB: $p = 0.008$).

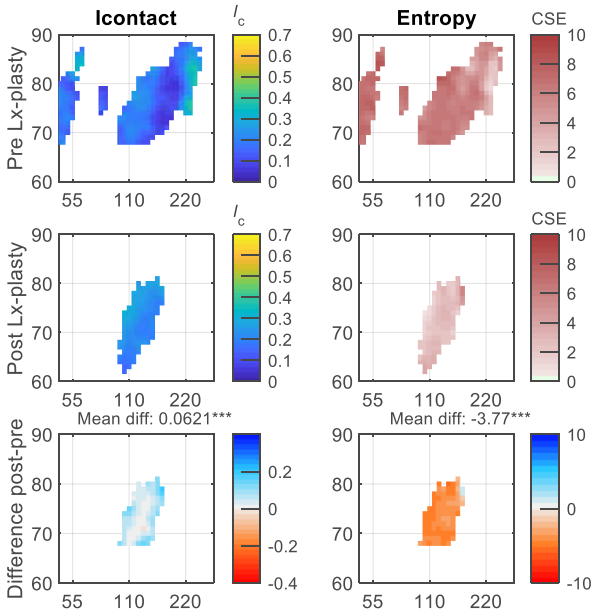


***Fig.2**: Maps of Ic (left) and CSE (right) for patient M8 before (top), and immediately after (middle) injection laryngoplasty (LxP), and the difference maps (bottom) of the overlapping cells. Ic quantifies the amount of VF contacting, while. CSE quantifies the variability in EGG wave-shape from cycle to cycle. In the overlapping f_0–SPL region, Ic increased somewhat and CSE decreased substantially, indicating increased regularity and stability of VF vibration post-injection. Note also how the creaky region below 55 Hz disappeared. The horizontal is f_0 in Hz, the vertical is SPL in dB(C) @ 0.3 m.*

Delayed effect of injection laryngoplasty (Cohort 2)

Perceptual evaluation: In the second cohort, which included four female patients treated with autologous fat injection laryngoplasty under general anesthesia, the GRBAS perceptual evaluation revealed consistent improvements at one month post-operatively (T1b). All four patients demonstrated a reduction in the overall grade of dysphonia (G) and in breathiness (B). Asthenia (A) was reduced in three patients, while roughness (R) decreased in one and increased in another. Notably, unlike the first cohort, none of the patients exhibited an increase in strain following the injection. At three months post-operatively (T2b), the perceptual profile remained largely stable, with the exception of changes in asthenia: two patients showed further reduction, likely due to the initiation of speech therapy, whereas in one patient asthenia was increased.

Voice mapping: Qualitative analysis of voice maps between T0 and T1b assessments revealed consistent trends among the four patients. The most prominent changes included an increase in CPP in three out of four patients, suggesting enhanced harmonic structure and vocal clarity, and a decrease in CSE in all four,

indicating more stable and regular vocal fold contact. At T2b these improvements were generally maintained. CPP further increased in three patients, while CSE remained stable.

IV. Discussion

The results indicate that voice mapping is feasible and informative even in severely dysphonic patients, despite the variability typically associated with pathological voices and the small sample size. Previous studies have shown that voice maps tend to be consistent within an individual but differ markedly across speakers, and this may be particularly impactful in the presence of vocal fold pathologies [4]. Nevertheless, several parameters in our study displayed coherent trends across patients, particularly those derived from the EGG signal and associated with vocal fold contact.

CSE emerged as the most consistent metric, decreasing in most patients both immediately after the injection and at follow-up recordings, indicating more stable vibratory patterns. Q_{Δ} and I_c increased in most patients after treatment, reflecting a stronger and more distinct vocal fold contact. These quantitative findings were corroborated by perceptual improvements, especially in the global rating of dysphonia and the degree of breathiness. In Cohort 1, recordings were performed immediately post-injection to capture the early effects of the procedure. Despite a mild increase in roughness, likely due to edema and temporary vibratory imbalance, voice mapping consistently revealed enhanced glottic contact, supporting the method's feasibility and sensitivity to immediate physiological changes.

When making difference maps pre-post, it is necessary to ensure overlap in the SPL–f_0 space between pre- and post-treatment recordings. In some cases, particularly in those with breathy or unstable preoperative voices, the voice range shifted considerably after treatment both in the SPL and in the f_0 domains, limiting direct comparisons. To facilitate analysis of pathological voices, we lowered the clarity threshold from the default value of 0.96 to 0.8, so as to obtain a number of observed cycles sufficient for mapping [4,9]. While this adjustment allowed map generation in dysphonic patients, it may limit inter-subject comparability and should be interpreted with caution.

Several open questions remain. Future studies should investigate which parameters are most clinically the most relevant to specific voice problems, establish effect sizes for meaningful changes, and develop user-friendly tools to support clinical interpretation. Our findings suggest that voice mapping may enrich clinical interpretation by offering a visual and objective representation of phonatory changes, with the potential to complement traditional perceptual and acoustic tools by quantifying treatment-related improvements. Such visual feedback may help clinicians not only verify the immediate success of medialization procedures but also monitor phonatory function over time. In particular, tracking specific parameters such as CSE and I_c could enable early detection of reabsorption of temporary injected materials before these changes become perceptually or endoscopically evident. This anticipatory use of voice maps might support more timely therapeutic decisions and individualized patient follow-up strategies. To support future clinical use, key aspects of the voice mapping procedure (such as metric thresholds, vocal tasks, and interpretation criteria) should be progressively standardized.

References

[1] C. Walton, P. Carding, E. Conway, K. Flanagan, and H. Blackshaw, "Voice outcome measures for adult patients with unilateral vocal fold paralysis: A systematic review," *Laryngoscope*, vol. 129, no. 1, pp. 187–197, Jan. 2019.

[2] A. W. Chen, C. H. Chen, T. M. Lin, A. C. Chang, T. P. Tsai, and S. Y. Chang, "Office-based structural autologous fat injection laryngoplasty for unilateral vocal fold paralysis," *J Clin Med*, vol. 11, no. 16, p. 4806, Aug. 2022.

[3] C. C. Wang, S. H. Wu, Y. K. Tu, W. J. Lin, and S. A. Liu, "Hyaluronic acid injection laryngoplasty for unilateral vocal fold paralysis—a systematic review and meta-analysis," *Cells*, vol. 9, no. 11, p. 2417, Nov. 2020.

[4] S. Ternström and P. Pabon, "Voice maps as a tool for understanding and dealing with variability in the voice," *Applied Sciences*, vol. 12, p. 11353, 2022.

[5] H. Cai, S. Ternström, P. Chaffanjon, and N. Henrich Bernardoni, "Effects on voice quality of thyroidectomy: A qualitative and quantitative study using voice maps," *Journal of Voice*, May 6, 2024, in press.

[6] R. R. Patel and S. Ternström, "Quantitative and qualitative electroglottographic wave shape differences in children and adults using voice map-based analysis," *Journal of Speech, Language, and Hearing Research*, vol. 64, no. 8, pp. 2977–2995, Aug. 2021.

[7] S. Ternström, S. D'Amario, and A. Selamtzis, "Effects of the lung volume on the electroglottographic waveform in trained female singers," *Journal of Voice*, vol. 34, no. 3, pp. 485.e1–485.e21, May 2020.

[8] F. M. B. Lã and S. Ternström, "Flow ball-assisted voice training: Immediate effects on vocal fold contacting," *Biomedical Signal Processing and Control*, vol. 62, p. 102064, Nov. 2020.

[9] Ternström, S. (2024). Update 3.1 to FonaDyn — a system for real-time analysis of the electroglottogram, over the voice range. *SoftwareX*, *26*, 101653.

[10] Ternström, S. (2019). Normalized time-domain parameters for electroglottographic waveforms. *The Journal of the Acoustical Society of America*, *146*(1), EL65–EL70.

MACHINE LEARNING STRATIFICATION OF PHONATION NEUROMOTOR ALTERATION IN PARKINSON'S DISEASE

A. Gómez-Rodellar[1], J. Mekyska[2], A. Álvarez-Marquina[3], D. Palacios-Alonso[4], P. Gómez-Vilda[3,4]

[1] St. Louis Missouri University, Madrid Campus, Madrid, Spain, *andres.gomez@slu.edu*
[2] Applied Neuroscience Research Group, CEITEC, Masaryk University & Brno University of Technology, Brno, Czech Republic, *mekyska@vut.cz*
[3] Universidad Politécnica de Madrid, 28220 Pozuelo de Alarcón, Madrid, Spain, *pedro.gomezv@upm.es, agustin.alvarez@upm.es*
[4] Universidad Rey Juan Carlos, Móstoles, Madrid, Spain, *daniel.palacios@urjc.es, pedro.gomezv@urjc.es*

Abstract:
The stratification of Parkinson's disease severity remains an unresolved challenge, as existing grading scales, such as UPDRS, fail to accurately capture speech-related neuromotor degeneration. Statistical Machine Learning approaches, particularly Random Forests (RF), offer a promising solution by providing an objective assessment of disease progression while delivering clinically interpretable insights that enhance symptom classification. This aligns closely with Explainable Artificial Intelligence (XAI) within the framework of clinical diagnostic decision-making. The effectiveness of RF-based stratification is validated through neuromotor competence tests and by evaluating the impact of dopaminergic treatment.

Keywords: **Larynx neuromotor activity, hypokinetic dysarthria, random forests, explainable AI.**

I. INTRODUCTION

Parkinson's disease (PD), the second most prevalent neurodegenerative disorder, is placing increasing pressure on Europe's healthcare systems, exacerbated by an aging global population. Phonation, a highly affected behavioral function in PD, offers a valuable, non-invasive, and cost-effective tool for detecting subtle neuromechanical changes linked to neurological decline. Traditional longitudinal assessments rely on subjective evaluations, primarily through the Unified Parkinson's Disease Rating Scale (UPDRS), particularly its Part III [1], which assesses neuromotor activity. To improve accuracy, researchers propose objective evaluation methods independent of subjective ratings. In this context, integrating Explainable Artificial Intelligence (XAI) with PD stratification methods could enhance phonation-based diagnostics. While Deep Learning (DL) methods are effective in AI-driven solutions, Decision Trees (DTs) offer a structured, intuitive, and clinically aligned approach to stratifying PD. Their hierarchical design naturally integrates with clinical differential diagnosis, making them a valuable tool for PD classification. The study explores the integration of laryngeal neuromechanical correlates as predictive features in DT training. Scientific evidence highlights the agonist-antagonist relationship between the cricothyroid (CT) and thyroarytenoid (TA) muscles, which have opposing effects on vocal fold dynamics. Since these muscles play a key role in modulating vocal fold tension, their inclusion in predictive modeling is highly relevant [2]:

- Cricothyroid muscle: This muscle acts as the vocal fold stretcher. When it contracts, it tilts the thyroid cartilage forward and downward, increasing the tension and length of the vocal folds.
- Thyroarytenoid muscle: This one acts like a vocal fold relaxer and thickener. It shortens and slackens the folds.

Previous research has shown that CT and thyroarytenoid TA neuromotor activity can be indirectly estimated through sustained vowel phonation [3]. The present study seeks to enhance phonation stratification using RFs, utilizing dynamic neuromotor alteration (DNMA) correlates from CT and TA activity as key predictive features.

II. METHODS

The neuromechanical activity of the CT and TA neuromotor pathways, respectively $s_{ct}(t)$ and $s_{ta}(t)$, may be indirectly estimated from the vocal fold body stiffness (VFBS) $\xi_b(t)$ by full-wave rectification after differentiation:

$$s_{ct}(t) = \frac{1}{2B_{ct}}\left(\left|\frac{\partial \xi_{bd}(t)}{\partial t}\right| - \frac{\partial \xi_{bd}(t)}{\partial t}\right);$$
$$s_{ta}(t) = \frac{1}{2B_{ta}}\left(\left|\frac{\partial \xi_{bd}(t)}{\partial t}\right| + \frac{\partial \xi_{bd}(t)}{\partial t}\right) \quad (1)$$

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

The amplitude probability distributions of $s_{ct}(t)$ and $s_{ta}(t)$, are estimated by means of normalized histograms. The whole process of feature extraction and classification depends on the following steps:

- Four-second nuclear segments of sustained utterances of vowel [a:], produced at a natural tone and comfortable effort, were extracted from each participant. The initial two seconds were omitted to eliminate vowel insertion effects.
- Adaptive inverse filtering was used to remove effects of the oro-naso-pharyngeal tract from each vowel segment to produce a glottal residual. The glottal source was estimated by integration of the glottal residual.
- A 2-mass biomechanical model of the vocal fold was used to estimate the VFBS exerted on the *musculus vocalis*. Full-wave rectification of the VBFS yielded correlates of the forces exerted by the CT (positive DNMAct) and TA (negative DNMAta), aligned with the neuromechanical model of cricothyroid articulation. The probability densities (pdfs) of the DNMAct, and DNMAta were estimated using amplitude normalized histograms.
- The study utilized three sample subsets for each participant gender set: MPD and FPD for male and female PDs, MHC and FHC for male and female healthy controls (HC), and MRS and FRS for male and female reference samples (RS). Jensen-Shannon Divergence (JSD) between each sample's pdf $p_s(\zeta)$ with respect to the centroid of the respective reference subset $p_r(\zeta)$ was estimated:

$$\Delta_{JS}\left(p_s(\zeta)\middle|p_r(\zeta)\right) = \\ = \frac{\Delta_{KL}\left(p_s(\zeta)\middle|p_a(\zeta)\right) + \Delta_{KL}\left(p_r(\zeta)\middle|p_a(\zeta)\right)}{2}; \quad (1)$$

$$p_a(\zeta) = \frac{p_s(\zeta) + p_r(\zeta)}{2}$$

where the variable ζ represents the normalized amplitude of the DNMA ($0 \leq \zeta \leq 1$) and Δ_{KL} is the Kulback-Leibler Divergence (Cover & Thomas, 2012) between $p(\zeta)$ and $q(\zeta)$:

$$\Delta_{KL}(p(\zeta)|q(\zeta)) = \int_0^\infty p(\zeta) abs\left\{log\frac{p(\zeta)}{q(\zeta)}\right\} d\zeta \quad (2)$$

- JSDs from DNMAct and DNMAta pdfs were used as predictors to feed an RF of DTs with three possible categories as target labels: P for PDs, C for HCs, and R, for RSs. Leave-one-out training was used on the whole male and female subsets to produce a trained ensamble model of the RF. The best performing DT was selected in each case to estimate the sensitivity, specificity, accuracy, and classification performance F1-score.

The inclusion of a reference set is crucial, as aging processes in some HC participants may align their phonation characteristics with those of PD participants, complicating differentiation between the two groups, as discussed in Section IV.

Table 1 Age distribution of participants

Males	MHC	MPD	MRS
No. Part.	16	16	16
mean age	65	70	44
std. age	9.7	8.0	13.2
Females	FHC	FPD	FRS
No. Part.	16	16	16
mean age	62	69	42
std. age	10.6	10.1	13.5

Sixteen samples for each gender and subset were selected from PD and HC samples recorded at St. Anne's University Hospital, Brno, in Czech Republic. RSs were collected at Hospital Universitario Gregorio Marañón in Madrid (HUGMM). A brief demographic description is given in Table 1.

The classification method by a specific DT predictive model assigns class labels to observations by recursively splitting the feature space [4]. The average accumulated mismatch (AAM), or missing rate, served as the performance measurement criterion:

$$AAM = 1 - \frac{\sum_{i=1}^{K} H(\boldsymbol{V}_{Ti}, \boldsymbol{V}_{Pi})}{K(S_P + S_C + S_R)} \quad (3)$$

Where each sample target vector $\boldsymbol{V}_{Ti}$ and predicted vector $\boldsymbol{V}_{Pi}$ are used to compute their Hamming distance $H(\boldsymbol{V}_{Ti}, \boldsymbol{V}_{Pi})$, normalized by the respective subset sizes $S_P + S_C + S_R$ and the number of folds in the cross-validation K. Standard definitions for sensitivity, specificity, accuracy, and F1-score are used as classification benchmarks [5]. Besides, individual JSDs were correlated (Pearson's) with levodopa equivalent dosage (LED), and UPDRS-III motor scale (UIII).

III. RESULTS

The trees showing the lowest missing rate have been selected as potential candidates (optimum trees) for the study of the stratification process on the male and female subsets, separately. The number of maximum splits have been fixed to 5, and all combinations have been included, given the limited size of the training subsets. The stratification consisted in classifying all the subsets with the optimum trees. The resulting trees have been plotted in Figure 1.

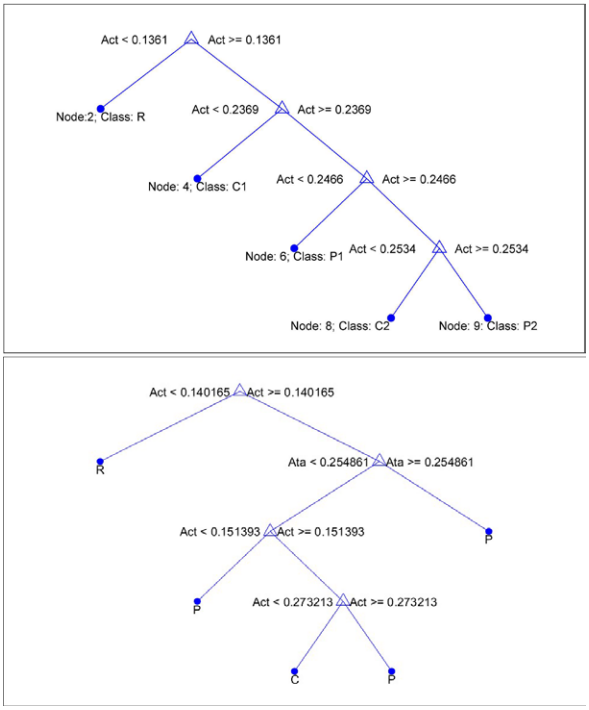


Figure 1 Optimum decision trees. Top: male subsets. Bottom: female subsets.

The performance of the optimal classification trees was evaluated based on pathological condition detection (positives) in the PD subsets and non-pathological condition detection (negatives) in the HC and RS subsets. The classification results are given in Table 2 and Table 3 for the male and female subsets.

Table 2 Male set classification performance

Missing Rate		0.1046	
	pR	**pC**	**pP**
tR	16	0	0
tC	0	13	3
tP	0	1	15
Sen	**Spe**	**Acc**	**F1**
0.94	0.91	0.92	0.88

Table 3 Female set classification performance

Missing Rate		0.1888	
	pR	**pC**	**pP**
tR	16	0	0
tC	1	14	1
tP	0	6	10
Sen	**Spe**	**Acc**	**F1**
0.63	0.97	0.85	0.74

It may be seen that both optimum decision trees separate the PD subsets onto several subsets labelled according to the leaf nodes of each tree, therefore stratifying the participant samples accordingly, as listed in Table 4 and Table 5 for the male and female subsets.

Table 4 Male PD subset stratification results. LED: Levodopa equivalent dose (LED), UPDRS-III: Unified Parkinson Disease Rating Scale Tier III.

Sample	Age	LED	UPDRS-III	Node	Global JSD
P2059	63	1340	25	C1	0.303
P2037	86	1185	36	P1	0.333
P2039	64	1058	37	P1	0.341
P2061	70	875	17	P1	0.340
P2089	75	2275	38	P1	0.336
P2113	70	750	21	P1	0.331
P2012	72	2185.5	35	P2	0.387
P2015	69	1324	22	P2	0.342
P2030	71	767	8	P2	0.369
P2034	74	870	15	P2	0.370
P2038	71	1330	55	P2	0.359
P2054	55	1836	13	P2	0.368
P2067	62	639	31	P2	0.361
P2070	59	300	18	P2	0.375
P2090	82	997.5	33	P2	0.363
P2104	69	1665	25	P2	0.381

Table 5 Female PD subset stratification results. LED: Levodopa equivalent dose (LED), UPDRS-III: Unified Parkinson Disease Rating Scale Tier III.

Sample	Age	LED	UPDRS-III	Node	Global JSD
P1006	59	875	24	P1	0.404
P1022	72	800	6	P1	0.433
P1058	71	463.75	20	P1	0.383
P1076	69	460	5	P1	0.435
P1078	83	718.75	21	P1	0.388
P1085	69	560	22	P1	0.408
P1100	84	332.5	36	P1	0.373
P1103	60	517.5	27	P1	0.429
P1027	65	740	8	P2	0.414
P1008	78	1444	23	C	0.300
P1052	49	700	33	C	0.352
P1064	60	660	11	C	0.297
P1073	65	2101.5	30	C	0.252
P1080	65	400	3	C	0.319
P1095	79	2047.5	18	C	0.262
P1097	71	1040.5	29	P3	0.372

The study explores whether the predictive power of optimal classification trees, as revealed by the JSD, could be linked to key anamnesis metadata.

Table 6 Correlation coefficients between the Global JDS, LED, and the UPDRS-III

Male subset	ρGJSD/LED	ρGJSD/UIII
P1	-0.08	0.22
P2	0.27	-0.10
Female subset	ρGJSD/LED	ρGJSD/UIII
P1	-0.29	0.55
P2, P3, C	-0.70	0.20

Specifically, it examines associations with dopaminergic medication intake (measured by the levodopa equivalent dose, LED**)** and the severity of motor symptoms (evaluated using the UPDRS-III scale)**.** These relationships are detailed in Table 6**.**

IV. DISCUSSION

The features utilized by the optimal trees for classification and stratification vary. In the male DT, only CT activity appears to have been used for stratification, whereas in the female DT, both CT and TA activities were employed. The male DT separated the reference subset (R) when the CT deviated from the reference group by less than 0.1361. A sequence of splits at thresholds 0.2369, 0.2466, and 0.2534 distinguished the first HC subset (C1), the first PD subset (P1), the second HC subset (C2), and the second PD subset (P2). In the female subsets, both CT and TA divergences were used, resulting in a distinct stratification pattern. The first split occurred at CT < 0.1402, separating the reference group (R). The second split, based on TA > 0.2549, distinguished P1. A subsequent split at CT < 0.1514 separated PD (P2), while the final split at CT < 0.2752 differentiated HC (C) from PD (PD3). The stratification in the male subset achieved a sensitivity of 0.94, specificity of 0.92, accuracy of 0.92, and an F1-score of 0.88. The classification performance for the female subsets, presented in Table 3, was slightly lower, achieving a sensitivity of 0.63, specificity of 0.97, accuracy of 0.85, and an F1-score of 0.74. The most compelling results emerged from the relationship between global divergences (global JSD) for each leaf node set and key factors such as dopaminergic medication (LED) and neuromotor alteration (UPDRS-III), as detailed in Table 6 through Pearson's correlation coefficients. While the male subset exhibited small correlations with opposite signs in both nodes, the female subset showed strong correlations across its stratification subsets—specifically, node P1 correlated with UPDRS-III, whereas subsets P2, P3, and C correlated with LED, with the latter showing a counter-correlation. These subsets were distinguished by a low TA divergence, reflecting the relaxing effect on the musculus vocalis induced by dopaminergic conditions. This finding highlights the potential of XAI-powered DTs in improving the interpretation of results within clinical contexts. Interestingly, a potential interpretation of the P1, P2, P3 groups could somehow refer to PD severity, although this question is out of the scope of the present study. An especially compelling question arises: why was the activity of the sole CT muscle sufficient to achieve stratification exclusively within the male cohort? This finding invites speculation about whether underlying neurological or laryngological factors uniquely affect male patients. Addressing these possibilities brings an open direction for future research.

V. CONCLUSIONS

The study highlights that decision trees (DTs) are effective simple tools for stratifying phonation samples affected by hypokinetic dysarthria. Using neuromotor activity in the laryngeal muscular system enables differentiation between normal and abnormal phonation, in the detection of anomalous patterns. Additionally, explainable AI (XAI) can enhance clinical interpretation of results. The findings suggest that stratification before transversal classification could improve accuracy and support longitudinal monitoring.

ACKNOWLEDGMENTS

This study was funded by project no. LX22NPO5107 (MEYS), from the European Union – Next Generation EU, by project no. CZ.02.01.01/00/23_025/0008726 (A Lifetime with Language: The Nature and Ontogeny of Linguistic Communication), co-funded by the European Union, by grant PID2023-152984OB-I00 (CaRaHAVoz) , co-funded by the Agencia Estatal de Investigación—(AEI) of Spain, and under the sponsorship of the Latin American Network of Neuroscience CYTED NT4SM (#225RT0169).

REFERENCES

[1] C. G. Goetz, et al. "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results", *Movement Disorders*, 23(15), 2129–2170. 2008. https://doi.org/10.1002/mds.22340.

[2] W. Jiang, et al., "A computational study of the influence of thyroarytenoid and cricothyroid muscle interaction on vocal fold dynamics in an MRI-based human laryngeal model", *Biomech. & Model. in Mechanobiol.*, 23, pp. 1801–1813, 2024. https://doi.org/10.1007/s10237-024-01869-9.

[3] P. Gómez, et al., "Assessing Laryngeal Neuromotor Activity from Phonation", *Int. J. of Neural Systems*, vol. 35 (6) 2550029 (19 pages), 2025, https://doi.org/10.1142/S0129065725500297.

[4] G. James, et al., *An introduction to statistical learning*, New York, Springer, 2013.

[5] R. Trevethan (2017) "Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice", *Front. Public Health*, 5:307, 2017. https//:doi.org/10.3389/fpubh.2017.00307.

SPONTANEOUS SPEECH PRODUCED BY BRAZILIAN AND PORTUGUESE COCHLEAR IMPLANT USERS

A. A. Maia[1], A. N. P. Almeida[1], A. C. A. M. Ghirardi[2] and Luis M. T. Jesus[3]

[1] Department of Speech Therapy, Health Sciences Center (CCS), Federal University of Espírito Santo (UFES), Vitória, Brazil
[2] Department of Speech Therapy, Health Sciences Center (CCS), Federal University of Santa Catarina (UFSC), Florianópolis, Brazil
[3] School of Health Sciences (ESSUA), Institute of Electronics and Informatics Engineering of Aveiro (IEETA), Intelligent Systems Associate Laboratory (LASI), Center for Languages, Literatures, and Cultures (CLLC), University of Aveiro, Portugal
andrea.maia@ufes.br, aline.n.almeida@ufes.br, carolina.ghirardi@ufsc.br, lmtj@ua.pt

Abstract: **Understanding the integration between the biofeedback provided by cochlear implant (CI) devices and speech production and perception systems, based on acoustic features of spontaneous speech, is a step forward towards better rehabilitation strategies for the hearing-impaired population. Objective: To analyse how auditory feedback in unilateral CI users influences spontaneous speech in two Portuguese language varieties; Brazilian Portuguese (BP) and European Portuguese (EP). Method: Cross-sectional, case-control speech production study of 18 postlingually deaf unilateral CI users (9 BP and 9 EP speakers). The acoustic signal for each group was annotated and analysed with Praat. The Long-Term Average Spectrum (LTAS) and Cepstral Peak Prominence (CPP) were estimated for each sentence; the fundamental frequency (f_o) and difference in amplitude between the first and second harmonics (H1-H2) were calculated for each production of vowel /a/. Conclusions: The impact of suboptimal CI feedback on CPP and f_o was similar in BP and EP, but differed in terms of the LTAS and H1-H2. The CPP could be used as a reliable cross-linguistic marker for describing voice quality and highlights convergent articulatory strategies in CI users. Furthermore, the LTAS and H1-H2 differences between BP and EP speakers may be related to their unique prosodic styles.**

Keywords: **Voice quality, acoustic phonetics, hearing loss, cochlear implant**

I. INTRODUCTION

Analysis of vocal control and speech production in cochlear implant (CI) users sheds light on the speech perception-production interaction and how auditory-motor processing is influenced by a partially restored and suboptimal auditory feedback [1]. When analysing the speech production deviations of CI users in a literature review, Ashjaei et al. [1] only reported consistent results for the fundamental frequency (f_o).

Acoustic measures provide indirect information about the configuration of the vocal tract when producing speech sounds. The auditory system helps control voice production, so adults with post-lingual hearing loss (HL), with different auditory feedback conditions, use specific strategies regarding the synchronisation of adjustments of supraglottis, glottis and general muscle tension resulting in target undershoot [1]. Therefore, an analysis of spontaneous speech samples may contribute to the description of measures that clarify the complex dynamics of physical and psychophysical attributes that characterise the human voice and its relationship with the auditory system.

The purpose of this study was to analyse how auditory feedback in unilateral CI users influences spontaneous speech in Brazilian Portuguese (BP) and European Portuguese (EP), aiming to understand source-filter interactions.

II. METHOD

This cross-sectional, case-control study acoustically describes the spontaneous speech of Brazilian and Portuguese adult unilateral CI users. Ethical approval was obtained from independent Research Ethics committees in Brazil (6.563.499) and Portugal (39-CED/2024).

Eighteen Portuguese native speakers (9 BP and 9 EP) with post-lingual HL, unilateral CI users, were recruited. The groups were balanced by age [BP – Mean (M) = 60.11 years, Standard deviation of the mean (SD) = 10.83 years; EP – M = 62.00 years, SD = 9.73 years]; sex [BP and EP – 7 women and 2 men]; audiological profile [BP and EP – 7 progressive HL and 2 HL abrupt aetiology]. The mean duration of CI use was 32.33 months for the BP speakers and 58.33 months for the EP speakers.

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

The participants' voices were recorded in a quiet room, with a Shure PG48 dynamic microphone positioned at a 45° angle from the participant's mouth and 5 cm away from the lips. The recordings were sampled at 44100 Hz with 16-bit per sample. Acoustic data were collected from the answers to the following request: Tell me about a special day for you.

The acoustic signal for each group was imported into Praat 6.4.39; a TexGrid object with two interval tiers (sentences and phones) was created and the start and end of all sentences and all the productions of vowel /a/ were manually annotated using perceptual and acoustic evidence. Only the vowels with a duration above 50 ms were annotated.

Power Spectral Density (PSD) estimates of all annotated sentences were obtained with the Long-Term Average Spectrum (LTAS) object type of Praat using the `Ltas (pitch-corrected)` function call [2].

The PSD estimates were then analysed using Functional Principal Component Analysis (FPCA) to explore variation [3] of the LTAS for the BP/ EP pair. PSD estimates were processed using a script based on two R packages developed by Happ-Kurz [4]. Both PC1 and PC2 scores (s1 and s2 shape descriptors), were used to linearly model the curves (`lm` function in R) using the following reconstruction formulas:

$$\text{predCurve}_{BP}(f) = \mu(f) + s1_{BP} \cdot PC1(f) + s2_{BP} \cdot PC2(f) \quad (1)$$

$$\text{predCurve}_{EP}(f) = \mu(f) + s1_{EP} \cdot PC1(f) + s2_{EP} \cdot PC2(f) \quad (2)$$

A purpose-built Praat script, was used to extract, the mean f_o from the annotated /a/ vowel segments.

The CPP [5] was estimated for each sentence using a plugin developed for Praat by Murray et al. [6], with the same settings described in their paper, except for the minimum and maximum CPP peak f_o search values that were set to $f_o$mean - 50 (Hz) and $f_o$mean + 50 (Hz) respectively. We calculated the CPP using a voicing activity detection algorithm [6].

The difference in amplitude between the first and second harmonics (H1-H2) was estimated for each vowel [7] using the Spectral Tilt Script for Praat developed by DiCanio [8].

Mixed effects regression models were developed using the `lmer` function from the lme4 version 1.1-37 package, with the `fo`, `CPP` and `H1-H2` as outcome variables, considering `variety_id` as a fixed effect, `speaker_id` and `sex_id` as random effects. Results from likelihood ratio tests of the models with the `variety_id` effect against the models without the `variety_id` effect were also analysed.

The regression models satisfied the normality assumption (i.e., its residuals were approximately normally distributed) and the constant variance assumption (homoscedasticity).

Praat 6.4.39 was used for signal processing and analysis; R version 4.5.1 running in RStudio Version 2024.04.2+764 were used for statistical analysis and data visualisation. The models' predictions and lines spanning the 95% confidence interval were drawn using the sjPlot 2.9.0 package.

III. RESULTS

The results reported in this section are based on the participants' spontaneous speech samples that were used to generate the LTAS and to calculate the CPP values from sentences. The /a/ vowels from the same sentences were used to estimate the f_o and H1-H2.

Functional Principal Component Analysis (FPCA) of the Long-Term Average Spectrum (LTAS)

FPCA was used to explore the main dimensions of variation of the PSD estimates of PB and EP for all sentences (shown in Fig. 1).

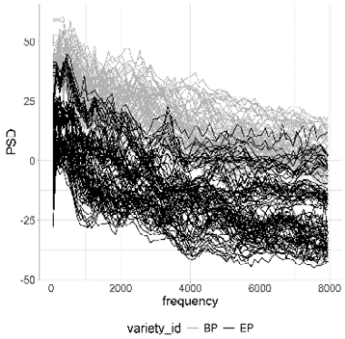


Fig.1: PSD estimates in dB (frequency in Hz).

The PC1 and PC2 components, shown in Fig. 2, had the following impact on the shape of the curves (proportion of explained variance): PC1 = 98.6 %; PC2 = 1.4 %.

The reconstructing of the curves (shown in Fig. 3) was based on s1 and s2 scores with the following standard deviations: s1 – 1535.85; s2 – 176.85. The proportions of variance explained by the regression models predicting s1 were: $R^2 = 0.658$, $p < 0.001$. The proportions of variance explained by the regression models predicting s2 were: $R^2 = 0.042$, $p = 0.007$.

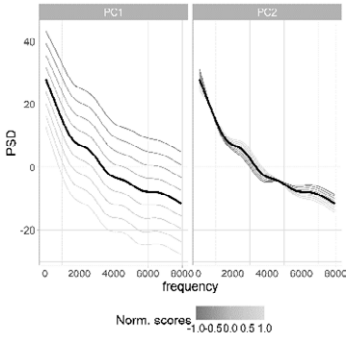


Fig. 2: Curves for components PC1 and PC2, and the effect of scores.

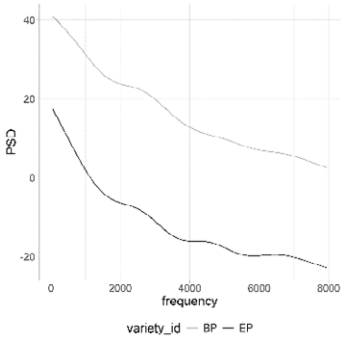


Fig. 3: Reconstructed curves.

CPP, f_o and H1-H2 values

The CPP, f_o and H1-H2 values are shown in Fig. 4.

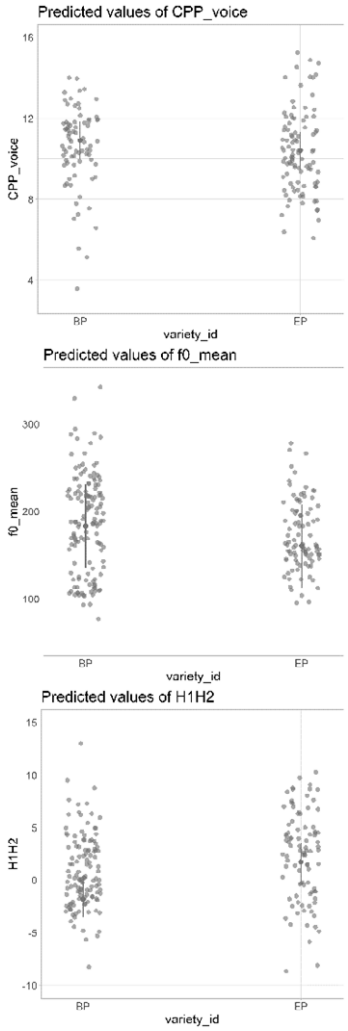


Fig. 4: CPP (dB), f_o (Hz) and H1-H2 (dB) raw values (grey dots) for the two varieties (BP and EP), model predictions and vertical lines spanning the 95% confidence intervals.

The following mixed effects regression models predicted the values shown in Fig. 4: `CPP ~ variety_id + (1|speaker_id) + (1|sex_id)`; f_o `~ variety_id + (1|speaker_id) + (1|sex_id)`; *H1-H2* `~ variety_id + (1|speaker_id) + (1|sex_id)`. Likelihood ratio tests of the models with the `variety_id` effect against the models without the `variety_id` effect only revealed a significant difference between the *H1-H2* models, i.e., there was only a significant difference between *H1-H2* values of the two varieties: CPP – $\chi^2(1) = 0.582$, $p = 0.446$; f_o – $\chi^2(1) = 1.669$, $p = 0.196$; H1-H2 – $\chi^2(1) = 6.178$, $p < 0.013$.

IV. Discussion

The analysis of spontaneous speech production in adult CI speakers of BP and EP, showed that the CPP and f_o measures were similar between the groups. The CPP values were higher than the references for neutral voice quality [9], suggesting deviations for both groups and that the CPP could be a cross-linguistic marker describing voice quality in speakers with auditory feedback partially restored by CI. The similar f_o of vowel /a/ in spontaneous speech observed for both groups can reflect Portuguese language-specific prosodic patterns [10]. The LTAS and H1-H2 captured phonetic and prosodic differences between the BP and EP varieties of Portuguese.

The pitch-corrected LTAS [2] showed significant differences between BP and EP varieties (see Fig. 3), the most noticeable difference being related to PSD levels (spectral slopes and spectral peak locations were similar). We believe these differences are related to a greater range of articulator movement and a prosody style that produces greater loudness in BP, which was corroborated by the lower H1-H2 in BP than in EP (see Fig. 4). Also, a previous study [11] suggested that cochlear implantation had positive effects on the LTAS of a voice with normal characteristics.

The absence of significant differences in CPP values between the two groups indicates that this parameter is more directly related to how voice quality is affected by the altered auditory feedback than to the differences in vocal tract use by both groups. CPP has been shown to be correlated to perceived voice quality and has proven to be a useful tool to differentiate dysphonic from non-dysphonic voices [9], [12]. In CI users, limitations in spectral resolution and auditory feedback often influence vocal control [13], which may lead to increased roughness or breathiness. However, comparable CPP values between groups can be interpreted as evidence of convergent vocal tract use strategies. Although the two varieties of Portuguese differ in phonetic and prosodic patterns, CPP values suggest that CI speakers rely on comparable articulatory and phonatory adjustments to maintain a

periodic and harmonic voice signal under conditions of altered auditory feedback. Moreover, the stability of CPP across BP and EP varieties suggests that phonatory stability was preserved independently of these prosodic differences. This convergence may reflect the physiological constraints imposed by the CI, which shape how speakers control the interaction between the source and filter, leading to strategies that are similar in both varieties of the Portuguese language. Thus, we suggest that the CPP may be a reliable tool in the clinical assessment of individuals with CI, regardless of language variety.

The f_o of vowel /a/ in spontaneous speech did not vary significantly between the BP and EP groups. These results, also corroborate the idea that the auditory mechanisms, replaced by the CI, act directly on muscle control and vocal tract adjustments, which improves the quality of speech and voice as a whole, but do not influence the sound source itself, as previously shown for the f_o values in different auditory profiles [14].

It may therefore be argued that both spectral and periodicity aspects of phonation may be less sensitive to language differences than to CI-related f_o modulations. Clinically, the f_o should be interpreted within the context of language-specific pitch norms, so as to avoid "over-pathologizing" prosodic variations, while the CPP serves as a more reliable, cross-linguistic marker of phonatory efficiency.

The auditory system assists in controlling voice production, and this study illustrates this relationship through a cross-linguistic approach, contrasting two varieties of Portuguese. This highlights the clinical importance of considering vocal patterns in the communication rehabilitation process for individuals with HL. However, to better clarify the therapeutic objectives of this process, studies that include an auditory-perceptual evaluation of voice are necessary.

V. Conclusions

When describing the spontaneous speech of Brazilian and Portuguese adult unilateral CI users, we observed that the impact of suboptimal CI feedback on voice production were similar, based on the CPP and f_o voice quality measures, and differed in terms of the energy level as reflected in the LTAS and H1-H2 values, suggesting the presence of greater phonatory effort in the Brazilian population.

This work was financially supported by CNPq - Brazil – Proc. 443150/2023-0, FCT - UIDB/00127/2020 and UID/04188/Centro de Línguas, Literaturas e Culturas.

REFERENCES

[1] S. Ashjaei, R. Behroozmand, S. Fozdar, R. Farrar, and M. Arjmandi, "Vocal control and speech production in cochlear implant listeners: A review within auditory-motor processing framework," Hear Res, vol. 453, 2024.

[2] P. Boersma and G. Kovacic, "Spectral characteristics of three styles of Croatian folk singing," J Acoust Soc Am, vol. 119, no. 3, pp. 1805-1816, 2006.

[3] J. Cronenberg, M. Gubian, J. Harrington, and H. Ruch, "A dynamic model of the change from pre- to post-aspiration in Andalusian Spanish," J Phon, vol. 83, 2020.

[4] C. Happ-Kurz, "Object-Oriented Software for Functional Data," J Stat Softw, vol. 93, no. 5, 2020.

[5] C. Watts, S. Awan, and Y. Maryn, "A Comparison of Cepstral Peak Prominence Measures from Two Acoustic Analysis Programs," Journal of Voice, vol. 31, no. 3, pp. 387.e1-387.e10, 2017.

[6] E. Murray, A. Chao, and L. Colletti, "A Practical Guide to Calculating Cepstral Peak Prominence in Praat," Journal of Voice, vol. 39, no. 2, pp. 365-370, 2025.

[7] Y. Chai and M. Garellek, "On H1-H2 as an acoustic measure of linguistic phonation type," J Acoust Soc Am, vol. 152, no. 3, pp. 1856-1870, 2022.

[8] C. DiCanio, "Spectral Tilt Script for Praat," 2012, University at Buffalo, USA. Available from https://www.acsu.buffalo.edu/~cdicanio/scripts.html

[9] O. Murton, R. Hillman, and D. Mehta, "Cepstral Peak Prominence Values for Clinical Voice Evaluation," Am J Speech Lang Pathol, vol. 29, no. 3, pp. 1596-1607, 2020.

[10] S. Frota and M. Vigário, "On the correlates of rhythmic distinctions: The European/Brazilian Portuguese case," Probus, vol. 13, no. 2, 2001.

[11] M. Yüksel and B. Gündüz, "Long-term Average Speech Spectra of Postlingual Cochlear Implant Users," Journal of Voice, vol. 33, no. 2, pp. 255.e19-255.e25, 2019.

[12] L. Brinca, A. Batista, A. Tavares, I. Gonçalves, and M. Moreno, "Use of Cepstral Analyses for Differentiating Normal from Dysphonic Voices: A Comparative Study of Connected Speech Versus Sustained Vowel in European Portuguese Female Speakers," Journal of Voice, vol. 28, no. 3, pp. 282-286, 2014.

[13] D. Horn et al., "Effects of age and hearing mechanism on spectral resolution in normal hearing and cochlear-implanted listeners," J Acoust Soc Am, vol. 141, no. 1, pp. 613-623, 2017.

[14] A. Maia, A. Almeida, A. Ghirardi, and L. Jesus, "Acoustic signal correlates of vocal quality and voice dynamics in an adult with hearing loss," in Proceedings of IberSPEECH 2024, 2024, pp. 21-25.

MU-BAND ACTIVITY DESYNCHRONIZATION BEHAVIOR FOUND IN PHONATION FROM TWO CASES OF AUTISM SPECTRUM DISORDER

A. Gómez-Rodellar[1], M. Jodra-Chuan[2,3], P. Gómez-Vilda[4,5]

[1] St. Louis Missouri University, Madrid Campus, Madrid, Spain, *andres.gomez@slu.edu*
[2] Department of Personality, Assessment and Clinical Psychology, Universidad Complutense de Madrid, Madrid, Spain, *majodra@ucm.es*
[3] Asociación Nuevo Horizonte, Las Rozas de Madrid, 28231 Madrid, Spain
[4] Universidad Politécnica de Madrid, 28220 Pozuelo de Alarcón, Madrid, Spain, *pedro.gomezv@upm.es*
[5] Universidad Rey Juan Carlos, Móstoles, Madrid, Spain, *pedro.gomezv@urjc.es*

Abstract:

This study explored potential behavioral indicators of phonation in two individuals with Autism Spectrum Disorder (ASD), focusing on desynchronization signatures derived from EEG-band analysis. Longitudinal data revealed significant differences in the synchronization index (SI) during attentive vocalization, with normotypical controls exhibiting marked desynchronization, most notably among males. While the majority of phonations fell within central quartiles, select outliers demonstrated heightened desynchronization. Correlational findings linked increased desynchronization to greater pitch variability, along with higher kurtosis in males and greater skewness in females, indicating distinctive phonation qualities across cases. The contrast between the male's imitative speech and the female's spontaneous vocalizations underscores the difficulty of standardizing protocols, reinforcing the importance of tailored, empathetic approaches to vocal data collection in ASD research.

Keywords: **Speech Analysis; Autistic Spectrum Disorder; μ-Band Desynchronization, Longitudinal Phonation Assessment.**

I. INTRODUCTION

Autism Spectrum Disorder (ASD) encompasses a broad range of neurocognitive differences whose causes remain elusive, making its study vital for understanding brain diversity and tailoring interventions that enhance individual lives. ASD presents high variability in behavior and communication—some individuals are highly articulate while others rely on alternative methods—complicating the use of speech as a diagnostic tool [1]. Challenges in social communication, sensory processing, and language development point to complex neural mechanisms, often reflected in unique phonation patterns [2]. As cross-sectional studies fall short of capturing this complexity, longitudinal approaches—like the six-year analysis of phonation and EEG-related vocal correlates featured in this study—offer deeper insights into the neurological and expressive dynamics of ASD, reinforcing the importance of personalized therapeutic strategies and continued research into neurodivergent conditions [3].

The EEG μ-band (typically 8–12 Hz) has been extensively studied in relation to autism spectrum disorder (ASD), and findings consistently point to alterations in its activity that may reflect underlying neurophysiological disruptions. Many studies report decreased μ-band power in individuals with ASD, especially during resting-state EEG. This reduction is often interpreted as a sign of impaired cortical inhibition, possibly linked to GABAergic dysfunction [4]. ASD EEG profiles often show excess power in low (ϑ) and high (β-γ) frequencies, with reduced μ in the middle, forming a "U-shaped" pattern [5]. Mu rhythms, neural oscillations linked to sensorimotor activity, play a crucial role in integrating perception and action, particularly through the transformation of sensory input into motor output [6]. Their suppression during action execution or observation reflects activity in the mirror neuron system (MNS), which is essential for understanding others' actions and imitation learning. In individuals with Autism Spectrum Disorder (ASD), μ suppression often appears weakened during observed actions but remains intact during self-initiated movements, suggesting selective disruptions in social-motor resonance. EEG studies indicate that this phenomenon is frequency-specific and influenced by developmental factors, with certain sub-bands (e.g., 10–13 Hz) more sensitive to suppression effects than others. Rather than implying a wholly dysfunctional MNS in ASD, these findings underscore complex neural variability and highlight the relevance of alternative networks and top-down modulation pathways in social cognition [7]. Moreover, the distinction between synchronization at rest and desynchronization during motor tasks—core conditions for μ suppression—suggests that a whole-brain EEG approach focusing on both μ and α rhythms may offer deeper insights into ASD-related neural dynamics [8].

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

This study tracks the vocal progression of two individuals with ASD from Asociación Nuevo Horizonte (ANH: Las Rozas de Madrid, Spain), using semiannual recordings of sustained [a:] vowels from 2019 to 2025. EEG-based biomarkers from the laryngeal neuromotor system; ϑ, μ, and α bands were extracted via phonation inversion to build a comparative dashboard against sixteen neurotypical reference subjects. The ultimate aim of this study is to assess the extent to which EEG-related activity, indirectly inferred from neuromotor phonation correlates, reveals the presence of U-shaped profiles or low/high frequency band ratios in abnormal power patterns associated with autism spectrum disorder (ASD). Such findings may substantiate the potential utility of voice recordings in the prodromal and longitudinal monitoring of ASD.

II. METHODS

Two participants diagnosed with ASD, namely a 45-year-old male and a 40-year-old female at the start of the study, were continuously monitored through recordings of sustained vowel utterances ([a:]) between 2019 and 2025. Their profiles summarized in Table 1 were carefully documented as part of a structured research program overseen by specialized professionals within ANH. All recordings were conducted under consent from legal representatives in strict adherence to the Helsinki protocol, ensuring ethical issues.

Table 1 Characteristics of participants; F: female; M: Male; CARS: childhood autism rating scale (non autistic condition <30, moderate 30-36, severe: >36); DEX: Dysexecutive Questionnaire (optimum condition <10, normal condition 10-18, moderate dysexecutive 19-28, severe cognitive alteration >28)

Gender	F	M
Age	45	40
Diagnosis	ASD	ASD
CARS	30	30
DEX	45	44
Comorbid.	Intellec. Disab.	Sev. Intellec. Dis.

Audio was recorded using a Sennheiser SK 300 G2A lavalier mic (15 cm from mouth), linked to an EM 300 G2 receiver and Motu Traveller interface on a portable PC. Speech was captured at 44.1 kHz/16-bit in uncompressed WAV format. Sessions occurred in a quiet, comfortable room to reduce ASD-related stress, with each lasting 3–4 minutes to avoid fatigue. Caregivers played a key role in easing the process and enhancing participant cooperation. Besides, two subsets of vowel [a:] samples from sixteen speakers of each gender drawn from Hospital Universitario Gregorio Marañón in Madrid (HUGMM) were used for the reference set, as given in Table 2. The study was conducted on segments of 1 s (excluding onset and decay) were extracted from the reference database; for ASD recordings, selections ranged from 600 ms to 1 s.

Table 2 Age distribution of normotypical subsets.

Male subset	MRS
No. Participants	16
mean age	44
std. age	13.2
Female subset	FRS
No. Participants	16
mean age	42
std. age	13.5

Adaptive inverse filtering was used to remove effects of the oro-naso-pharyngeal tract (ONPT) from each vowel segment produce a glottal residual. The glottal residual $\varepsilon_g(t)$ was estimated from the phonation signal $s_p(t)$ by the inversion of the vocal tract, and the glottal source $u_g(t)$ was estimated integrating the glottal residual $\varepsilon_g(t)$. A 2-mass biomechanical model of the vocal fold was used to estimate the vocal fold body stress (VFBS) $\xi_b(t)$ as a result of the inversion of the biomechanical model on the glottal source $u_g(t)$. The EEG-band components of the VFBS were the results of band-filtering and normalizing the VFBS $\mathcal{F}_\Omega\{\xi_b(t)\}$ for the respective EEG frequency bands (Ω) defined as:

$$\begin{aligned}
\xi_\delta(t) &= \mathcal{F}_{\Omega_\delta}\{\xi_b(t)\}/\bar{\xi}_b\,;0<\Omega_\delta\leq 4\,Hz\\
\xi_\vartheta(t) &= \mathcal{F}_{\Omega_\vartheta}\{\xi_b(t)\}/\bar{\xi}_b\,;4<\Omega_\vartheta\leq 8\,Hz\\
\xi_\alpha(t) &= \mathcal{F}_{\Omega_\alpha}\{\xi_b(t)\}/\bar{\xi}_b\,;8<\Omega_\alpha\leq 16\,Hz\\
\xi_\beta(t) &= \mathcal{F}_{\Omega_\beta}\{\xi_b(t)\}/\bar{\xi}_b\,;16<\Omega_\beta\leq 32\,Hz\\
\xi_\gamma(t) &= \mathcal{F}_{\Omega_\gamma}\{\xi_b(t)\}/\bar{\xi}_b\,;32<\Omega_\gamma\leq 64\,Hz\\
\xi_\mu(t) &= \mathcal{F}_{\Omega_\mu}\{\xi_b(t)\}/\bar{\xi}_b\,;8<\Omega_\mu\leq 12\,Hz
\end{aligned} \tag{1}$$

where $\bar{\xi_b}$ was the average of the VFBS. In EEG studies, definitions of the α band vary across publications, typically limited to 8–12 Hz, whereas this manuscript adopts a broader range of 8–16 Hz, overlapping with the μ band, also defined as 8–12 Hz. This overlap arises from the distinct scalp regions each band covers: the sensorimotor areas for μ and the occipital zone for α. To resolve ambiguity in voice-based EEG extraction, the μ band is defined as 8–12 Hz, while α spans 8–16 Hz, enabling a decoupling of high-α activity ($12<\Omega_{\alpha h}\leq 16\,Hz$) from μ rhythms on time-averaged analysis:

$$\bar{\xi}_{\alpha h} = 2\,\bar{\xi}_\alpha/\bar{\xi}_\vartheta - \bar{\xi}_\mu/\bar{\xi}_\vartheta \tag{2}$$

The synchronization index (SI) was estimated from the average VFBS on the ϑ, μ, and high-α bands:

$$SI = \left(\bar{\xi}_\mu - \left(\bar{\xi}_\vartheta + \bar{\xi}_{\alpha h}\right)/2\right)/\bar{\xi}_\vartheta \tag{3}$$

The expected behavior of the SI will show the following patterns

$$\begin{aligned} &\text{DC1: } \bar{\xi}_{\mu} < \bar{\xi}_{\vartheta};\ \bar{\xi}_{\mu} < \bar{\xi}_{\alpha h};\\ &\text{DC2: } \bar{\xi}_{\mu} < \bar{\xi}_{\vartheta};\ \bar{\xi}_{\mu} > \bar{\xi}_{\alpha h};\\ &\text{SC1: } \bar{\xi}_{\mu} > \bar{\xi}_{\vartheta};\ \bar{\xi}_{\mu} > \bar{\xi}_{\alpha h};\\ &\text{SC2: } \bar{\xi}_{\mu} > \bar{\xi}_{\vartheta};\ \bar{\xi}_{\mu} < \bar{\xi}_{\alpha h}; \end{aligned} \tag{4}$$

DC1 and DC2 representing desynchronization states, with mid μ band activity falling below the low-frequency ϑ band; DC1 aligns with the U-shaped pattern. Conversely, SC1 and SC2 represent synchronization states, μ activity exceeding ϑ levels.

III. RESULTS

The longitudinal evolution of SI from the longitudinal sample collection described before on the male and female participants have been plotted in Figure 1.

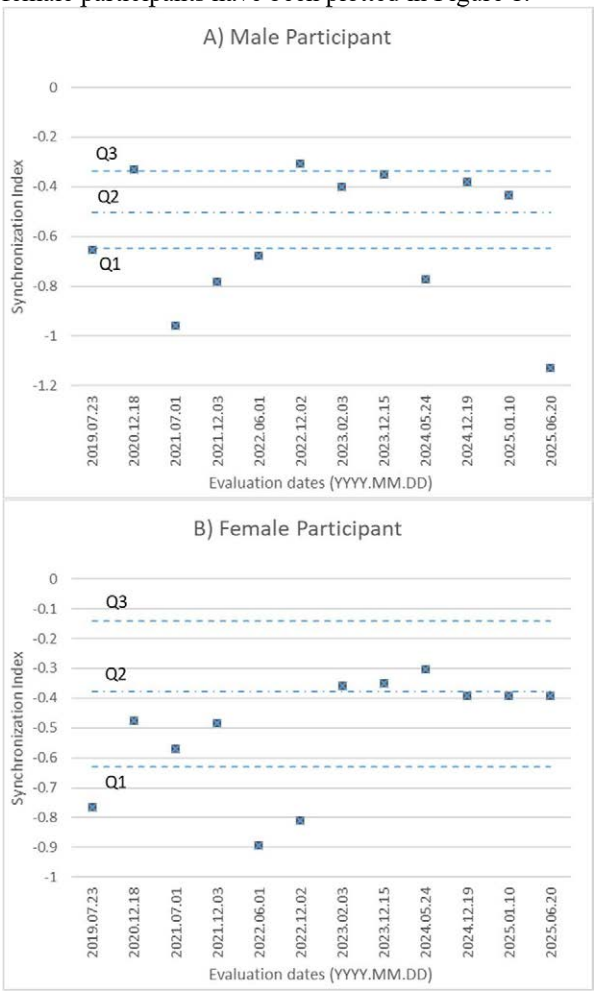


Figure 1 Values of SI from the evaluation process. The median (Q2) of the respective reference distributions is given by the dash-dot line (−·−). The upper and lower dash lines (− −) represent the respective first and third quartiles (Q1 and Q3) of the reference distributions.

A central question is whether the Synchronization Index (SI) accurately reflects the degree of μ band desynchronization triggered by intense vocal attention, specifically, whether focusing on phonation can alter vocal quality beyond baseline levels. To explore this, correlations between the SI and key voice parameters were examined, including fundamental frequency (f_0) and cepstral peak prominence (CPP). This relationship examines correlations between the degree of conscious attention from phonation stability represented by fluctuations in f_0, and harmonic/inharmonic ratio stability represented by fluctuations in the CPP.

Table 3 Pearson correlation between the SI and descriptive features of voice quality indices from f0 and cpp; mean: average; std: standard deviation; skw: skewness; kur: kurtosis; Q2: median; IQ: inter-quartile interval. Relevant values are given in bold.

Correlation	Male part.	Female part.
$\rho SI/f0_{mean}$	-0.154	0.135
$\rho SI/f0_{std}$	**-0.362**	**-0.361**
$\rho SI/f0_{skw}$	0.191	**0.486**
$\rho SI/f0_{kur}$	**-0.314**	-0.029
$\rho SI/f0_{Q2}$	-0.174	0.064
$\rho SI/f0_{IQ}$	-0.204	**-0.418**
$\rho SI/cpp_{mean}$	-0.259	0.051
$\rho SI/cpp_{std}$	0.001	0.133
$\rho SI/cpp_{skw}$	-0.061	**-0.358**
$\rho SI/cpp_{kur}$	**-0.375**	0.260
$\rho SI/cpp_{Q2}$	-0.163	0.060
$\rho SI/cpp_{IQ}$	0.247	**0.578**

IV. DISCUSSION

The longitudinal evolution in Figure 1 reveals that SI in neurotypical samples shows negative values, linked to DC1 (U-shape), suggesting desynchronization during attentive phonation. Distributions are nearly symmetrical, with Q2 slightly lower in males than females. Most phonations cluster within the Q1–Q3 range, except for a few outliers, four in males, three in females, indicating intensified attentive desynchronization.

Regarding the results associating the synchronization index and the phonation quality in terms of f_0 stability and harmonic/inharmonic contents, the results reveal consistent patterns linking synchronization index (SI) and phonation quality:

- f_0 stability: Both participants show moderate negative correlation between SI and f_0 standard deviation (ρ = –0.362 male, –0.361 female), indicating greater desynchronization relates to increased pitch variability. This trend is reinforced in the female's f_0 interquartile range (ρ = –0.418).
- f_0 distribution dhape: Higher synchronization corresponds to greater skewness in pitch (ρ =

0.486), notably in the female, as well as increased desynchronization correlates with elevated kurtosis in males (ρ = –0.314).

- CPP Measures: Female CPP skew grows with desynchronization (ρ = –0.358), and male CPP kurtosis rises similarly (ρ = –0.375). Both show positive correlation between lower desynchronization and larger CPP interquartile spread (ρ = 0.578), strongest in the female.

These trends suggest that synchronization levels influence vocal signal regularity and distribution shape, with clearer effects in female participants.

V. CONCLUSIONS

The study concentrated on pointing out some possible behavioral indicators from the phonation of two cases of ASD affected people potential desynchronization hallmarks derived from the EEG-band description of phonation. In this sense, it must be concluded that longitudinal data revealed notable differences in the synchronization index (SI) during attentive phonation, with normotypical samples showing desynchronization, especially pronounced in males. Most phonations clustered within central quartiles, though a few outliers displayed intensified desynchronization. Correlational analysis showed that greater desynchronization was associated with increased pitch variability, kurtosis (in males), and skewness (in females), highlighting differences in phonation quality across participants. This variability, expressed by the contrast between the male's imitative speech and the female's spontaneous vocalizations, illustrates the challenges in applying uniform protocols and emphasizes the merit of careful, individualized sample collection in ASD research. These results highlight the potential of EEG-derived synchronization indices during phonation as objective biomarkers for ASD, revealing sex-specific differences in vocal and neural patterns that challenge uniform assessment protocols. The associations between desynchronization and acoustic features like pitch variability emphasize the need for individualized, longitudinal data to capture ASD's heterogeneity. Clinically, this supports advancing tailored diagnostic and therapeutic strategies that integrate neurophysiological and acoustic measures, enhancing precision in assessment and intervention. However, caution is necessary given the complexity of the method and the statistically underpowered sample size. To take these preliminary findings into consideration, the approach should be validated on larger cohorts.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the sponsorship of the Latin American Network of Neuroscience CYTED NT4SM (#225RT0169), and the support provided by Asociación Nuevo Horizonte, on their continued encouragement and trust in our work. We are also grateful to the hosting provided by Universidad de Las Palmas de Gran Canaria to the Neurolinguistic Mindfulness Initiative supporting this research (https://neuminnet.ulpgc.es/).

REFERENCES

[1] Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR). American Psychiatric Association, 2022. https://doi.org/10.1176/appi.books.9780890425787.

[2] K. Scaler Scott, J. A. Tetnowski, J. Flaitz, J. S. Yaruss, "Preliminary study of disfluency in school-age children with autism", *International Journal of Language and Communication Disorders*, 2014, 49(1), 75-89. https://doi.org/10.1111/1460-6984.12048.

[3] F. Töpfer, U. Wiesing, "The medical theory of Richard Koch I: theory of science and ethics", *Medicine, Health Care and Philosophy*, 2005, 8, 207-219. https://doi.org/10.1007/s11019-004-7445-5.

[4] M. Milovanovic, R. Grujicic R, "Electroencephalography in Assessment of Autism Spectrum Disorders: A Review", *Front. Psychiatry* 12, art. 686021. https//doi.org/10.3389/fpsyt.2021.686021.

[5] J. Wang, J. Barstein, L. E. Ethridge, M. W. Mosconi, Y. Takarae, J. A. Sweeney, "Resting state EEG abnormalities in autism spectrum disorders", J. of Neurodev. Dis., 2013, 5 (24). https://doi.org/10.1186/1866-1955-5-24.

[6] J. A. Pineda, "The functional significance of mu rhythms: translating "seeing" and "hearing" into "doing"". *Brain research reviews*, 2005, 50 (1), 57-68. https://doi.org/10.1016/j.brainresrev.2005.04.005.

[7] A. K. Lockhart, C. F. Sharpley, V. Bitsika, "Mu Desynchronisation in Autistic Individuals: What We Know and What We Need to Know", Rev. J. Autism Dev. Disord., 2024 11, 595–606. https://doi.org/10.1007/s40489-023-00354-w.

[8] G. Dumas, R. Soussignan, L. Hugueville, J. Martinerie, J. Nadel, "Revisiting mu suppression in autism spectrum disorder", *Brain Research*, 2014, 1585, 108-119. https://doi.org/10.1016/j.brainres.2014.08.035.

SPECIAL SESSION I
THE ROLE OF AI IN THE FIELD OF PHONIATRICS AND LARYNGOLOGY: PRESENT AND FUTURE

Organized by G. Cantarella

EVALUATING STATE OF THE ART VOICE CONVERSION MODELS FOR DYSPHONIC AND ELECTRO-LARYNX SPEECH

Benedikt Mayrhofer[1,3], Martin Hagmüller[1,3], Franz Pernkopf[1], Philipp Aichinger[2,3]
[1]Signal Processing and Speech Communication Laboratory, Graz University of Technology
[2]Department of Otorhinolaryngology, Div. Phoniatrics-Logopedics, Medical University of Vienna
[3]Comprehensive Centre for AI in Medicine, Medical University of Vienna
{benedikt.mayrhofer, hagmueller, pernkopf}@tugraz.at, philipp.aichinger@meduniwien.ac.at

Pathological speech, caused by dysphonia or produced via electro-larynx devices, often suffers from poor intelligibility and unnatural prosody. In this paper, we investigate the potential of four state-of-the-art voice conversion models: FreeVC, QuickVC, LLVC, and XVC for restoring healthy-sounding speech. All models are fine-tuned on Austrian-German datasets and evaluated using objective and subjective metrics. Results show substantial gains in intelligibility, naturalness, and perceived vocal health. QuickVC, FreeVC, and XVC perform similarly and achieve the highest preference scores, exceeding unprocessed pathological speech by up to 200%. These findings highlight the potential to improve communication for individuals with voice disorders and motivate further development of efficient, high-quality conversion systems.
Keywords: **voice conversion, speech rehabilitation, pathological speech, electro-larynx, machine learning**

I. INTRODUCTION

Voice has an important role in human communication. Beside its functional use, it is connected with our identity, shaping how we perceive ourselves and how others perceive us [1]. Voice disorders, such as dysphonia, significantly impact individuals' ability to communicate, leading to reduced social relationships, feelings of isolation, and psychological burdens [2]. In severe cases, such as those caused by laryngeal cancer, patients may require a laryngectomy, after which electro-larynx (EL) devices are commonly employed, producing speech that is characterized by limited prosodic expressiveness, and degraded intelligibility [3].

With the rapid advancement of speech and machine learning technologies, artificial intelligence is a promising approach for improving voice quality of impaired speech [4]. Voice conversion (VC) refers to the process of modifying a speaker's voice to resemble that of another target speaker while preserving the original linguistic content [5].

In this paper, VC is investigated as a potential method for enhancing speech quality of individuals with voice disorders. The novelty lies in testing and evaluating state-of-the-art VC models for converting both dysphonic and electro-larynx generated speech into healthy-sounding voices. The experiments focus on fine-tuning the models on Austrian-German datasets and assessing their performance through objective and subjective evaluations.

II. RESOURCES

Four VC models were selected: FreeVC[1] [5], QuickVC[2] [6], LLVC[3] [7], and XVC[4] [8]. All models are built upon the Generative Adversarial Network (GAN) framework [9], which enables direct waveform synthesis. FreeVC is designed for high-quality VC, while QuickVC emphasizes computational efficiency. LLVC is further optimized for low-latency, real-time conversion. XVC, on the other hand, focuses on cross-lingual VC, enabling speech conversion across different languages. FreeVC, QuickVC, and XVC are similar in their architecture and were chosen because they represent the state-of-the-art in general VC technology, are publicly available, and strong results have been reported. LLVC, on the other hand, was selected for its unique ability to perform real-time voice conversion on consumer-grade CPUs, enabling an assessment of low-resource performance in comparison to larger models. Unlike the other models, which rely on disentangling speaker identity from linguistic content, LLVC applies a conversion mask to the acoustic feature vectors extracted from the source speech, in order to simulate target speaker characteristics.

A. Database

The database covers a wide range of Austrian-German speech in different conditions, including healthy voices [10], pathologies [11], [12], and speech recorded with electro-larynx [13] devices. It also includes audio recordings provided by the Medical University of Vienna, where parts are non-public due to data protection regulations.

The final database has a total duration of 10.5 hours. Dysphonic speech accounts for 26.87%, electro-larynx recordings for 47.6%, and healthy speech for 25.90%.

[1]https://github.com/OlaWod/FreeVC
[2]https://github.com/quickvc/QuickVC-VoiceConversion
[3]https://github.com/KoeAI/LLVC
[4]https://github.com/ConsistencyVC/ConsistencyVC-voive-conversion

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

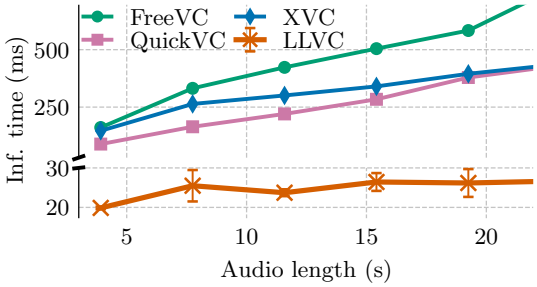


Fig. 1: Mean inference (Inf.) time of the models vs. input audio length (on NVIDIA Geforce RTX 3050 Laptop GPU). 99% confidence interval (CI) shown for LLVC.

The use of the audio recordings for evaluation and presentation purposes has been approved by the Institutional Review Board (IRB) of the Medical University of Vienna (number 2005/2022).

B. Computational resources

The models differ significantly in their computational requirements and sizes. LLVC has 3.3M parameters, requiring 38 MB of storage. In contrast, FreeVC, QuickVC, and XVC range from 39.3M to 41.3M parameters, requiring 450–473 MB of storage.

To evaluate computational performance, inference times were measured for 6.520 VCs per model (Fig. 1). LLVC is the only model supporting true streaming; the others (FreeVC, QuickVC, and XVC) require complete audio phrase inputs. For comparability, LLVC was also evaluated using complete audio phrase inputs. It shows the lowest inference time, i.e. under 30 ms even for inputs exceeding 20 seconds. QuickVC and XVC stay below 200 ms for shorter utterances but scale with audio length. FreeVC exhibits the highest inference times.

III. TRAINING

All models were initially pre-trained on English data. The pre-trained model-weights are fine-tuned on subsets of the database.

Pathological-to-normal VC ideally requires parallel data to align pathological content with healthy speaker characteristics. However, FreeVC, QuickVC, and XVC are designed for non-parallel tasks [5], [6], [8], and incorporating pathological speech during training degrades their performance by encouraging the adoption of pathological traits. To avoid this, they are fine-tuned exclusively on healthy speech, while pathological data is reserved for evaluation. This approach focuses fine-tuning on an unseen language and promotes generation of high-quality, healthy-sounding speech.

In contrast, LLVC requires parallel data for training and is limited to converting source speech into a single target voice [7]. To address this, FreeVC is used as a teacher model to convert all input samples into a consistent healthy-sounding voice, which is then used to train LLVC. This teacher-student strategy allows LLVC to approximate FreeVC's output while using its own architecture. Five separate fine-tuning runs are performed, each producing a distinct LLVC model configured for one male or female target voice.

The database is split 80/20 into training and testing set. All models are trained using the original hyperparameters. FreeVC, QuickVC, and XVC requires 20k-30k steps (batch size 64). LLVC, is trained for 100k steps (batch size 9). Training is performed on an NVIDIA A40 GPU.

IV. EVALUATION

An online survey was done to compare synthesized speech with the dysphonic and electro-larynx speech. Participants rated intelligibility, naturalness, perceived healthiness, rhythm, intonation and preference.

Participants were required to be healthy, at least 18 years old, and free from significant voice or hearing impairments. The survey was distributed among students, colleagues in speech or audio fields, and the international research community.

A. Test design

Ten random recordings from pathological datasets are processed by each model, generating 170 samples (10 $\times$ 4 target voices $\times$ 4 models + 10 originals). Two female and two male target voices are used (same- and cross-gender); one reflects a pre-illness voice, the others serve as representative healthy exemplars. Participants rate a randomized subset to ensure broad coverage while keeping the survey duration manageable. Tasks include:

- **Demographic information:** Participants provide sociodemographic data and confirm eligibility.
- **Intelligibility:** Participants orthographically transcribe unique audio samples. The Word Error Rate (WER) is calculated to assess intelligibility. Lower WER indicates better intelligibility.
- **Quality rating:** Using a 1–5 MOS scale, participants rate naturalness, healthiness, and rhythm & intonation; 1 = worst, 5 = best.
- **Preference rating:** Participants rank pathological and synthesized samples by preference using a 0–100 visual scale (equal ratings allowed); 0 = very unpleasant, 100 = very pleasant.

Participants are randomly assigned to Group A or B. Group A rates phrases 1–5 for intelligibility and 6–10 for quality; Group B does the opposite. This prevents overlap of phrases across tasks.

V. RESULTS

93 participants (Mean age = 36.4 years, SD = 16.6) take part; 52.4% are female, 47.8% male. Most (89.2%) are native German speakers. Mild hearing problems are reported for 9.7% (e.g., mild tinnitus, age-related loss).

TABLE I: MOS with (±) 95% CI for pathological (Path.) speech, including dysphonic (Dysph.) and electro-larynx (E-Larynx), and for conversions considering three quality measures. Best scores are marked **bold**▲, worst with ▼.

Model	Naturalness (non-synthetic)		Healthiness		Rhythm & Intonation	
	Dysph.	E-larynx	Dysph.	E-larynx	Dysph.	E-larynx
Path.	3.73 ± 0.29	1.00 ± 0.00▼	1.95 ± 0.23▼	1.49 ± 0.51▼	3.62 ± 0.27	1.98 ± 0.53
FreeVC	3.80 ± 0.25	**2.33 ± 0.77**▲	3.70 ± 0.27	2.33 ± 0.67	3.60 ± 0.25	**2.46 ± 0.78**▲
QuickVC	3.91 ± 0.27	2.16 ± 0.52	**4.15 ± 0.24**▲	**2.90 ± 0.54**▲	3.57 ± 0.28	1.89 ± 0.39
LLVC	2.35 ± 0.30▼	1.50 ± 0.44	2.38 ± 0.29	1.90 ± 0.45	3.15 ± 0.28▼	1.50 ± 0.28▼
XVC	**3.99 ± 0.24**▲	**2.36 ± 0.54**▲	3.83 ± 0.28	2.15 ± 0.53	**3.77 ± 0.28**▲	2.06 ± 0.43

A. *Intelligibility*

Dysphonic and electro-larynx recordings achieve mean WERs of 10% and 25%, respectivley (Fig. 2). FreeVC slightly increases WER for dysphonic speech (13%) but reduces it for electro-larynx (17%). QuickVC increases WERs to 23% and 40%. LLVC performs worst (34% and 40%), while XVC achieves the lowest WERs: 9% (dysphonic) and 20% (electro-larynx), indicating improvements.

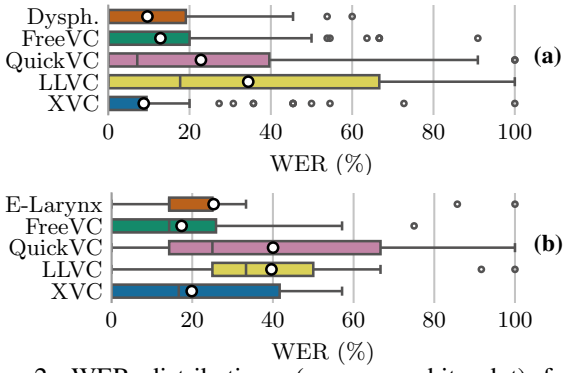


Fig. 2: WER distributions (mean = white dot) for **(a)** dysphonic, **(b)** electro-larynx, and converted speech

B. *Quality rating*

TABLE I shows MOS ratings for naturalness, healthiness and rhythm & intonation. Dysphonic speech has moderate scores (naturalness: 3.73, healthiness: 1.95, rhythm & intonation: 3.62), while electro-larynx samples score consistently low (naturalness: 1.00 healthiness: 1.49, rhythm & intonation: 1.98). For dysphonic input, XVC achieves the highest naturalness (3.99) and rhythm & intonation (3.77), while QuickVC leads in healthiness (4.15). LLVC performs worst across all quality measures. For electro-larynx speech, QuickVC leads in healthiness (2.90), FreeVC and XVC show similar results, and LLVC scores lowest.

C. *Preference rating*

Preference ratings (Fig. 3) confirm a clear improvement of VC models over pathological speech. All three top-performing models (FreeVC, QuickVC, XVC) show preference gains across all severity classes of dysphonic speech (mild: 61.3, moderate: 25.3, severe: 22.7), with the largest improvement gains observed in moderate cases (e.g., FreeVC: 74.5, QuickVC: 71.94, XVC: 71.2). Mild cases yield the smallest gain but highest score (FreeVC: 83.6, QuickVC: 77.6, XVC: 77.7), reflecting their high baseline ratings and limited room for improvement. For severe cases, while improvements remain substantial (FreeVC: 52.44, QuickVC: 61.8, XVC: 57.9), the scores do not reach moderate or mild qualities, likely due to heavily distorted inputs compromising the models' conversion capabilities. For electro-larynx speech, all models achieve high comparable gains with lower total scores (FreeVC: 36.7, QuickVC: 42.8, XVC: 42.3), due to the low baseline.

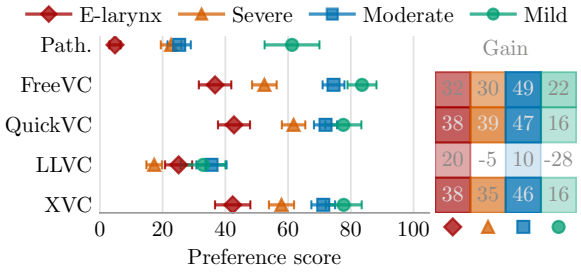


Fig. 3: Preference scores (mean ± 95% CI), dysphonic speech grouped by hoarseness severity (severe, moderate, mild) and electro-larynx speech. The heatmap shows absolute gains over pathological speech.

VI. DISCUSSION

Across all metrics VC models consistently improve speech quality of both dysphonic and electro-larynx inputs. However, the effectiveness is higher for dysphonic speech. WER results (Fig. 2) show that selected models preserve intelligibility for dysphonic speech and even improve it for electro-larynx. Quality ratings (TABLE I) show strong gains in healthiness without compromising naturalness or prosody. Preference ratings (Fig. 3) confirm these trends by achieving consistent improvements across all severity levels of dysphonic speech.

Spectral analyses further substantiate these findings. Dysphonic input (Fig. 4, top), often characterized by high noise levels, slow speaking rates and heavy breathing, lacks clear formants and F0 contours. The conversion restores both, resulting in spectra qualitatively similar to healthy speech. However, speaking rate remains unchanged, which may explain why severely

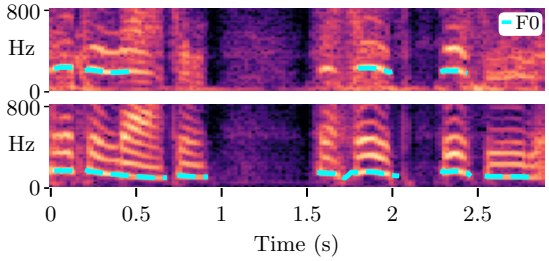


Fig. 4: Mel-spectrogram including $F0$ contour for unprocessed dysphonic speech (top), and converted speech using FreeVC (bottom).

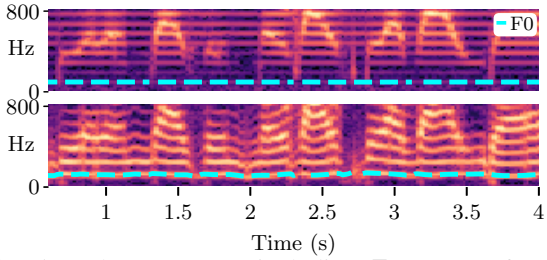


Fig. 5: Mel-spectrogram including $F0$ contour for unprocessed electro-larynx speech (top), and converted speech using FreeVC (bottom).

impaired samples yield lower scores compared to moderate or mild ones, as they exhibit slower speaking rates and heavier breathing. Similarly, original electro-larynx speech (Fig. 5, top) shows strong excitation artifacts and a flat F0 contour, which produces monotonous and inexpressive speech. Although the conversion significantly reduces these artifacts and reconstructs plausible formant patterns, the F0 remains flat, suggesting that monotonic excitation limits prosody recovery. Addressing this requires explicit modeling of prosody, through other training strategies or F0 modulation.

LLVC performs worst across metrics due to its lightweight architecture and use of a conversion mask instead of content-speaker disentanglement. This limits its ability to model fine-grained acoustics and suppress artifacts, leading to low synthesis quality. This is evident from TABLE I, where LLVC got low naturalness scores and Fig. 3, where LLVC is rated worst regardless of the input severity. This shows the need for future research into efficient yet high-quality VC models.

The present study is limited to German speakers. XVC, being cross-lingual, is expected to generalize well to other languages, whereas FreeVC and QuickVC would likely benefit from language-specific fine-tuning.

VII. CONCLUSION

This paper explores the potential of VC technologies (FreeVC, QuickVC, LLVC, XVC) to enhance dysphonic and electro-larynx generated speech. FreeVC, QuickVC, and XVC improved preference and MOS ratings, with FreeVC and XVC also enhancing intelligibility. LLVC showed no meaningful improvements. The effectiveness is notably higher for dysphonic inputs, especially at moderate and mild severity levels. Electro-larynx speech presents greater challenges due to its constant excitation signal. While models restore formants and F0 contours, temporal characteristics like speaking rate and prosodic expressiveness remain limited in severe cases. LLVC's under-performance highlights the limitations of lightweight architectures but also emphasize the need for future research into more resource-efficient VC models.

VIII. ACKNOWLEDGMENTS

We thank our late colleague Matthias Metelka for his key role in the survey design, and all participants. This research was funded in part by the Austrian Science Fund (FWF) [10.55776/PAT5948223] and utilized the Austrian Scientific Computing (ASC) infrastructure. ChatGPT (OpenAI) was used for proofreading.

REFERENCES

[1] M. Tiwari and M. Tiwari, "Voice - How humans communicate?" *J. Nat. Sci. Biol. Med.*, vol. 3, no. 1, pp. 3–11, Jan. 2012.

[2] S. M. Cohen, W. D. Dupont *et al.*, "Quality-of-Life Impact of Non-Neoplastic Voice Disorders: A Meta-Analysis," *Ann. Otol. Rhinol. Laryngol.*, vol. 115, no. 2, pp. 128–134, Feb. 2006.

[3] A. K. Fuchs, M. Hagmüller *et al.*, "The New Bionic Electro-Larynx Speech System," *IEEE J. Sel. Top. Sig. Process.*, vol. 10, no. 5, pp. 952–961, Aug. 2016.

[4] A. Bhardwaj *et al.*, "Transforming pediatric speech and language disorder diagnosis and therapy: The evolving role of artificial intelligence," *Health Sci. Rev.*, vol. 12, no. 100188, pp. 2772–6320, Sep. 2024.

[5] J. Li, W. Tu *et al.*, "FreeVC: Towards High-Quality Text-Free One-Shot Voice Conversion," in *Proc. ICASSP*, Rhodes Island, Greece, June 2023, pp. 1–5.

[6] H. Guo, C. Liu *et al.*, "QUICKVC: A Lightweight VITS-Based Any-to-Many Voice Conversion Model using ISTFT for Faster Conversion," in *Proc. ASRU*, Taipei, Taiwan, Dec. 2023, pp. 1–7.

[7] K. Sadov, M. Hutter *et al.*, "Low-latency Real-time Voice Conversion on CPU," *arXiv*, 2023.

[8] H. Guo, C. Liu *et al.*, "Using Joint Training Speaker Encoder With Consistency Loss to Achieve Cross-Lingual Voice Conversion and Expressive Voice Conversion," in *Proc. ASRU*, Taipei, Taiwan, Dec. 2023, pp. 1–8.

[9] M. Krichen, "Generative Adversarial Networks," in *Proc. ICCCNT*, Delhi, India, July 2023, pp. 1–7.

[10] B. Schuppler, M. Hagmüller *et al.*, "GRASS: the Graz corpus of Read And Spontaneous Speech," in *Proc. LREC*, Reykjavik, Iceland, May 2014, pp. 1465–1470.

[11] T. Nawka and L. C. Anders, "Die auditive Bewertung heiserer Stimmen nach dem RBH-System," *Thieme*, 1996.

[12] W. J. Barry and M. Pützer, "Saarbruecken voice database," 2007, Institute of Phonetics, Saarland University.

[13] A. K. Fuchs, J. A. Morales-Cordovilla *et al.*, "ASR for Electro-Laryngeal Speech," in *Proc. ASRU*, 2013, pp. 234–238.

THE ROLE OF ARTIFICIAL INTELLIGENCE IN LARYNGOLOGY: CURRENT EVIDENCE AND FUTURE PERSPECTIVES

E. Bellini[1,2], C. Sampieri[3,4,5], F. Mora[1,2], G. Peretti[1,2]

[1] Unit of Otorhinolaryngology – Head and Neck Surgery, IRCCS Ospedale Policlinico San Martino, Genoa, Italy
[2] Department of Surgical Science (DISC), University of Genoa, School of Medicine, Genoa, Italy
[3] Department of Otorhinolaryngology, Hospital Clinic, Barcelona, Spain
[4] Head and Neck Cancer Unit, Hospital Clinic, Barcelona, Spain
[5] Department of Experimental Medicine (DIMES), University of Genoa, School of Medicine, Genoa, Italy
Email: e.e.elisabellini@gmail.com

Abstract: **Artificial intelligence (AI) is rapidly transforming the field of laryngology, particularly in the context of endoscopic examination of the upper aerodigestive tract (UADT). AI-based computer-aided diagnosis (CADx) systems can achieve diagnostic accuracy comparable to expert laryngologists, with applications spanning lesion classification, boundary segmentation, real-time informative frame selection, and automated detection of malignancies. Beyond diagnostic improvements, AI promises to enhance workflow efficiency, democratize access to care in resource limited settings, and support medical education. This paper summarizes the current evidence and outlines the opportunities and challenges in integrating AI into clinical laryngology.**

Keywords: Artificial Intelligence; Laryngology; Endoscopy; Deep Learning; Head and Neck Cancer.

I. INTRODUCTION

Laryngeal lesions constitute the most commonly encountered pathology within the upper aerodigestive tract (UADT) [1]. Endoscopic assessment of the UADT is fundamental for both the detection and longitudinal follow-up of such lesions. Precise identification of pathological changes, together with reliable discrimination between benign and malignant entities, remains crucial. Epidemiological evidence suggests that a variable proportion of premalignant UADT lesions progress to invasive carcinoma. Accordingly, early and accurate differentiation between benign and potentially malignant lesions is essential. Laryngoscopy, through High-Definition Videolaryngoscopy (HD-VLS) continues to represent the standard diagnostic modality for the evaluation of laryngeal lesions. The detection and characterization of laryngeal lesions have been the focus of scientific investigation for over a century, with the earliest [2-3] systematic descriptions dating back to the nineteenth century. Despite technological progress with high definition imaging and narrow-band imaging (NBI) [4-5], interpretation remains highly operator-dependent, with requiring a substantial learning curve and is hindered by the absence of standardized quality metrics. In this context, artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), emerges as a promising tool, offering automated, objective, and reproducible approaches to overcome these limitations[6]. Recent translational projects, including AIRCARE, have demonstrated the feasibility of AI integration into point-of-care diagnostics, combining real-time analysis, multimodal data, and cloud-based platforms to enhance clinical decision-making and accessibility [6].

II. METHODS

This work reviews recent multicenter validation studies, systematic reviews, and experimental applications of AI in laryngology. We analyzed CADx systems for lesion classification, segmentation models for tumor delineation, and real-time CADe tools in videoendoscopy. Specific attention was given to dataset characteristics, training methodologies, and cross institutional validation. Studies implementing data harmonization and bias mitigation, such as AIRCARE's federated training approach, were emphasized as examples of clinically scalable solutions.

III. RESULTS

AI models trained on large, annotated datasets achieved diagnostic accuracies between 85–92%, comparable to expert laryngologists. Segmentation frameworks such as SegMENT-Plus, U-Net derivatives, and transformer-based architectures successfully delineated tumor boundaries in both WL and NBI modalities, reaching real-time inference speeds (~25–30 FPS) [7-9]. Hybrid AI pipelines integrating CADe and CADx modules have improved detection of premalignant lesions and automatic identification of diagnostically relevant frames. Additionally, deep super-resolution YOLO variants enhanced visualization of microvascular patterns, aiding early dysplasia detection [10]. Despite these advances, challenges persist— dataset bias, limited generalizability, and model overfitting remain

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

key obstacles requiring systematic solutions through multicenter data sharing and standardization [11].

IV. Discussion

The clinical implementation of AI in laryngology offers opportunities to standardize diagnostic workflows, reduce interobserver variability, and expand access to high-quality care. AI-assisted workflows can be integrated into clinical routines, from image acquisition to diagnostic interpretation and follow-up, improving efficiency and consistency [12-13]. For instance, AIRCARE's approach demonstrates how portable AI-powered endoscopic units can provide real-time lesion scoring and remote follow-up via cloud platforms—bridging care disparities between urban and rural centers. These innovations also hold educational value, offering trainees objective feedback and supporting evidence-based learning. However, most current models rely on retrospective datasets and may suffer from overfitting, bias, or lack of external validation. Dataset diversity, representativeness, and labeling quality are crucial to ensure generalizable AI. Ethical and regulatory frameworks must evolve to address data privacy, algorithm transparency, and validation standards. Furthermore, interdisciplinary collaboration—linking clinicians, data scientist. [14-15]

V. Conclusion

Artificial intelligence is rapidly reshaping clinical laryngology through its potential to enhance diagnostic precision, efficiency, and accessibility. Multicenter initiatives such as AIRCARE illustrate a tangible roadmap for clinical translation, combining AI, robotics, and telemedicine into an integrated ecosystem for head and neck care. Future work must focus on prospective validation, regulatory approval, and continuous evaluation to ensure safe, equitable, and effective AI deployment across healthcare systems.

REFERENCES

[1] Wilmes CMLH, BSc AG, Marres HAM, Wellenstein DJ, van den Broek GB. A Systematic Review of the Clinical Impact of Implementing Artificial Intelligence in Upper Aerodigestive Tract Endoscopy. Head Neck. 2025 Jun 18. doi: 10.1002/hed.28213. Epub ahead of print. PMID: 40530669.

[2] Green H. On the Surgical Treatment of Polypi of the Larynx, and Oedema of the Glottis. Vol xi. G.P. Putnam; 1852:9-124.)(Garcia M. Observations on the human voice. Proc R Soc Lond. 1855;7:399-408

[3] Sawashima M, Hirose H. New laryngoscopic technique by use of fiber optics. J Acoust Soc Am. 1968;43(1):168-169.) (Aviv JE, Takoudes TG, Ma G, Close LG. Office-based esophagoscopy: a preliminary report. Otolaryngol Head Neck Surg. 2001;125(3):170175.

[4] Piazza C, Cocco D, De Benedetto L, Del Bon F, Nicolai P, Peretti G. Narrow Band Imaging and High Definition Television in the Assessment of Laryngeal Cancer: A Prospective Study on 279 Patients. Eur Arch Oto-Rhino-Laryngol 2009 2673 (2010) 267:409–14. doi: 10.1007/S00405-009-1121-6

[5] Vilaseca I, Valls-Mateus M, Nogués A, Lehrer E, López-Chacón M, Avilés-Jurado FX, et al. Usefulness of Office Examination With Narrow Band Imaging for the Diagnosis of Head and Neck Squamous Cell Carcinoma and Follow-Up of Premalignant Lesions. Head Neck (2017) 39:1854–63. doi: 10.1002/hed.24849

[6] Dunham ME, Kong KA, McWhorter AJ, Adkins LK. Optical Biopsy: Automated Classification of Airway Endoscopic Findings Using a Convolutional Neural Network. Laryngoscope (2022) 132 Suppl:S1–8. doi: 10.1002/lary.28708

25. Fehling MK, Grosch F, Schuster ME, Schick B, Lohscheller J. Fully Automatic Segmentation of Glottis and Vocal Folds in Endoscopic Laryngeal High-Speed Videos Using a Deep Convolutional LSTM Network. PloS One (2020) 15:1– 29. doi: 10.1371/journal.pone.0227791

[7] C.M.L.H. Wilmes, A. Goril, H.A.M. Marres, D.J. Wellenstein, G.B. van den Broek, "A systematic review of the clinical impact of implementing artificial intelligence in upper aerodigestive tract endoscopy," Head & Neck, 2025; 1–21. doi:10.1002/hed.28213.

[8] C. Sampieri, F. Mora, G. Peretti, et al., "Multicenter clinical validation of an artificial intelligence diagnostic classification model for laryngoscopy images," Otolaryngology–Head and Neck Surgery, 2025.

[9] C. Sampieri, C. Baldini, M.A. Azam, et al., "Artificial intelligence for upper aerodigestive tract endoscopy and laryngoscopy: a guide for physicians and state-of-the-art review," Otolaryngology–Head and Neck Surgery, 2023; 1–19. doi:10.1002/ohn.343.

[10] C. Sampieri, M.A. Azam, A. Ioppi, et al., "Realtime laryngeal cancer boundaries delineation on white light and narrow-band imaging laryngoscopy with deep learning," Laryngoscope, 2024; 00:1–9. doi:10.1002/lary.31255.

[11] C. Baldini, M.A. Azam, C. Sampieri, et al., "An automated approach for real-time informative frames classification in laryngeal endoscopy using deep learning," Eur Arch Otorhinolaryngol, 2024. doi:10.1007/s00405-024-08676-z.

[12] C. Baldini, L. Migliorelli, D. Berardini, et al., "Improving real-time detection of laryngeal lesions in endoscopic images using a decoupled super-resolution enhanced YOLO," Comput Methods Programs Biomed, 2025; 260:108539. doi:10.1016/j.cmpb.2024.108539.

[13] C. Sampieri, G. Peretti, “Democratizing cancer detection: artificial intelligence-enhanced endoscopy could address global disparities in head and neck cancer outcomes,” Eur Arch Otorhinolaryngol, 2025. doi:10.1007/s00405-025-09257-4.
[14] Kang YF, Yang L, Xu K, Hu BB, Cai LJ, Liu YH, Lu X. A lightweight intelligent laryngeal cancer detection system for rural areas. Am J Otolaryngol. 2024 Nov-Dec;45(6):104474. doi: 10.1016/j.amjoto.2024.104474. Epub 2024 Aug 8. PMID: 39137696
[15] Yao P, Usman M, Chen YH, German A, Andreadis K, Mages K, et al. Applications of Artificial Intelligence to Office Laryngoscopy: A Scoping Review. Laryngoscope (2021) 00:1–24. doi: 10.1002/lary.29886.

VOCAL BIOMARKERS OF DYSPHONIA AND THEIR INTERPRETABILITY: CHALLENGES FOR AN UNDERSTANDABLE AI

Federico Calà[1]

[1] Department of Information Engineering, Università degli Studi di Firenze, Firenze, Italy

federico.cala@unifi.it

Abstract: **Acoustic recordings of vocal emissions are multidimensional signals that can describe human health with a cost-effective and contactless approach. Their analysis relies on the extraction of diverse features across time, frequency, cepstral, and nonlinear dynamic domains, each capturing complementary aspects of voice production. These descriptors, ranging from perturbation measures to cepstral parameters, were used as inputs for artificial intelligence (AI) models to distinguish healthy from pathological voices or predict clinical severity scores. Despite achieving high accuracies in experimental studies, such systems remain underutilized in clinical practice, largely due to limited interpretability. Black-box predictions do not meet clinicians' needs for detailed, physiologically meaningful information to guide decisions. To address this gap, explainable AI (XAI) strategies have been also applied in acoustic analysis to define pathological voice properties. These methods provide novel and quantitative insights into which acoustic features define models' predictions, supporting the decision-making process and the usage of such systems. These efforts are leading to the affirmation of a new branch, called audiomics, that focuses not only predictive performance but also on interpretability, bridging computational results with clinical expertise to foster future applicability.**

Keywords: voice pathology, feature extraction, explainable AI, audiomics

I. INTRODUCTION

The human voice is a complex biosignal that reflects both physiological mechanisms and pathological alterations of the phonatory apparatus. Both can be descripted through acoustic feature extraction, i.e., the transformation of raw audio recordings into a set of numerical features that retain the most important information. Different analytical domains capture complementary aspects of vocal production [1, 2]. Time-domain features provide metrics related to speech timing and rhythm, such as vocal emission duration, voiced unit length, and pauses. These indicators often reflect breath control, vocal effort, or speech fluency. Frequency-domain parameters quantify the harmonic structure of the voice, with parameters like fundamental frequency (F0), jitter, shimmer (i.e., F0 perturbations in frequency and amplitude, respectively), and formant frequencies (F1-F3). These descriptors are sensitive to vocal fold vibrations and resonances occurring in the fixed and variable shape cavities located along the vocal tract (e.g., pharynx, oral and nasal cavities). This domain also includes measures computing the ratio between the harmonics and the background noise in speech, such as the harmonics-to-noise ratio (HNR) and the normalized noise energy (NNE). Cepstral features (e.g., Cepstral Peak Prominence, CPP) are obtained from the inverse Fourier transform of the acoustic signal's logarithmic spectrum [3]. Several studies highlighted their strong correlations with perceptual assessments of dysphonia, specifically breathiness and hoarseness. Cepstrum has been also extensively used to calculate the Mel-Frequency Cepstral Coefficients (MFCCs), i.e., metrics that exploit the knowledge of human auditory principles and the decorrelating property of the cepstrum to decompose acoustic signals with a filter-bank approach [1, 2]. Such features also present several advantageous properties, including its appropriateness for analysing both sustained vowels and running speech, and the ability to characterise speech signals without relying on the F0 estimation [2]. Nonlinear dynamic features share similar benefits with the latter, and they also account for well-known nonlinearities that in speech production (e.g., air turbulence, rheological characteristics of the vocal fold tissue, and asymmetry in left and right vocal fold movements and collision) [4]. Among these, three categories may be identified:

1) Measures describing the state space geometrical properties and trajectories (e.g., fractal dimensions and the Largest Lyapunov Exponent). Recurrence quantification analysis investigates how trajectories return to specific regions of the state space [5]. They require voice dynamics to be purely deterministic, not accounting for physiological (and pathological) randomness.

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

Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

2) Information theory measures do not make assumption on the nature of the signal. Entropy parameters (e.g., Approximate, Sample Fuzzy) fall into this category.
3) Self-similarity characteristics as detrended fluctuations analysis (DFA) quantify the self-affinity of an acoustic signal as an alternative to entropy to study chaotic vibrations.

Notably, these features assume that the scale invariance does not depend on time and space. However, such variations do usually occur in voice and speech. Hence, multiscale approaches have been proposed to better examine the multiple components, interactions, and scales that the voice production system involves over time.

Taken together, such parameters provide a multidimensional description of voice that serves as the input for advanced computational models.

II. ARTIFICIAL INTELLIGENCE IN THE DETECTION AND ASSESSMENT OF DYSPHONIA

The integration of artificial intelligence (AI) with acoustic analysis has markedly advanced voice pathology research. AI models are trained to recognise hidden patterns within acoustic biomarkers and generate accurate predictions about vocal health.

Two widely used examples of predictive strategies are regression and classification. The first estimates continuous values, for example predicting a patient's score on perceptual scales such as GRB (Grade, Roughness, Breathiness) [6]. The latter aims at distinguishing categories, such as differentiating healthy controls from patients, as well as separating individual pathologies between each other.

These models can be trained under supervised learning, where labelled data guide the learning process, or unsupervised one, where hidden clusters and structures emerge without predefined categories.

Studies combining acoustic features with AI have achieved high detection accuracies (often above 90%) in identifying dysphonia [7, 8, 9]. However, despite these promising results, such tools are rarely adopted in routine clinical practice. Common limitations concern data quality, such as numerosity and balance between classes, absence of comorbidities that may worsen the vocal output and algorithm design. However, the main obstacle that hinder the wide applicability of such models is represented by the lack of the interpretability of the AI outcome.

III. AUDIOMICS AND MODEL INTERPRETABILITY

AI systems in healthcare are often described as "black boxes", since only their inputs and outputs are visible while the internal reasoning remains opaque. Highly complex models may achieve strong predictive performance, but their decisions are difficult for clinicians to interpret and trust. From an ethical and practical perspective, medical AI is therefore required to be explainable enough to allow physicians to identify potential errors and contest system decisions. Transparency and interpretability are essential prerequisites for introducing AI into clinical practice, since clinicians must understand the motives behind predictions to reliably integrate them into care [10]. Indeed, a simple prediction (whether an observation belongs to one class or another) is not sufficient to be accepted as clinically relevant [10, 11, 12]. Hence, to be considered practical tools and to avoid limiting their effectiveness, explainable AI (XAI) has been proposed to allow field experts to understand and manage AI model development and results.

Furthermore, XAI holds significant promise for addressing the limitations inherent to acoustic voice analysis. Traditional measures, such as jitter and shimmer, are highly sensitive to external factors including background noise, microphone quality, room acoustics, and inconsistencies in speech tasks [13]. Moreover, speaker variability, arising from age, gender, emotional state, fatigue, health status, and dialect, introduces further ambiguity, making it challenging to distinguish pathological from non-pathological voices. XAI can enhance transparency by identifying which features contribute most to diagnostic decisions and under what conditions these features remain robust. This could allow clinicians to better understand when an acoustic parameter is reflecting pathology versus being confounded by extraneous factors, improving the use of AI-assisted voice diagnostics.

In radiomics, medical images are classified after being decomposed into critical regions and described by several features, concerning shape, intensity and texture, to support diagnostic decisions. By analogy, the field of audiomics has emerged, aiming to bring the same paradigm to acoustic analysis [14]. Audiomics emphasizes not just prediction accuracy, but also interpretability, enabling models to highlight the acoustic parameters most responsible for their output. Moreover, it aims at associating widely used parameters that lack a direct link with physiological processes, such as the MFCC, more understandable themselves. Thus, considering the intended use of computer-aided diagnostic systems, explainability relies also on the meaning extracted features. Specifically, it would be important to, on the one hand, to know which acoustic features were considered most relevant for a classifier for the distinction of two or more classes and, on the other hand, that such metrics could have a reliable physiological meaning [15]. One XAI strategy corresponds to post-hoc techniques that

provide text, visual or local explanations for interpretability. Here, Shapley values are among the most common implemented methodologies. They are a game theory concept that calculates for each feature its marginal impact exerted on the model prediction [16]. They were successfully used to understand which acoustic parameters were mostly relevant in distinguishing healthy controls from subject with Alzheimer's disease [17], Parkison's disease [16, 18], multiple sclerosis [16], functional speech disorder [19]. Shapley values were also used comparing pre- and post-treatment voices in detecting robust biomarkers of vocal improvements in post-thyroidectomy disorder [20] and in predicting voice-related quality of life scores [21]. They have been implemented also to separate patients diagnosed with polyps and unilateral vocal fold paralysis (UVFP) [22]. Another approach consists of the Local-Interpretable Model-agnostic Explanation. This technique explains the individual predictions of any machine learning model by approximating the latter with a simple, local, and understandable model. It was adopted to evaluate the relationships between perceptual and acoustic features of spasmodic dysphonia [23]. Moreover, when using ensemble models (e.g., AdaBoost, XGBoost), interpretability can be actuated through feature importance. This can be computed by means of diverse procedures, e.g., that measure the decrease of accuracy of the forest when a variable is randomly permuted or the decrease of impurity of a nodes where the given variable is used for splitting [15]. Such techniques were deployed for understanding the separation between patients diagnosed with benign lesions of the vocal folds and UVFP, as well as pre- and post-treatment patients [15], and between healthy controls and individuals with asthma [24] and with several voice disorders (considered as a whole) [25].

Research has only recently started to properly address the explainability gap, however, these promising results suggest that XAI can foster trust, usability, and achieve medical relevance, allowing interpretability to transform classifiers from black boxes into decision-support instruments that augment, rather than replace, clinical judgment.

IV. NEW PERSPECTIVES ON DEEP LEARNING

Deep learning (DL) is a subfield of machine learning that uses artificial neural networks with multiple layers to learn patterns from vast amounts of data, mimicking how the human brain processes information. Their high performance has recently led to a greater usage and popularity of these models, also in laryngology [26], however, their interpretability remains a critical limitation and has been only addressed in the last few years [27]. XAI in DL can be performed with adequately adapted version of the methods illustrated in Section III. For instance, Shapley values were used to distinguish an altered articulation (i.e., dysarthria) between healthy controls and people with cerebral palsy and amyotrophic lateral sclerosis [28].

However, DL offered the possibility to develop new methods as the Gradient-weighted Class Activation Mapping (Grad-CAM), which employs the gradient of the feature classification score from a convolutional layer [29], or Occlusion Sensitivity maps [12]. They were successfully employed into understanding which regions of a long-term mel spectrogram were more useful for the model to distinguish between healthy controls, mild and severe Parkinson's patients [30], as well as normophonic and pathological voices [12, 31].

V. CONCLUSION

AI-driven acoustic analysis offers the potentiality to detect subtle, multidimensional biomarkers of voice pathology with accuracy often surpassing traditional perceptual assessments. Yet, clinical integration requires more than raw performance. Similarly to radiologists, laryngologists and speech therapists need smart tools that explain thoroughly their reasoning, linking acoustic properties to underlying pathophysiology.

Audiomics represents this new frontier. By combining advanced feature extraction, robust AI algorithms, and interpretable models, it can provide clinicians with a transparent, data-driven lens on vocal health. Such systems will not only improve diagnostic confidence and patient monitoring but also pave the way for personalized interventions in the broader healthcare ecosystem.

REFERENCES

[1] Fagherazzi, G., … & Despotovic, V. (2021). Voice for health: the use of vocal biomarkers from research to clinical practice. *Digit biomark*, *5*(1), 78-88.

[2] Gómez-García, J. A., Moro-Velázquez, L., & Godino-Llorente, J. I. (2019). On the design of automatic voice condition analysis systems. Part II: Review of speaker recognition techniques and study on the effects of different variability factors. *Biomed Signal Process Control*, *48*, 128-143.

[3] Yadav, S., … & Meena, P. (2023). A review of feature extraction and classification techniques in speech recognition. *SN Computer Science*, *4*(6), 777.

[4] Calà, F., … & Lanata, A. (2025). On the complexity matching and multiscale nonlinear perspective of voice restoration via fat injection laryngoplasty in unilateral vocal fold paralysis. *Sci Rep*, *15*(1), 31801.

[5] Little, M., … & Moroz, I. (2007). Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *Nat Preced*, 1-1.

[6] Gómez-García, J. A., … & Godino-Llorente, J. I. (2019). Emulating the perceptual capabilities of a human evaluator to map the GRB scale for the assessment of voice disorders. *Eng Appl Artif Intell*, *82*, 236-251.

[7] Al-Nasheri, A., … & Ibrahim, M. F. (2017). Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions. *IEEE Access*, *6*, 6961-6974.

[8] Islam, R., Abdel-Raheem, E., & Tarique, M. (2022). Voice pathology detection using convolutional neural networks with electroglottographic (EGG) and speech signals. *Comput Methods Programs Biomed Update*, *2*, 100074.

[9] Verma, V., ... & Chui, K. T. (2023). A novel hybrid model integrating MFCC and acoustic parameters for voice disorder detection. *Sci Rep*, *13*(1), 22719.

[10] Lauritsen, S. M., ... & Thiesson, B. (2020). Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nat commun*, *11*(1), 3852.

[11] Xu, H., & Shuttleworth, K. M. J. (2024). Medical artificial intelligence and the black box problem: a view based on the ethical principle of "do no harm". *Intell Med*, *4*(1), 52-57.

[12] Özcan, F. (2025). Differentiability of voice disorders through explainable AI. *Sci Rep*, *15*(1), 18250.

[13] Bottalico, P., ... & Rubin, A. D. (2020). Reproducibility of voice parameters: The effect of room acoustics and microphones. *J Voice*, *34*(3), 320-334.

[14] Bensoussan, Y., Elemento, O., & Rameau, A. (2024). Voice as an AI biomarker of health—introducing audiomics. *JAMA Otolaryngol Head Neck Surg*, *150*(4), 283-284.

[15] Calà, F., … & Lanata, A. (2025). Towards an explainable Artificial intelligence system for voice pathology identification and post-treatment characterisation. *Biomed Signal Process Control*, *104*, 107530.

[16] Vizza, P., ... & Veltri, P. (2025). Through the Speech and Vocal Signals Hidden Secrets: An Explainable Methodology for Neurological Diseases Early Detection. *J Healthc Inform Res*, 1-34.

[17] Oiza-Zapata, I., & Gallardo-Antolín, A. (2025). Alzheimer's Disease Detection from Speech Using Shapley Additive Explanations for Feature Selection and Enhanced Interpretability. *Electronics*, *14*(11), 2248.

[18] Rahman, W., ... & Hoque, E. (2021). Detecting parkinson disease using a web-based speech task: Observational study. *J Med Internet Res, 23*(10), e26305.

[19] Freeburn, J. L., … & Rezaii, N. (2025). Using Digital Speech Markers to Classify Functional Speech Disorder: A Proof-of-Concept Pilot Study. *J Mov Dis*.

[20] Celepli, S., ... & Erogul, O. (2025). SHAP-Based Identification of Potential Acoustic Biomarkers in Patients with Post-Thyroidectomy Voice Disorder. *Diagnostics*, *15*(16), 2065.

[21] Park, J. H., … & Lee, J. Y. (2025). Prediction of Voice Therapy Outcomes Using Machine Learning Approaches and SHAP Analysis: A K-VRQOL-Based Analysis. *Appl Sci*, *15*(13), 7045.

[22] Seedat, N., Aharonson, V., & Hamzany, Y. (2020, December). Automated and interpretable m-health discrimination of vocal cord pathology enabled by machine learning. In *CSDE* (pp. 1-6). IEEE.

[23] Calà, F., ... & Cantarella, G. (2023). Machine learning assessment of spasmodic dysphonia based on acoustical and perceptual parameters. *Bioengineering*, *10*(4), 426.

[24] Lyu, Y., ... & Xu, J. (2025). Non-invasive acoustic classification of adult asthma using an XGBoost model with vocal biomarkers. *Sci Rep*, *15*(1), 28682.

[25] Au, Y. C., & Ng, M. L. (2025). Developing a smart system for binary classification of disordered voices using machine learning. *Am J Otolaryngol*, 104672.

[26] Barlow, J., … & Kirke, D. N. (2024). The use of deep learning software in the detection of voice disorders: a systematic review. *Otolaryngol Head Neck Surg*, *170*(6), 1531-1543.

[27] Teng, Q., … & Lu, Y. (2022). A survey on the interpretability of deep learning in medical diagnosis. *Multimed Syst*, *28*(6), 2335-2355.

[28] Hassan, E., … & Shams, M. Y. (2025). Enhanced dysarthria detection in cerebral palsy and ALS patients using WaveNet and CNN-BiLSTM models: A comparative study with model interpretability. *Biomed Signal Processi Control*, *110*, 108128.

[29] Selvaraju, R. R., … & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc IEEE Int Conf Comput Vis* (pp. 618-626).

[30] Malekroodi, H. S., Madusanka, N., Lee, B. I., & Yi, M. (2024). Leveraging deep learning for fine-grained categorization of Parkinson's disease progression levels through analysis of vocal acoustic patterns. *Bioengineering*, *11*(3), 295.

[31] Jegan, R., & Jayagowri, R. (2025). Pathological voice detection using optimized deep residual neural network and explainable artificial intelligence. *Multimed Tools Appl*, *84*(19), 21863-21889.

SESSION II
SINGING VOICE ANALYSIS

THE INFLUENCE OF ESTILL VOICE TRAINING FIGURES ON ACOUSTIC AND ELECTROGLOTTOGRAPHIC PARAMETERS

M. Frič, A. Dobrovolná

Musical Acoustics Research Centre, Music and dance faculty of Academy of Performing Arts in Prague, Praha, Czechia
marekfric@centrum.cz, alena.dobrovolna@gmail.com

Abstract: **Voice production can be shaped deliberately by the Estill Voice Model's Figures. We examined how four of them—Thyroid Tilt, Laryngeal Vertical Position, Aryepiglottic Sphincter, and Body–Cover—affect the sound and vocal-fold contact across one-octave pitch range. Seven female singers sustained vowels at five notes while we simultaneously recorded acoustics and electroglottography (EGG). From each token we derived fundamental frequency (fo), SPL, spectral slope, SPR, formants, spectral centroids, jitter/shimmer, CPPS/HNR, and time-resolved EGG descriptors and cycle-averaged waveforms. Body–Cover most strongly altered EGG shape and noise; a higher larynx shifted F1, F2, spectral centroids, and a narrowed aryepiglottic region boosted 1.5–7 kHz energy with de-contact changes in EGG. These Figures yielded distinct, reproducible signatures that support objective differentiation and biofeedback-guided training; larger datasets with aerodynamics/imaging will refine predictive models.**

Keywords: **Estill Voice Model, Estill figures, electroglottography, acoustical parametrisation, EGG waveform**

I. INTRODUCTION

The Estill Voice Model (EVM) frames voice production as the trainable sequence of the positions of thirteen anatomical structures. These separate positions of the anatomical structures can be intentionally combined to create basic qualities making EVM a practical bridge between physiology and pedagogy [1]. In the EVM, a term 'Figure' denotes a specific, isolatable laryngeal or vocal tract configuration, such as the Thyroid Cartilage Figure (with its vertical and tilt conditions) or the Aryepiglottic Sphincter Figure (with its wide/narrow options). Figures are the building blocks of the model, i.e., structural manoeuvres that can be practiced individually and then combined to shape different vocal qualities. Empirical work on EVM has expanded rapidly across methods. Aerodynamic–acoustic studies show that body–cover qualities differ in subglottal pressure, airflow, SPL, and perturbation measures, though simple EGG closed-phase indices have been less sensitive [2]. Spectral studies in trained commercial singers demonstrate that performers can keep f_o stable while volitionally redistributing spectral energy captured by SPR/LTAS [3]. Imaging adds mechanistic context: MRI identifies consistent aryepiglottic and often oropharyngeal narrowing during twang with a frequently higher larynx [4,5], while synchronous videoendoscopy links pitch raising to laryngeal elevation, pharyngeal narrowing, and velar elevation [6]. High-speed laryngoscopy has further differentiated vibratory regimes associated with stylistic settings [7]. EVT has emerged as a particularly useful method for voice therapy and rehabilitation [8,9].

Despite this progress, systematic, Figure-specific mapping of both acoustic and detailed EGG signatures across pitch remains limited. Prior studies often examined either global qualities or isolated gestures, and EGG was typically summarized by a single ratio. The present study addresses these gaps by quantifying how four Figures—Thyroid Tilt (THY), Laryngeal Vertical Position (LVP), Aryepiglottic Sphincter (AES), and Body–Cover (BC)—shape acoustic and EGG derived metrics at five controlled pitches.

II. METHODS

Participants and tasks: Seven female participants (completed a regular, structured 6-month training program targeting the EVM Figures) were included in the study. Each participant produced different settings of the EVM Figures: THY (vertical, tilted), LVP (low, mid, high), AES (wide, narrow), and BC (thick, thin, stiff). For Thyroid Tilt (THY) and Aryepiglottic Sphincter (AES), an additional 'mid' category was occasionally assigned, reflecting that these configurations can be controlled in a continuous manner and some productions were perceptually judged as intermediate. Tasks were performed at five target pitches within one octave (A3, C4, D#4, F#4, A4) on the vowel /a/ at a comfortable medium dynamic. In total, 1,108 tokens were collected across the four Figures (247 for THY, 352 for LVP, 171 for AES, and 338 for BC).

Procedure: Synchronous recordings of the acoustic signal (at 30 cm) and electroglottographic (EGG) signal

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

(Laryngograph D-200, AGC disabled) were obtained in a sound-treated room. For each token, a 500-ms segment from the most stable portion of the sustained phonation was extracted. Perceptual validation was carried out by the first author.

Signal analysis: Acoustic signals were analyzed in Praat (version 6.3.16). Extracted parameters included fo, sound pressure level (SPL), formant frequencies (F1, F2), spectral slope, Singing Power Ratio (SPR), spectral centroids SC(0–2.5 kHz), SC(2–5 kHz), Jitter, Shimmer, CPPS, noise-to-harmonic ratio (NHR), and harmonic level relations (L1, L1–L2).

Electroglottographic signals were analyzed for both averaged amplitude and temporal patterns of vocal fold contact. Parameters such as Qx, Qci, SQx, and Ap–p were calculated following [10]. Cycle boundaries were detected with a wavegram-based alignment approach [11]; cycles were time-normalized to 0–100% of period length. The EGG was processed using two cycle-averaged EGG waveform (CAEW) methods [12]: (i) shape-normalized (SH-CAEWs, per-cycle min–max scaled between 0 and 1), and (ii) speech-baseline-referenced (SBR-CAEWs, amplitude scaled to a per-subject reference from the 'speech-like' baseline: LAR mid, BC thick, AES wide, THY vertical).

The cycle-synchronous EGG pulses were averaged within each token (500 ms), then averaged across tokens within each Figure setting before group averaging.

Statistical analysis: All acoustic and EGG parameters, including cycle-averaged EGG waveforms and spectra, were submitted to ANOVA to evaluate the effect of Figure settings. For CAEWs, values within 1% of the glottal cycle were compared, while spectral levels were assessed in 100-Hz bands. Bonferroni-corrected post hoc tests were applied where appropriate.

III. RESULTS

A. Spectral and EGG waveforms analysis

Fig. 1 presents comparative graphs of shape-normalized and speech-baseline-referenced CAEWs as well as acoustic spectra for different settings of EVM Figures: BC, AES, LVP, and THY. THY exerted only minimal influence on the shape of the EGG cycles. SBR-CAEWs showed only minor statistical differences in the maximum EGG amplitude. By contrast, the spectra revealed substantial differences at lower frequencies, close to the fundamental frequency, as well as in the amplitude of the first and second formant regions.

LVP proved a major factor influencing the EGG shape in the pre-contact phase of the vocal folds (0–20% of cycle length). Only minimal changes were observed in the cycle intervals 50–75% and 90–100%. SBR-CAEWs revealed highly significant effects across the entire vocal fold contact interval (20–55%). Spectral analysis showed pronounced changes in the second formant region (1.1–2.6 kHz) and in the higher range between 4.9–5.5 kHz.

Adjustments of the AES were mainly reflected in the EGG shape during the de-contact phase (55–74%). SBR-CAEWs revealed only minor statistical differences from maximum contact to maximum de-contact (35–75%). Spectrally, however, narrowing of the AES produced a broadband increase in the frequency range 1.5–7.2 kHz, with the most pronounced differences between 1.5–3 kHz.

The BC configurations exerted the strongest influence on the SH-CAEWs across most of the cycle, with the smallest effect between 65–90%. On SBR-CAEW, the most significant differences were observed in the ranges 0–50% and 85–100% of the period length. Spectrally, the entire frequency spectrum was affected.

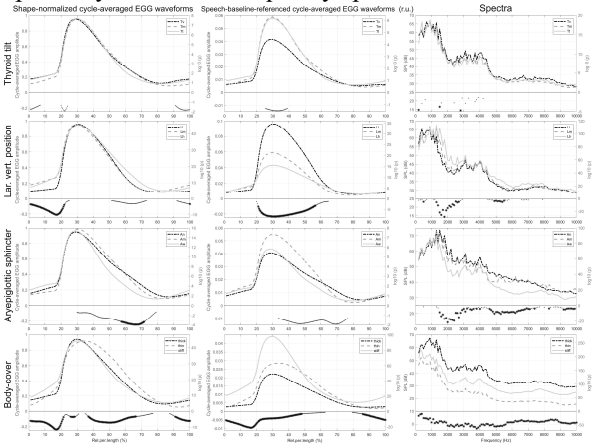


Fig. 1 *Top panels show condition-wise grand means of shape-normalized (SH-CAEW) and speech-baseline-referenced (SBR-CAEW)(per-subject) cycle-averaged EGG waveforms and averaged spectra (left y-axis); bottom panels display log10(p) from ANOVA (right y-axis) for Figure settings—THY, LVP, AES, and BC.*

B. Acoustic and electroglottographic parameters

In the upper parts of the graphs in Fig. 2, p-values from ANOVA are shown for different settings of the studied EVM Figures of the measured parameter. SPL-dependent metrics are interpreted with caution and as part of the observed Figure effects. **THY** exerted a moderate influence on spectral slope, CPPS, SPR, F_1, and HNR. **LVP** had highly significant effects on spectral slope and the formant positions (F_1 and F_2). These changes also influenced spectral centroids and HNR. Smaller but significant effects were observed on Shimmer, Qx, and EGG amplitude, with a moderate effect on Jitter. Adjustments of the **AES** produced very significant effects on spectral slope, SPR, F_1, F_2, and spectral centroids, a moderate effect on HNR and SPL, and a minor effect on Shimmer, Qx, and Qci.

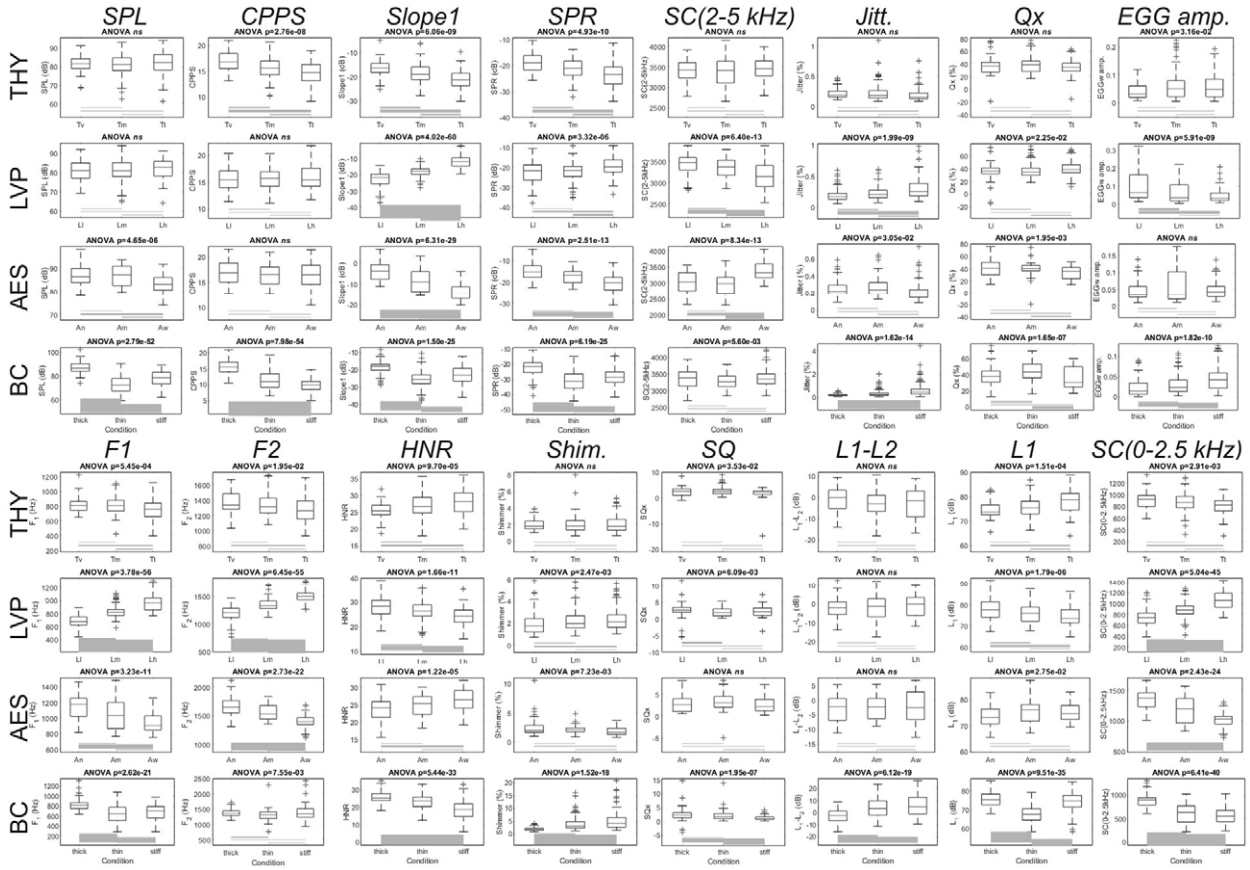


Fig. 2 *Comparison of measured parameter values across different settings of EVM Figures—THY (Tv - vertical, Tm - middle, Tt - tilt), LVP (Ll - low, Lm - middle, Lh - high), AES (An - narrow, Am - middle, Aw - wide), and BC (thick, thin, stiff)—together with p-values from the analysis of variance (ANOVA).*

BC adjustments exerted the most extensive influence on the measured parameters. They affected SPL and noise-related measures and produced the strongest impact on EGG-based parameters among all studied Figures. BC substantially influenced the position of the first formant and, consequently, the spectral centroid between 0–2.5 kHz

The boxplot graphs in Fig. 2 show the distributions of measured parameters for different Figure settings, enabling pairwise comparisons. The results provide a more detailed view of specific pairwise contrasts. Horizontal lines between conditions denote statistically significant differences, with line thickness proportional to the level of significance.

Slope and SC(0–2.5 kHz) were extremely sensitive to LVP, AES, and BC, whereas in THY the effects were only of moderate strength. SPR effectively distinguished BC and AES conditions and moderately separated THY. SC(2–5 kHz) strongly distinguished LVP and AES. Jitter, EGG amplitude, and HNR were highly sensitive to BC and LVP. Qx only moderately distinguished thin versus stiff conditions. F_1 and F_2 were sensitive to LVP and AES width, with F_1 also affected by BC. Shimmer, SQ, L_1–L_2, and L_1 were particularly sensitive to BC.

Systematic trends were evident across varying degrees of THY. With increasing tilt, CPPS, spectral slope, SPR, F_2, and SC(0–2.5 kHz) consistently declined, whereas HNR showed systematic increases. Increasing LVP led to systematic increases in spectral slope, SPR, Jitter, F_1, F_2, Shimmer, L_1–L_2, and SC(0–2.5 kHz). At the same time, parameters such as SC(2–5 kHz), EGG amplitude, HNR, SQ, and L_1 decreased. Narrowing of the AES systematically increased spectral slope, SPR, F_1, F_2, and SC(0–2.5 kHz), while HNR decreased. Adjustments of BC along the thick–thin–stiff continuum led to systematic decreases in CPPS, HNR, and SC(0–2.5 kHz). In contrast, Jitter, EGG amplitude, Shimmer, and L_1–L_2 showed systematic increases.

IV. Discussion and Conclusion

We observed strong BC effects on SPL and noise/perturbation (↑jitter, ↑shimmer; ↓HNR, ↓CPPS from thick→thin→stiff) and robust, phase-specific changes in EGG (Qx, SQx, amplitude; waveform across most of the cycle). This converges with [2], who

discriminated Thick/Thin/Stiff via Psg, airflow, SPL, and perturbations, but found limited sensitivity of a single EGG closed-phase metric. Our multi-parameter EGG analysis shows that time-resolved descriptors capture BC-dependent contact dynamics that closed-phase alone may miss.

Higher LVP produced steeper spectral slope, increased SPR and SC(0–2.5 kHz), and raised F1, F2, with decreases in SC(2–5 kHz), EGG amplitude, and HNR. These filter-side shifts mirror supraglottic adjustments reported during pitch elevation (laryngeal elevation, pharyngeal narrowing, velar elevation) [6]. Narrowing AES yielded broadband energy increases from ~1.5–7 kHz with clear rises in SPR and SC(0–2.5 kHz), modest HNR reductions, and EGG changes during de-contact. These spectral fingerprints match evidence of epilaryngeal narrowing in twang [4,5] and provide frequency-localized acoustic markers that correspond to the morphologic constriction. Thyroid tilt showed minimal effects on EGG cycle shape and moderate narrow-band spectral changes.

Novelty and implications: This study offers a parallel, pitch-controlled comparison of four Estill Figures within a single protocol, enabling direct contrasts of effect sizes and parameter specificity. Beyond prior acoustic work on body–cover [2,3], we add comprehensive cycle-averaged EGG waveforms to localize where in the cycle differences arise (pre-contact for LVP, de-contact for AES, and broad effects for BC). Frequency-resolved outcomes reveal LVP/AES influences that coarse metrics such as SPR alone may miss. Together with imaging and endoscopy showing aryepiglottic narrowing and laryngeal elevation [4–6], our results place the Figures within a coherent source–filter–impedance account.

Limitations and use: A small cohort (7 female singers). The study did not include independent visual verification of vocal tract configurations (e.g., endoscopy or MRI); thus, the precise physiological gestures cannot be fully guaranteed. Instead, our design followed established EVT pedagogy, grounded in prior experimental work showing reliable associations between Figures and laryngeal or articulatory configurations. Further multimodal studies are needed to establish direct physiological ground truth. While the primary aim was pedagogical, the observed patterns may also serve as baseline data for clinical research, especially in the treatment of functional (non-organic) voice disorders.

Acknowledgements: This publication was written at the Academy of Performing Arts in Prague as part of the project "*The influence of Estill Voice Training (EVT) on voice characteristics and technical and performance skills*" with the support of the Institutional Endowment for the Long-Term Conceptual Development of Research Institutes, as provided by the Ministry of Education, Youth and Sports of the Czech Republic.

REFERENCES

[1] Steinhauer KM, McDonald Klimek M, Estill J. *The Estill Voice Model: Theory & Translation*. Pittsburgh, PA: Estill Voice International; 2017.

[2] Barone NA, Ludlow CL, Tellis CM. Acoustic and Aerodynamic Comparisons of Voice Qualities Produced After Voice Training. *J Voice*. 2021;35(1):P157.E11-157.E21.

[3] Fantini M, Fussi F, Crosetti E, Succo G. Estill Voice Training and voice quality control in contemporary commercial singing: an exploratory study. *Logop Phoniatr Vocology*. 2017;42(4):146-152.

[4] Perta K, Bae Y, Obert K. A pilot investigation of twang quality using magnetic resonance imaging. *Logop Phoniatr Vocology*. 2020;0(0):1-9.

[5] Jelinger J, Perta K, Lee J, Wiksten N, Bae Y. Oropharyngeal and Aryepiglottic Narrowing for Twang: A Magnetic Resonance Imaging Study. *J Voice*. Published online 2024:1-10.

[6] Yanagisawa E, Yanagisawa E, Estill J, Talkin D. Supraglottic contributions to pitch raising: Videoendoscopic study with spectroanalysis. *Ann Otol Rhinol Laryngol*. 1991;100(1):19-30.

[7] Frič M, Dobrovolná A, Amarante Andrade P. Comparison of laryngoscopic, glottal and vibratory parameters among Estill qualities – Case study. *Biomed Signal Process Control*. 2024;87(April 2023).

[8] Grillo EU, Wolfberg J, Perta K, Van Stan J, Steinhauer K. Connecting Auditory-Perceptual Prompts Used in Voice Therapy to Anatomy and Physiology: Application to the Estill Voice Model and the Rehabilitation Treatment Specification System. *J Voice*. Published online 2024:1-16.

[9] Steinhauer K, Eichhorn K. Effect of Practice Structure and Feedback Frequency on Voice Motor Learning in Older Adults. *J Voice*. Published online 2023.

[10] Ternström S. Normalized time-domain parameters for electroglottographic waveforms. *J Acoust Soc Am*. 2019;146(1):EL65-EL70.

[11] Herbst CT, Fitch WTS, Švec JG. Electroglottographic wavegrams: A technique for visualizing vocal fold dynamics noninvasively. *J Acoust Soc Am*. 2010;128(5):3070-3078.

[12] Aaen M, Frič M. Going Beyond the Register—Vocal Mode Categorization Across Four Octaves in Professional Male and Female Singing Voice Using Voice Range Profile, EGG, Acoustic, and Vibroacoustic Measurements: Double-Case Study. *J Voice*. Published online 2025. doi:10.1016/j.jvoice.2025.06.003

PERCEPTUAL AND ACOUSTIC EVALUATION OF VIBRATO IN DIFFERENT SINGING STYLES

E. Globerson[1,2], O. Amir[3], O. I. Ronen[3], N. Amir[3]

[1]Technion-Israel Institute of Technology, Dept. of Humanities and Arts, Haifa, Israel
[2]Jerusalem Academy of Music and Dance, Jerusalem, Israel
[3]Dept. of Communication Disorders, Gray Faculty of Medical and Health Sciences, Tel Aviv University, Israel
eitan.globerson@gmail.com, oferamir@tauex.tau.ac.il, oryahronen@gmail.com, noama@tauex.tau.ac.il

Abstract: **Vibrato, a periodic variation in frequency and intensity, is a commonly found property of musical performance, adding a sense of flexibility and richness. Prior studies comparing vibrato in different singing styles employed mainly laboratory recordings, focusing on two principle features, rate and extent. The current study examined vibrato characteristics of classical, jazz and pop singing, employing commercial recordings of prominent artists. This provided a highly ecological description of artistic style. The presence of vibrato was determined perceptually by a panel of five professional musicians. Time-frequency analysis was then performed on pitch contours of sustained notes. A range of thresholds was applied to the time-frequency matrix to improve SNR. Several measures were derived from the thresholded matrix: vibrato extent, rate, consistency and onset time. Vibrato rate and extent were found to be higher in opera singing, compared to jazz and pop singing styles. Vibrato onset was found to be delayed mainly in jazz and pop singing. These results reinforce the importance of vibrato as a stylistically defining parameter and may serve as a tool in future research and vocal pedagogy, as well as in automatic recognition of singing styles.**

Keywords: **Singing, vibrato, singing style**

I. INTRODUCTION

Vibrato is a widespread phenomenon, found in playing of various musical instruments as well as in the singing voice. It is an effective musical tool, often associated with expressiveness, tone richness and affect in music performance [1,2,3]. There are several acoustic definitions for vibrato, all describing vibrato as regular changes in various acoustic properties, including frequency, timbre and intensity [4]. Most prior studies on vibrato focused on the characterization of two main parameters derived from the fundamental frequency: a) vibrato **rate**: the frequency of F0 undulation, and b) vibrato **extent**, defined as the range of F0 values in a single sustained tone. Vibrato rate in singing was previously found to be in the range of 5 to 9 Hz [5], and vibrato extent has been measured between 50 to 150 cents [6]. Most previous studies on vibrato focused on one style of singing, highlighting some basic features of vibrato defining this style [7,8,9]. Very few studies have employed the same methodology to compare vibrato characteristics in different singing styles. One comparative study by Bezerra et al. compared vibrato characteristics of opera and a traditional Brazilian style known as Sertanejo [10]. A significant difference in rate was found between the two groups, but not in extent. Another study by Manfredi et al. compared acoustic features of vibrato in jazz and opera singing, in students and professional jazz and opera singers. Their results demonstrated a wider vibrato extent and greater regularity in opera singing [11]. The study by Spradling & Binek found that jazz vibrato is often initiated after the onset of the vowel and can follow a straight tone (i.e., a tone with no vibrato), whereas classical vibrato is initiated directly at the onset of the vowel and usually does not change in rate during a sung vowel [7].

Some of the prior studies on vibrato employed recordings of professional singers as their database, e.g. [1,12,13] Recordings of prominent artists are considered a model of ideal singing, and therefore, can serve as optimal examples for any systematic study of vibrato. However, to our best knowledge, no prior studies employed recordings of prominent singers for a systematic comparison of vibrato features across different singing styles. The current study attempts to bridge this gap by examining a large corpus of commercial recordings of prominent female singers in three musical styles: classical singing of opera and art song, jazz, and pop singing. Since commercial recordings usually contain accompanying music, this introduces an additional challenge in signal processing, which is addressed in the current study by filtering the recordings and thresholding the time-frequency matrix.

Vibrato is not necessarily a regular periodic variation in fundamental frequency. In artistic practice it can take more complicated forms: its acoustic characteristics and even its presence can change during a note. In some cases, these changes may be systematic and intentional, whereas in some cases they may be the result of insufficient training or control [14]. This raises a need

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

Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

for defining robust analytical measures which can highlight time-dependent variations in vibrato characteristics. The current study addresses this issue, employing novel analytical methods which enable an exploration of various quantitative features such as rate, consistency and onset.

In addition to the acoustic analysis, the current study included a panel of professional musicians - vocal coaches and singers, to evaluate existence or lack of vibrato in all recordings. Since undulation in F0 has been found to exist ubiquitously in human singing [15], a perceptual evaluation is mandatory in order to differentiate between samples perceived as containing or lacking vibrato. The linkage between perception and acoustic characteristics of vibrato can help both researchers in the field of vocal analysis, as well as music pedagogues and performers seeking objective measures which can help reach a more artistically convincing production of various musical styles.

II. METHODS

A. Recordings

Vibrato samples were taken from commercially released albums of 30 female singers, performing in three different musical genres: 10 opera singers, 10 jazz singers and 10 pop singers. Singers' ages were between 23 and 50, and all excerpts were taken from albums recorded between 1990 and 2010. Opera singing was taken from recordings of operatic and lieder singing. All jazz singers performed in English. Four of the pop singers sang in Hebrew and six in English.

The performances of each singer were scanned manually for sustained singing of the vowel /a/ lasting more than 800ms. For each singer, a maximum of 20 excerpts of such sustained tones was obtained, with the minimum being seven.

B. Perceptual Evaluation

All musical excerpts were evaluated by a panel of five professional musicians. This panel included an opera conductor, an opera soprano singer, an opera coach, a jazz composer and a pop singer. A majority vote of three affirmatives was set as the threshold for presence of vibrato.

Overall, 429 excerpts were extracted, distributed in an approximately even manner over the different singing styles. Of these, 337 excerpts were judged unanimously to contain vibrato, and 25 were judged unanimously not to. Of the remaining 67 excerpts, 23 more were judged by only one or two judges to contain vibrato and were therefore considered to be without vibrato. While **all** opera singing excerpts were rated as containing vibrato, in the case of jazz and pop singing, only 80% and 84% respectively were rated as containing vibrato. The statistical analyses described below were therefore carried out only on the excerpts that were judged to actually contain vibrato.

C. F0 extraction

F0 extraction was complicated by the presence of background music, since the excerpts were taken from commercial recordings. To facilitate extraction, the pitch of the sung note was determined manually, followed by narrow band-pass filtering around the first or second harmonic. Subsequently, pitch was determined using Praat software [16], at 3ms. intervals.

D. Vibrato extent

The raw F0 contours obtained from Praat were then evaluated for extent. Instead of using range, which is susceptible to outliers, we opted for interpercentile range, expressed in semitones, as a more robust measure.

A. Time-Frequency analysis

Short-Time-Fourier-Transform (STFT) was performed on the pitch contour using a 400ms sliding window and an 80ms step size, resulting in a 2D matrix. Only values within the range of 3-8 Hz were considered for analysis. Subsequently, the values in the remaining STFT matrix were z-scored. All values below a z-score threshold were set to zero, conserving the more prominent vibrato-related values. We termed the resulting matrix the "Vibrato Matrix" (VM). A range of thresholds (1-10) was examined, to explore the association between threshold and the resultant vibrato feaures. This produced 10 VMs per excerpt. An example is shown in Fig. 1. For each combination of excerpt and threshold, the non-zero values remaining in the VM were used to calculate three derived features,:

1. Vibrato **consistency** (range=0-1): The duration of uninterrupted vibrato, divided by the total time of vibrato in the sound excerpt.
2. Mean vibrato **rate**: The mean frequency of the values in the thresholded vibrato matrix.
3. Vibrato **onset** (range=0-1): The first time point of non-zero values in the thresholded vibrato matrix, divided by the total time of the singing excerpt.

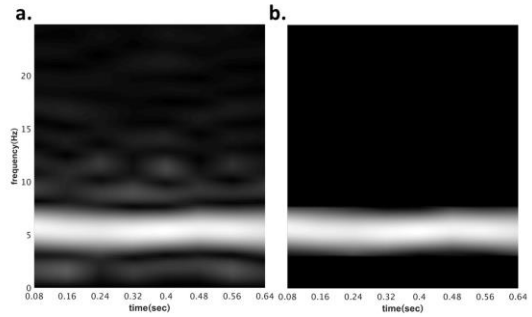


Fig. 1: a) STFT analysis of the F0 contour. b) Vibrato Matrix representing STFT values between 3-8 Hz, for a standard-score threshold of 10.

III. RESULTS

A. Extent

A box and whisker plot for vibrato extent, in semitones is presented in Fig. 2. In this case there appears to be a decrease in extent from opera, through jazz down to pop.

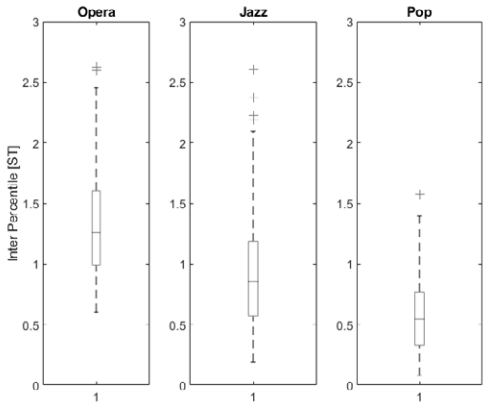


Fig. 2: Box and whisker plot of vibrato extent in each singing style.

A statistical analysis of both extent and duration was carried out after averaging these values individually per each singer. Duration, not directly related to vibrato, was analyzed to determine if there was any systematic difference of this value between the different styles.
Two one-way ANOVAs were performed on these variables, with a between-subject factor of singing style. A significant main effect was found for extent ($F(2,26)=21.125$, $p<0.001$). Post-hoc comparisons for mean extent revealed significant differences between each style and the others: Jazz and opera ($p=0.013$), jazz and pop ($p=0.006$), opera and pop ($p<0.001$), indicating that vibrato extent was highest in operatic singing, followed by jazz singing and then pop. No main effect was found for mean duration of excerpt, indicating that there was no specific trend for the excerpts to be longer or shorter in any specific style.

B. Consistency, rate and onset

Fig 3. Shows the additional three vibrato characteristics as a function of the VM thresholding. Opera singing appears to be most stable, in the sense that *consistency* and *onset* of opera singing are barely affected by the threshold value. However, for increasing thresholds, opera vibrato *rate* appears to diverge from jazz and pop vibrato rates. Increasing the threshold also indicates that pop and jazz tend on average to have later onset.

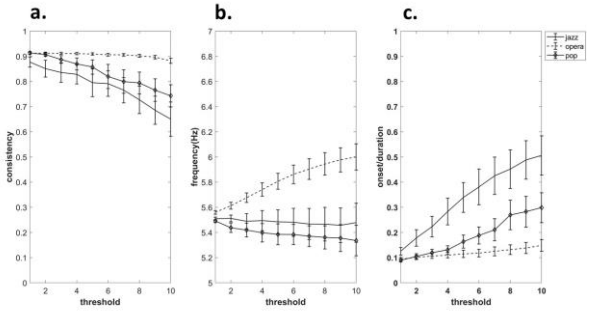


Fig. 6: Vibrato characteristics as a function of z-score thresholds: a) consistency, b) rate, c) onset (dotted line – Opera, thin line – jazz; thick line – pop).

IV. DISCUSSION

The current study found clear distinctions between the acoustic properties of opera singing and other singing styles. First, vibrato rate in opera singing was found to be higher than in jazz and pop singing. This is in contrast to a prior study [11] comparing jazz and opera singing, in which no such difference was found. This can be ascribed to major methodological differences between the two studies, both in how the recordings were obtained – lab recordings vs. commercial recordings - and also in the analysis methods which were applied. The current study employed a novel method, the thresholded vibrato matrix, which may contribute to the resolution of the analysis.
The difference in vibrato rate between opera singing and the other two musical styles can be attributed to a variety of factors, including deliberate discrepancies in vocal production between musical styles, as well as differences in vocal control and singing techniques. Future studies could look further into this phenomenon, by examining the effect of manipulation of vibrato rate on the perception of musical style in experienced listeners and professional singers.
In contrast to rate, for both vibrato extent and vibrato onset, differences were indeed found between jazz and pop singing also, adding an interesting perspective to the

characterization of these two musical styles. Regarding vibrato extent, the results of the current study replicate the findings of Manfredi et al. [11], demonstrating a larger vibrato extent for opera singing than for jazz. In addition, it was found that vibrato extent for pop singing was lower than for both opera and jazz. As for vibrato onset, a clear distinction was found between jazz and the two other singing styles. This finding is supported by many vocal methodologies, emphasizing the important role of delayed vibrato in jazz [7]. Hence, the results of the current study highlight difference between pop and jazz singing not previously documented. Also, one should take into account that pop singing is not as well methodologically defined as jazz. There are no golden-standard written methodologies on pop singing, while opera and jazz singing are based on well-defined traditions of performance and pedagogy.

An additional finding demonstrates a higher vibrato consistency in opera, compared to jazz singing. This can be explained by the longer and more formalized vocal training of opera singers, compared to any other existing style of singing. This result complements the results of Manfredi et al. [11], which showed lower (i.e., better) values of vibrato jitter and vibrato shimmer for opera singers as compared to jazz singers.

There are many practical implications to the results of the current study. For studies on the acoustics of singing, these results suggest that the analysis of commercial recordings can be used as a solid basis for studying the characteristics of vibrato in different singing styles. The methods and results of this study can also contribute to musicological studies focusing on different singing styles and provide professional singers and pedagogues with an accurate description of how leading artists implement their training in recorded performances.

The results of this study also highlight some acoustic differences between excerpts rated by human judgment as containing vibrato and those rated as “straight tones”. These results call for further studies on the psychoacoustics of vibrato, which could provide an accurate association between perception and the acoustics of vibrato.

REFERENCES

[1] Howes, P., Callaghan, J., Davis, P., Kenny, D., and Thorpe, W., “The relationship between measured vibrato characteristics and perception in Western operatic singing,” *Journal of Voice*, 18(2), pp. 216–230, 2004.

[2] Jansens, S., Bloothooft, G., and Krom, G. de., “Perception and acoustics of emotions in singing,” *Fifth European Conference on Speech Communication and Technology*, 1997.

[3] Seashore, C. E., “The Psychology of Music. VI. The Vibrato: (1) What is It?” *Music Educators Journal,* 23(4), pp. 30–33, 1937.

[4] Dejonckere, P. H., Hirano, M., and Sundberg, J., *Vibrato*, Singular Publishing Group, 1995.

[5] Howard, E., and Austin, H., *Born to Sing*, Music World, 2006.

[6] Horii, Y., “Acoustic analysis of vocal vibrato: A theoretical interpretation of data,” *Journal of Voice*, 3(1), pp. 36–43, 1989.

[7] Spradling, D., and Binek, J., “Pedagogy for the Jazz singer,” *The Choral Journal*, 55(11), pp. 6–17, 2015.

[8] Walker, G., “Good Vibrations: Vibrato, Science, and the Choral Singer,” *The Choral Journal*, 47(6), pp. 36–46, 2006.

[9] Mitchell, H. F., and Kenny, D. T., “Change in Vibrato Rate and Extent During Tertiary Training in Classical Singing Students,” *Journal of Voice*, 24(4), pp. 427–434, 2010.

[10] Bezerra, A., Cukier-Blaj, S., Duprat, A., Camargo, Z., and Granato, L., “The Characterization of the Vibrato in Lyric and Sertanejo Singing Styles: Acoustic and Perceptual Auditory Aspects,” *Journal of Voice*, 23, pp. 666–670, 2008.

[11] Manfredi, C., Barbagallo, D., Baracca, G., Orlandi, S., Bandini, A., and Dejonckere, P. H., “Automatic Assessment of Acoustic Parameters of the Singing Voice: Application to Professional Western Operatic and Jazz Singers,” *Journal of Voice*, 29(4), 517.e1-517.e9, 2015.

[12] Ferrante, I., “Vibrato rate and extent in soprano voice: A survey on one century of singing,” *The Journal of the Acoustical Society of America*, 130(3), pp. 1683–1688, 2011.

[13] Rothman, H., Diaz, J., and Vincent, K., “Comparing historical and contemporary opera singers with historical and contemporary Jewish cantors,” *Journal of Voice*, 14, pp. 205–214, 2000.

[14] Amir, O., Amir, N., Michaeli, O., “Acoustic and perceptual assessment of vibrato quality of singing students”, *Biomedical Signal Processing and Control*, 1, pp. 144-150, 2006.

[15] Wooding, R., and Nix, J., “Perception of Non-Vibrato Sung Tones: A Pilot Study,” *Journal of Voice*, 30(6), 762.e15-762.e21, 2016.

[16] Boersma, P., and Weenink, D., Praat: doing phonetics by computer [Computer program].

EFFECTS OF OVERTONE FLUTE BREATHING TRAINING ON VOICE RANGE PROFILES AND SPECTRAL OUTPUT

M. Frič[1], P. Amarante Andrade[1,2], J. Passerin[1,3], J. Kantor[3], M. Kučera[1,4]

[1]Musical Acoustics Research Centre, Music and dance faculty of Academy of Performing Arts in Prague, Czechia
[2]Curtin School of Allied Health, Perth, WA, Australia
[3]Institute of Special Education Studies, Faculty of Education, Palacky University Olomouc, Czechia
[4]Institute for the Treatment and Research of Communication Disorders, Ltd., Rychnov nad Kněžnou, Czechia
marekfric@centrum.cz, pedro.andrade@curtin.edu.au, johana.passerin@gmail.com,
jiri.kantor@upol.cz, hlascentrum@seznam.cz

Abstract: **This study investigated the impact of overtone flute training on female voice production. Ten vocally untrained women completed a five-week intervention with the shepherd's overtone flute, which requires airflow control without phonation. The aim was to test whether improved breath management influences vocal outcomes influence vocal outcomes.**

Voice recordings were collected before and after training. Tasks included habitual and stage reading, gradual call, glissando, and full singing voice range profiles (VRPs) at three dynamic levels. Long-term average spectrum (LTAS) was calculated from voiced segments. Comparisons were made using non-parametric tests.

Results showed significant increases in maximum phonation time, mean and maximum SPL in stage reading, and maximum f_0 in gradual call. Glissando and singing tasks revealed wider dynamic ranges, caused by higher SPL maxima and lower minima. LTAS analysis confirmed spectral strengthening above 2 kHz, most evident in stage speech, glissando, and medium-dynamic singing. These changes align with earlier findings linking high-frequency energy to epilaryngeal tube adjustments.

In conclusion, overtone flute training, despite being voiceless, produced measurable vocal improvements. The intervention enhanced breath control, and reinforced resonance in spectral regions important for projection. These outcomes suggest that overtone flute exercises could be valuable in voice education and rehabilitation.

Keywords: **Overtone flute, breath management, voice range profile (VRP), long-term average spectrum (LTAS), voice training**

I. Introduction

Breath management is regarded as a fundamental element of voice production and is often referred to as voice support [1]. Breathing exercises are widely applied in singing pedagogy and voice therapy. However, according to Herbst, support extends beyond breathing and encompasses a complex interaction of respiration, phonation, and resonance subsystems [2]. Within the respiratory system, lung volume and expiratory pressure provide the driving forces [3,4]; at the phonatory level, glottal resistance and adduction control airflow; and at the resonator level, vocal tract shape and length influence voice support [5].

Previous studies demonstrated that breathing exercises can increase sound pressure level (SPL) [6], extend maximum phonation time (MPT) [7], and enhance respiratory parameters [8]. They also affect phonatory outcomes such as pitch range and voice quality [6,8,9]. These improvements were noted in both singers and non-singers, and targeted interventions were also effective in clinical populations [10]. Despite these effects, consensus is lacking on the optimal strategy for training voice support, since professional singers employ variable breathing patterns [11,12].

The shepherd's overtone flute (koncovka in Czech) serves as a unique means of training breath management in isolation [13]. It is an intuitive instrument without finger holes; melodies are produced solely by modulating the exhalation, where changes in airflow intensity directly affect both pitch and loudness. The flute generates 7–8 overtone tones, and its exhalation pattern requires coordinated breathing cycles. Playing therefore demands precise expiratory control independent of phonation. Previous observations indicated systematic breathing patterns while playing the flute [14]. Clinicians have reported benefits for posture, respiration, and sensorimotor integration when using the instrument in therapy [13].

Based on earlier findings, the present study was designed to investigate whether overtone flute training without phonation can produce measurable changes in vocal outcomes among vocally untrained women. The aim was to evaluate changes in voice range profiles (VRPs) and long-term average spectrum (LTAS) after a five-week intervention.

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

Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

II. Methods

Ten vocally untrained normophonic women (aged 28–49 years, mean 39.1 ± 5.6) participated in this pre–post study. None of the participants had formal voice or music training, and all were free from vocal pathology at the time of testing. The study design deliberately avoided direct vocal training in order to isolate the effects of overtone flute breath practice on vocal output.

Breathing intervention: Participants engaged in a five-week training program using the shepherd's overtone flute. The instrument requires precise modulation of air pressure and flow but does not involve phonation. Training consisted of daily 10–15 min video-guided sessions focusing on airflow coordination, pressure control, and sustaining overtone series.

Voice range profiles (VRPs): VRPs were recorded in a treated studio using a calibrated Sennheiser MKE2 at 30 cm. The protocol comprised: habitual reading, stage reading simulating classroom projection, gradual call task, in which participants repeatedly called the word /ma:ma/ while progressively increasing loudness, mimicking a natural calling gesture, full singing $VRP_{\text{all dynamics}}$ on the syllable /va:/ across the physiological pitch range in three dynamics (softest,medium, loudest), and glissando at medium dynamics. Minimum, maximum, and mean values of fundamental frequency (f_o) and SPL were extracted, and VRP contours were constructed for each task.

LTAS was calculated from voiced portions of the reading and singing recordings for each task to assess global changes in energy distribution.

Statistics: Pairwise pre–post comparisons of VRP and LTAS parameters were conducted using non-parametric Mann-Whitney U-tests. Effect sizes (Cohen's d) were also calculated.

III. Results

Pairwise comparisons of VRP parameters before and after training are presented in Tab. 1. MPT showed a statistically significant increase of 3.4 seconds. For habitual reading, no significant changes were found. In the stage condition, mean and maximum SPL increased significantly. Maximum f_o also increased significantly during gradual call. In the glissando task, minimum SPL decreased, while maximum SPL and dynamic range increased.

In $VRP_{\text{all dynamics}}$, minimum SPL decreased, and training improved medium and softest dynamics by enhancing maximum f_o, f_o range, maximum SPL, dynamic range, and VRP area. VRP contours are illustrated in Fig. 1 (first and third columns). In the stage condition, SPL maxima in the A3–B3# f_o range increased significantly. Maximum f_0 and tone range increased during gradual call. In glissando, SPL values increased across nearly the entire C5–C6 octave. Minimum SPL between F3 and A4 decreased for the softest dynamic. For medium dynamics, both SPL minima and maxima increased across the upper portion of the f_o range. In the loudest dynamic, SPL maxima increased above F5 and SPL minima decreased between F3 and F4. In $VRP_{\text{all dynamics}}$, significant SPL maxima increases were observed in the E5–A5 region, while SPL minima decreased between C3 and C5.

LTAS analysis is shown in Fig. 1 (second and fourth columns). Spectrally, the habitual voice, gradual calling, and loudest dynamics exhibited no significant changes. Other tasks revealed a general increase in SPL levels across the entire frequency range for the post-training condition. Overall $VRP_{\text{all dynamics}}$ showed a systematic rise in the 2–6.5 kHz spectral band.

IV. Discussion and Conclusion

The average increase in MPT of 3.4 s supports the efficacy of breathing exercises in improving coordination between respiratory and phonatory systems. As phonation was not directly involved in the intervention, these changes are more plausibly linked to respiratory function rather than glottal adjustments. This aligns with findings from incentive spirometer training in children [7].

Although no changes occurred in habitual speech, significant increases in mean and maximum SPL were found in the stage task, with further gains in glissando and gradual call. Singing tasks showed extended dynamic ranges, explained by both increases in maximum SPL and decreases in minimum SPL. Similarly, maximum f_o and f_o range increased during gradual call and singing. These results suggest that overtone flute training benefits tasks requiring greater coordination and energy.

VRP contours resembled those in earlier study [15]. After training, peak SPL increased within $VRP_{\text{all dynamics}}$ (E5–A5 range), in the upper half of medium dynamics, and above C5 in glissando, while SPL minima decreased in C3–C5. These findings confirm an expanded vocal dynamic range. Pre-training VRPs resembled those of novices, while post-training contours shifted toward trained singers, especially in SPL maxima above C5 [16]. LTAS analysis revealed spectral increases above 2 kHz, particularly during glissando and softest-dynamic singing. Such changes may result from higher SPL, since louder phonation strengthens high-frequency harmonics. At the same time, the results are consistent with earlier reports linking strengthened high-frequency energy to epilaryngeal tube narrowing in classical resonance strategies, resonance tube and straw phonation [17], and loud twang-like voice [18].

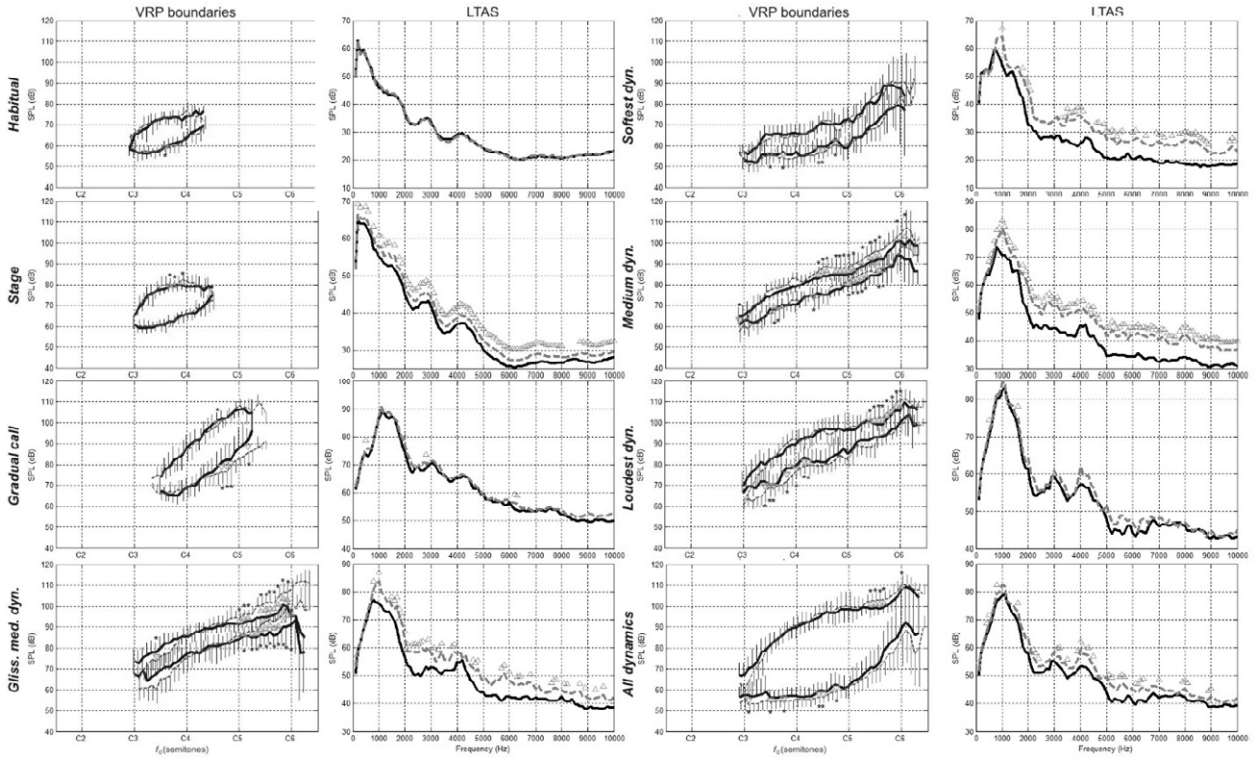


Fig. 1 *Comparison of the VRP contours and LTAS for the measured tasks. The pre-training condition is depicted by the thick black line, and the post-training condition is depicted by the gray dashed line. Gray triangles in the LTAS graphs and black stars in the VRP graphs indicate statistically significant differences determined by paired sample t-tests.*

Tab. 1 *Results of pairwise comparisons of VRP parameters for measured tasks using Mann–Whitney U-tests, showing only statistically significant differences before and after training.*

	Task	**Habitual**	**Stage**		**Calling**	**Glissando**			**VRP all dynamics**	**Softest dynamics**	
Par.	**Stat.**	**MPT (s)**	**SPL mean (dB)**	**SPL max. (dB)**	**f_o max. (Hz)**	**SPL min. (dB)**	**SPL max. (dB)**	**dynamic range (dB)**	**SPL min. (dB)**	**f_o max. (Hz)**	**f_o range (Hz)**
before	**median**	23.3	71	81	542	64	101	40	53	929	781
	25, 75pct.	15.3, 27.4	69, 72	79, 81	513, 596	58, 68	96, 103	26, 45	50, 57	747, 1103	612, 954
after	**median**	23.8	73	84	613	59	110	51	50	1026	885
	25, 75pct.	22.6, 33.2	71, 74	80, 86	565, 700	55, 65	105, 111	44, 53	49, 53	980, 1225	858, 1070
MWU	**p-val.**	***0.03***	***0.002***	***0.01***	***0.002***	***0.027***	***0.013***	***0.0098***	***0.042***	***0.0371***	***0.0195***
Effect size	**Cohen's d**	***0.45***	***1.52***	***1.01***	***1.56***	***0.71***	***0.99***	***0.94***	***0.68***	***0.80***	***0.84***

	Task	**Softest dynamics**			**Medium dynamics**						**Loudest dyn.**
Par.	**Stat.**	**SPL max. (dB)**	**dyn. range (dB)**	**VRP area (ST * dB)**	**f_o max. (Hz)**	**f_o range (Hz)**	**SPL min. (dB)**	**SPL max. (dB)**	**dyn. range (dB)**	**VRP area (ST x dB)**	**VRP area (ST x dB)**
before	**median**	85	32	304	992	851	62	97	38	270	376
	25, 75pct.	80, 90	26, 39	263, 362	832, 1260	694, 1130	60, 63	93, 102	31, 43	231, 316	272, 409
after	**median**	91	39	390	1117	977	60	104	45	359	462
	25, 75pct.	89, 98	35, 48	361, 419	980, 1297	821, 1200	57, 61	101, 108	42, 50	311, 374	395, 513
MWU	**p-val.**	***0.0273***	***0.0059***	***0.0488***	***0.0098***	***0.0059***	***0.0098***	***0.002***	***0.002***	***0.002***	***0.0137***
Effect size	**Cohen's d**	***0.80***	***1.16***	***0.87***	***1.18***	***1.15***	***0.81***	***1.30***	***1.45***	***1.06***	***0.91***

The overtone flute may therefore function similarly to semi-occluded vocal tract exercises, indirectly eliciting increased epilaryngeal impedance and resonance. To disentangle these mechanisms, future work will include a more detailed analysis integrating electroglottographic signals and separating the effects of SPL changes from potential resonance adjustments.

Limitations: This study should be regarded as preliminary due to the small number of participants. Although potential effects of shyness or task repetition cannot be fully excluded, these were minimized by providing all subjects with repeated VRP trials prior to the intervention. Nevertheless, the absence of a control group does not allow a definitive attribution of the observed improvements solely to overtone flute training, and future studies should address this limitation.

In summary, overtone flute training led to measurable improvements in vocal performance, including longer phonation times, greater loudness in stage speech, and expanded pitch and dynamic ranges, particularly in the upper register. The consistent spectral strengthening above 2 kHz further suggests improved resonance characteristics associated with epilaryngeal tube adjustments. These results indicate that overtone flute exercises may support respiro-phonatory coordination and have potential in voice education and rehabilitation.

Acknowledgements: This work was supported by the Institutional Endowment for Long-Term Conceptual Development of Research Institutes at the Academy of Performing Arts in Prague, funded by the Ministry of Education, Youth and Sports of the Czech Republic. The study was also financially supported by a project titled “Research of inclusion in individuals with special needs with respect to specific interventions” (IGA_PdF_2022_022) and a project titled “Concept Evidence-Based Practice in Special Education and Arts Therapies”, both projects funded by Palacky University Olomouc.

References

[1] Sonninen A, Laukkanen AM, Karma K, Hurme P. Evaluation of Support in Singing. J Voice. 2005;19(2):223-237.

[2] Herbst CT. A Review of Singing Voice Subsystem Interactions—Toward an Extended Physiological Model of “Support.” J Voice. 2017;31(2):249.e13-249.e19

[3] Cossette I, Fabre B, Fréour V, Montgermont N, Monaco P. From breath to sound: Linking respiratory mechanics to aeroacoustic sound production in flutes. Acta Acust united with Acust. 2010;96(4):654-667

[4] Collyer S, Kenny DT, Archer M. The effect of abdominal kinematic directives on respiratory behaviour in female classical singing. Logop Phoniatr Vocology. 2009;34(3):100-110

[5] Griffin B, Woo P, Colton R, Casper J, Brewer D. Physiological characteristics of the supported singing voice. A preliminary study. J Voice. 1995;9(1):45-56

[6] Schaeffer N. Pre- and Poststimulation Study on the Phonatory Aerodynamic System on Participants with Dysphonia. J Voice. 2017;31(2):254.e1-254.e9

[7] Choi JY, Rha D, Park ES. Changes in pulmonary function after incentive spirometer exercise in children with spastic cerebral palsy. Yonsei Medicak J. 2016;3(57):769-775.

[8] Ray C, Trudeau MD, McCoy S. Effects of Respiratory Muscle Strength Training in Classically Trained Singers. J Voice. 2018;32(5):644.e25-644.e34

[9] Roy N, Weinrich B, Gray SD, Tanner K, Stemple JC, Sapienza CM. Three treatments for teachers with voice disorders: A randomized clinical trial. J Speech, Lang Hear Res. 2003;46(3):670-688

[10] Smith ME, Ramig LO, Dromey C, Perez KS, Samandari R. Intensive voice treatment in parkinson disease: Laryngostroboscopic findings. J Voice. 1995;9(4):453-459

[11] Thomasson M, Sundberg J. Consistency of phonatory breathing patterns in professional operatic singers. J Voice. 2001;15(3):373-383

[12] Sundberg J, Thalén M. Respiratory and Acoustical Differences Between Belt and Neutral Style of Singing. J Voice. 2015;29(4):418-425

[13] Kučera M. Overtone Flute: A Traitional Czech and Slovak Instrument in Neurorehabilitation Practice. Published 2020. Accessed September 10, 2022. http://www.drmag.cz/wp-content/uploads/koncovka-eng-web.pdf, https://www.youtube.com/watch?v=M305S4Lu-Bo

[14] Frič M, Kučera M. [Optical analysis of breathing movements during exercise with a simple overton flute - koncovka]. In: Nové Trendy Akustického Spektra 2014. Technická Univerzita vo Zvolene; 2014:63-72.

[15] Åkerlund L, Gramming P, Sundberg J. Phonetogram and averages of sound pressure levels and fundamental frequencies of speech: Comparison between female singers and nonsingers. J Voice. 1992;6(1):55-63

[16] Frič M. Comparison of voice range profile parameters between males and females. Akustika. 2018;30(September):42-63.

[17] Guzman M, Laukkanen AM, Krupa P, Horáček J, Švec JG, Geneid A. Vocal tract and glottal function during and after vocal exercising with resonance tube and straw. J Voice. 2013;27(4):523.e19-523.e34

[18] Saldías M, Laukkanen AM, Guzmán M, et al. The Vocal Tract in Loud Twang-Like Singing While Producing High and Low Pitches. J Voice. 2021;35(5):807.e1-807.e23

IMPACT OF VOCAL TRACT RESONANCES ON OBOE PLAYING

Annika Koop[1], Malte Kob[2]

[1] Hanover University of Music, Drama and Media, Hanover, Germany
[2] Detmold University of Music, Erich Thienhaus Institute, Detmold, Germany
koopa@stud.hmtm-hannover.de, malte.kob@hfm-detmold.de

Abstract: **Oboists form vowels in their mouths while playing to improve the response of certain notes or to change the timbre. We developed an impedance measurement device that allows the measurement of the intraoral impedance while playing the oboe. Several students were measured while playing different pitches and articulating the vowels /e:/, /i:/ and /o:/. For some notes, the signal spectra show a strong influence on the strength of the partial tones under different articulation conditions, while some notes are not affected by this.**
Keywords: **vocal tract impedance, player-instrument interaction, reed instrument**

I. INTRODUCTION

The pitch and timbre of the oboe are largely determined by the instrument (resonator) and the reed (sound generator) [1]. If the player is dissatisfied with the timbre, they will make a new mouthpiece, which is a time-intensive process. However, when comparing beginners and teachers, it is noticeable that the teacher can produce a much more pleasant sound than the student with the same reed and the same instrument. So what do professional oboists do differently from beginners? Possible factors are air flow, mouth resonance or even the position of the lips holding the reed. In order to explain these factors more clearly, pupils should form the vowels /e:/ and /o:/ with their mouths while playing [2]. Some teachers use this technique not only to improve the response of the notes, but also to change the timbre, and so vowel formation has also found its way into contemporary playing techniques [3].

The influence of the vocal tract on wind instruments has been investigated several times using various methods. The results range from ‘no influence’ to ‘influence under certain conditions’ [1]. These studies were primarily conducted on single-reed instruments and recorders and suggest that the results also apply to double-reed instruments.

Different vowels are used depending on the note. Long-fingered notes produce a stable, long-standing wave throughout the instrument and are played with an /o:/ formed in the mouth. Short-fingered notes have keys that are already open after a few centimeters, which is why the wave produced in the instrument is very short. Such notes are played with /e:/ and sometimes also /i:/. It should be checked whether resonances occur in the oral cavity at all, or whether the embouchure when playing the oboe is a fixed position. Are the vowels imagined for psychological reasons and are they not actually there?

This investigation should help students understand what is happening in their mouths and how they can achieve a better sound. On the other hand, this understanding can help to develop new contemporary playing techniques.

II. METHODS

A. Study design

In a pre-test of the new measurement device, impedance measurements of the open mouth were conducted to test the reliability of the new set-up.

Four music students majoring in oboe were selected for the study, representing different semesters. Three of them were female, one participant was male, and all were between 20 and 27 years old. The most experienced player is studying oboe at the master's level, while the three education students are in their fourth master's semester, eighth bachelor's semester, and second bachelor's semester.

We conducted a series of measurements with each test subject with the oboe in place and a sustained note.

The notes e4, c5 and c6 were selected for further analysis in order to cover all three octaves on the oboe. The note e4 is considered stable, and it is assumed that the oboist can only exert minimal influence on it. Usually, the vowel /o:/ is formed when playing. For comparison, we also added the vowel /e:/, which is otherwise more commonly used in higher octaves. Conversely, the note c5 is classified as a short-fingered note. This indicates that the resonator is considerably shorter. There is no universal vowel here, which is why

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

we tried all three vowels. In the third octave, c6 was selected as the representative note, for which the vowel /i:/ is often used to enhance its resonance.

B. Impedance measurements

We used the BIAS system (Artim) to measure the impedance at the reed of the oboe for various fingerings using an adapter to allow an air-tight connection between the staple of the oboe and the BIAS measurement head [4]. The results were then exported and processed using a Python script.

Fig. 1: Photo of the impedance measurement device

For the measurement of the intraoral impedance a new device was developed using a 3D printer, a small loudspeaker (Visaton BF 45), a microphone capsule (Sennheiser KE 4-211-2) and a few screws.

The design is inspired by research of Chen et al. [5], in which a device was used with the microphone and speaker ducts built directly into the saxophone mouthpiece. In our approach, due to the small mouthpiece of the oboe, the measurement device was separated from the instrument, as shown in Fig. 1.

The schematic set-up is shown in Fig. 2. A swept sine is generated by the software WinMF [6] (Fouraudio) with Babyface (RME) audio interface, amplified, and fed to the loudspeaker which is mounted inside an air-tight housing with conical outlet. Absorbing foam is installed inside the conical duct to reduce internal standing waves. Another duct is constructed such that the microphone captures the sound very close to the loudspeaker outlet.

The size of the capillaries is based on the mouthpiece of the oboe to disturb the player as little as possible. The oval opening is the same size for the sound transmission of the microphone and the loudspeaker and is approximately 5 mm in length and 1 mm in width. The edge is approximately 1 mm thick. The capillary for the loudspeaker is 120 mm long, and the microphone is inserted into the 70 mm long capillary.

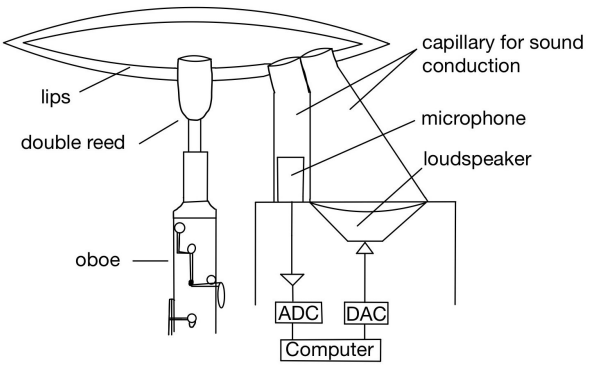


Fig. 2: Schematic structure of the experiment

Before measurement of the mouth cavity, a reference measurement is performed and stored without any structure attached. Subsequent measurements are divided by this reference measurement to obtain the normalized acoustic impedance (see [7] for details).

III. RESULTS

The results of the pre-test show that the impedance curves from measurements of oral resonances without sounding oboe vary significantly among all subjects.. However, it can be observed that the ability to keep the vowels constant is more pronounced in the two test subjects from the master's programme, while the other two show significant deviations.

All participants were able to significantly modify the partials of the note c5 due to the shaped vowel. As can be seen in Fig. 3, the levels of the overtones for the vowel shape /e:/ are increased with reference to the fundamental frequency, while for the vowel /o:/ they are reduced. Participants can therefore make the oboe sound brighter [8] by forming an /e:/. It is also noticeable that the fundamental is approximately 110 Hz above the first impedance peak of the oboe. This suggests that the player can influence not only the timbre but also the intonation on the c5.

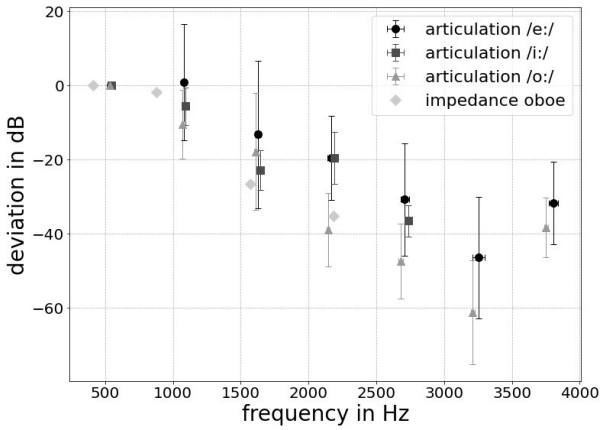


Fig. 3: This figure shows the mean values of the maxima from the impedance curves recorded in the oral cavity when playing the note c5. The values were normalized by setting f_o (the first partial) to 0 dB. The deviation in dB from f_o was plotted over the frequency. The size of the error bars corresponds to the standard deviations for all subjects. The impedance peaks of the oboe without a player are shown as a reference.

When evaluating the data, it is noticeable that some participants were able to achieve a significantly greater deviation in the level of the partials by changing the vowel than others. The standard deviation also given in Fig. 3 suggests that the aptitude to change the timbre is much more pronounced in some of the study participants than in others.

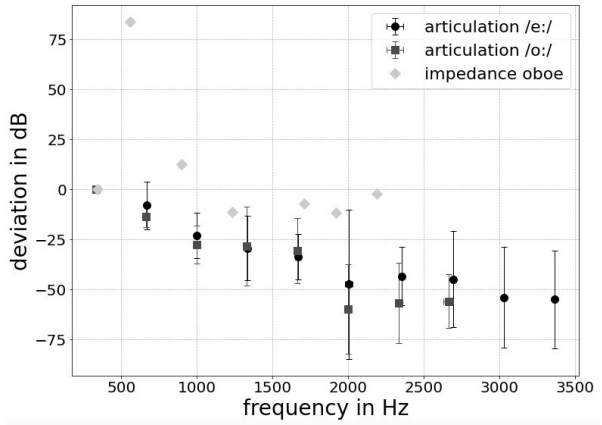


Fig 4: The figure shows the mean values of the maxima from the impedance curves recorded in the oral cavity when playing the note e4. The values were normalized by setting f_o (the first partial) to 0 dB. The deviation in dB from f_o was plotted over the frequency. The size of the error bars corresponds to the standard deviations for all subjects. Again, the impedance peaks of the oboe without a player are shown as a reference.

The evaluation for the long-fingered note e4 shows hardly any difference between the formed vowels /e:/ and /o:/ as displayed in Fig. 4. All partial tones are below the impedance of the oboe and show a significantly higher deviation from the fundamental tone than in the case of c5.

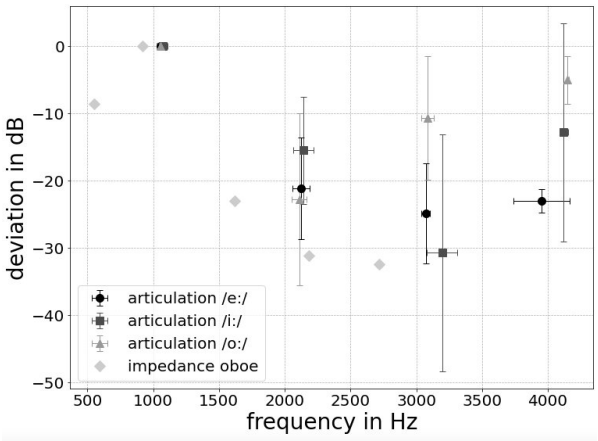


Fig. 5: This figure shows the mean values of the maxima from the impedance curves recorded in the oral cavity when playing the note c6. The values were normalized by setting f_o (the first partial) to 0 dB. The deviation in dB from f_o was plotted over the frequency. The size of the error bars corresponds to the standard deviations for all subjects. The impedance peaks of the oboe without a player are shown as a reference.

During the recordings for the note c6, there were increased levels of multiphonics, particularly with the vowel /o:/, which is why only a few recordings were available for evaluation. In Figure 5 it can be seen that the brightest sound is produced with the vowel /o:/ and that /i:/ causes the greatest deviation in the volume of the overtones. That is an interesting observation, as the fingering for c5 only differs from that for c6 in the octave key. There is therefore a connection between the shaping of the vocal tract, the fingering for the notes and the octave.

IV. Discussion

Due to the compact construction, unwanted resonances occurred in the range of 3500 Hz to 4000 Hz, which could not be completely eliminated using the absorbent filling and the reference measurement. More artefacts occurred at low frequencies below 400 Hz. Therefore, an improved setup would be desired to evaluate these ranges. Furthermore, in a coupled system consisting of the oboe, the double reed, and the player, it is challenging to evaluate the impact of just one modification such as the change of the vocal tract resonances. Any change in articulation also changes the position of the lips and the lip pressure on the double reed. Also, the variation of reed pressure would change the input impedance of the instrument as well as the

resonance structure in the mouth cavity. These variations are difficult to control.

In order to make a valid statement about the actual influence of the vocal tract on the sound of the oboe, further data from more test participants would be required. This would also enable us to investigate whether and to what extent the results deviate between professional players and amateurs, and whether the results are related to the players' performance level. The extended study should be accompanied by a psychological questionnaire to obtain subjective feedback from the players. This would allow us to find out whether each player develops their own technique or whether certain notes respond best to certain vowels.

In this paper, we examined whether the parallel measurement of impedance and sound spectra in the oral cavity works. In further experiments, we will also record the external sound under controlled conditions and compare its analysis to the oral sound. This also includes whether the players feel that they can influence their sound.

Furthermore, this experiment deals with the maximum change in the strength of the partial tones. This also leads to changes in the pitch of the fundamental tone and the partial tones, which has been omitted from this evaluation in order to focus on the timbre.

V. Conclusion

The new measurement method showed that forming the vowels /e:/, /i:/ and /o:/ while playing the oboe causes a change in the proportions of partial tones in the oral cavity. The vowel /e:/ produces the brightest sound on the note c5 while the partials of the e4 are rather stable which suggests that for the low register of oboe the timbre is not much influenced by the players' articulation. For short-fingured notes such as c5 and c6, the vocal tract has a greater influence on the partials than for long notes such as e4, as the vocal tract is small in relation to the resonating body, the oboe.

In addition, some players were able to keep their mouth shape more consistent than others, which is probably related to their experience.

By forming the vowel /o:/ on the c6, participants found it easy to produce multiphonics. This could simplify the learning of contemporary playing techniques.

To be able to use the results for teaching purposes, further research is needed to find out which notes sound best with which vowels.

ACKNOWLEDGEMENTS

We are grateful for the three subjects' interest and patience during the measurements. Luis Roca Paz is acknowledged for his support with the construction of an earlier version of the measurement device. Timo Grothe is acknowledged for his support with adaptation of the oboe to the BIAS system.

REFERENCES

[1] M. Oehler, *Die digitale Impulsformung als Werkzeug für die Analyse und Synthese von Blasinstrumentenklängen*, Berlin: Lang, 2008, pp.71-76.

[2] P. Veale, C.-S. Mahnkopf, W. Motz, and T. Hummel, *The techniques of oboe playing*, 4. ed, Kassel: Bärenreiter, 2018, p.146.

[3] K. I. Tracz, *Singing through the oboe – voicing and other vocal techniques within playing and teaching*, Greensboro: University of North Carolina, 2021, pp.20-22.

[4] T. Ossman, H. Pichler and G. Widholm, "BIAS: a computer-aided test system for brass wind instruments", *Journal of the audio engineering society*, vol. 2834, 1989.

[5] J. M. Chen, J. Smith, & J. Wolfe, "Experienced saxophonists learn to tune their vocal tracts", *Science*, vol. 319, pp. 776–776, 2008.

[6] M. Makarski, *WinMF – measurement software*, Four Audio GmbH & Co. KG, version 1.22, http://www.winmf.de, 2022.

[7] M. Kob, & C. Neuschaefer-Rube. "A method for measurement of the vocal tract impedance at the mouth," *Medical Engineering & Physics,* vol. 24(7), pp.467–471, 2002.

[8] C. Reuter, (2002). Klangfarbe und Instrumentation: Geschichte – Ursachen – Wirkung, *Die Oboe,* vol 5. Berlin: Lang, 2022, pp.91-132.

MULTIMODAL FEATURE ANALYSIS FOR DETECTING EXPRESSIVITY IN SINGING USING A MACHINE LEARNING APPROACH

N. Kotsani[1], V. Lyberatos[1], S. Kantarelis[1], A. Andreopoulou[2], G. Stamou[1], and A. Georgaki[2]

[1]National Technical University of Athens, Athens, Greece
[2]National and Kapodistrian University of Athens, Athens, Greece
nkotsani@corelab.ntua.gr, vaslyb@ails.ece.ntua.gr, spyroskanta@ails.ece.ntua.gr, aandreo@music.uoa.gr, gstam@cs.ntua.gr, georgaki@music.uoa.gr

***Abstract*—In this study, five singers performed under expressive and neutral conditions in front of a listener, while biometric signals including electroencephalography (EEG), galvanic skin response (GSR), and electrocardiography (ECG) were recorded from both performers and audience, alongside high-quality audio recordings of the singers, and emotion annotations. Audio features were extracted and analyzed. Physiological signals were processed to derive markers of relaxation, heart rate variability, and arousal. Four machine learning models were trained to classify expressive versus neutral performances, achieving up to 82% cross-validated accuracy, with pitch, loudness, and spectral features emerging as the most informative predictors. Multimodal analysis revealed additional discriminative patterns, as performers exhibited higher EEG-based relaxation during expressive performances, while audience heart rates reflected heightened arousal. Furthermore, emotional alignment between performers and listeners was observed predominantly in the expressive condition. Audio and behavioral features obtained from the multimodal analysis were consistent with the broader dataset collected within the witheFlow project, which included 16 instrumentalists, allowing for direct comparison with the current study. These findings indicate that vocal expressivity can be reliably quantified using combined audio and physiological analyses, providing a foundation for future studies on emotional communication in music performance.**

***Keywords:*—Expressive Singing, Audio Signal Processing, Physiological Signals, Machine Learning**

I. Introduction

According to Sundberg et al. [1] "expressiveness is generally regarded as a most important, though subjective, quality of performed music". Numerous studies have examined the audio features that shape singers' expressivity and allow them to communicate emotional coloring effectively to listeners. In Kotyar and Morosov's acoustic experiments [2], listeners identified intended emotions of eleven professional singers in about 80% of the cases on average, though only 56% for joy. Sundberg et al. [3], studying the prevailing emotional colors of performed excerpts, revealed differences in loudness, tempo, phonation type, and the rate of change of sound level. Siegwart and Scherer [4] using recordings from opera excerpts and eleven experienced listeners identified two component scores explaining 84% of variance in preference ratings. Sundberg et al. [1] focusing on F0 extraction, concluded that the singer sharpened tones more when he sang the examples as expressive as in a concert than when he sang them as void of expressivity as he could, and in their recent work on emotional coloring in professional singers performing scales with vowels in ten emotional categories [5], reported that LTAS parameters affect listener judgment of the enacted emotions and the accuracy of the intended emotional coloring.

In this study, five singers performed songs under expressive and neutral conditions while physiological signals (EEG, GSR, ECG) were recorded from both performers and listeners. Audio features including pitch, loudness, dynamics, and spectral measures were extracted, and the physiological data were subsequently analyzed.

II. Methodology

A. Participants and Dataset Creation

Five singers (two female, three male), with a mean age of 39.6 ± 3 years and varying levels of professional experience each, chose two public-domain songs. Five audience members (two female, three male) with a mean age of 35.2 ± 4.8 years listened without selecting which singer to attend. On arrival, all participants completed a demographic questionnaire and signed consent forms allowing the use of their recordings and biometric data for research. Each participant, both performers and audience members, wore three sensors: a four-electrode commercial electroencephalogram (EEG) headband, an electrocardiogram (ECG) sensor and a galvanic skin response (GSR) sensor attached to the index and middle fingers.

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

Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

B. Experimental Protocol

Audio recordings were obtained in a sound-isolated room using a two-channel USB audio interface and an industry-standard dynamic vocal performance microphone, recorded in mono. Each session was conducted in the presence of the performer, one audience member, and two researchers. Biosignals were captured on separate computers for the performer and the audience member, with synchronized timestamps applied to ensure precise dataset alignment.

The performers completed a brief warm-up and a two-minute rest (with dimmed lights, during which participants were asked not to keep their eyes closed) to ensure proper device calibration. Each then performed their two self-selected public domain songs under neutral (as unexpressive as possible) and expressive conditions across five sessions, each lasting about one hour. At the end of each recording, both the listener and the performer assigned at least one categorical emotion tag proposed by the GEMS-9 emotion framework [6] plus the ”neutral” tag. Selections were made without knowledge of the other party’s choices, allowing the procedure to serve as a blind assessment of emotional alignment between performers and listeners. This procedure was used to estimate the degree of emotional alignment between performers and listeners. After each session, participants rated their stress (1–5) and any discomfort from the sensors to control for possible psychological or physical confounds.

C. Data pre-processing and curation

The total duration of all mono recordings was approximately 2,607 seconds (∼43 minutes), with a median of 115.5 seconds per track. Expressive and neutral recordings were balanced (1,327 vs. 1,280 seconds). Audio was captured at 48 kHz, 24-bit, and trimmed to remove leading and trailing pauses. All signals were time-aligned, producing a synchronized dataset of audio, ECG, GSR, and EEG for multimodal analysis.

D. Feature Extraction

Audio recordings were segmented into 10-second excerpts labeled as “neutral” or “expressive”. Acoustic features captured both spectral and prosodic aspects, including pitch (mean, SD, range), loudness and dynamics (mean, SD, range), spectral centroid and flux, and voice quality measures (jitter, shimmer, HNR, CPP). Formant frequencies (F1–F3) and LTAS metrics summarized vowel resonances and overall spectral content. MFCCs (1–13) with delta coefficients captured fine-grained spectral patterns for timbre and phonetic characterization, while amplitude-in features provided an additional measure of signal energy dynamics.

Physiological signals (EEG, ECG, GSR) were incorporated into the multimodal analysis to complement

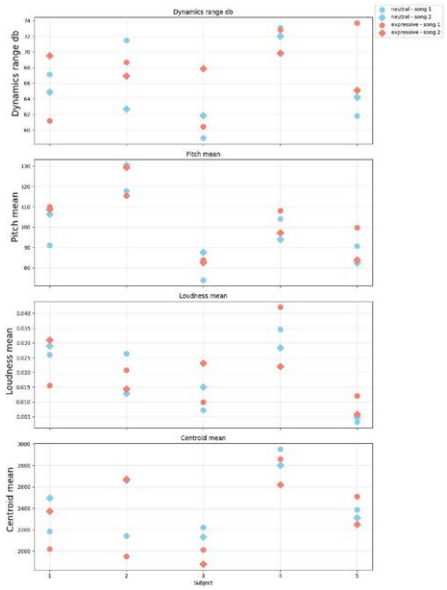


Fig. 1: Audio features extraction: Dynamic, Pitch, Loudness and Centroid

the audio data. EEG was recorded at 250 Hz from four electrodes (O1, O2, T3, T4), band-pass filtered, and analyzed in the alpha (8–13 Hz) and beta (13–30 Hz) ranges. Relaxation indices were derived by computing the proportion of alpha power relative to the combined alpha and beta power [7]. ECG was acquired at 1000 Hz, filtered, and artifact-corrected to extract RR intervals and compute heart-rate variability features using the scipy Python library, including the Baevsky Stress Index [8] as a marker of autonomic regulation. GSR was denoised using median and low-pass filtering (1 Hz) to capture tonic skin conductance fluctuations associated with arousal. Extracted features included heart rate, RR interval statistics, relaxation indices, and mean skin conductance level. To ensure cross-subject comparability, all features were z-normalized at the participant level.

E. Models

We combined statistical analyses with machine learning to compare expressive and neutral performances. Statistical tests revealed significant patterns in audio and physiological features, while supervised models (Random Forests, XGBoost, SVM, Logistic Regression) evaluated their predictive value. This approach balanced interpretability with predictive accuracy.

III. Results

A. Audio Analysis

A comparative analysis of four supervised learning models was conducted (Random Forests, XGBoost, Support Vector Machines and Logistic Regression) to

evaluate their performance in classifying expressive versus neutral speech. Features extracted from each recording were preprocessed with mean imputation for missing values and then split into training and test sets with an 80/20 ratio, maintaining class balance. Each model was trained on the training data and evaluated on the test set, with accuracy, confusion matrices, and classification reports computed. The results were then visually compared (Fig. 2), illustrating the relative performance of the models, where Random Forest and XGBoost demonstrated the highest classification accuracies, followed by Logistic Regression and SVM.

A feature importance analysis was conducted for all classifiers to identify the acoustic features most relevant for distinguishing neutral from expressive singing (Fig. 3). For Random Forest and XGBoost, the built-in feature importance measures were directly used, reflecting how much each feature contributed to the model's predictive power. For SVM and Logistic Regression, permutation importance was employed, assessing the impact of randomly shuffling each feature on model accuracy. A Random Forest classifier was trained with 5-fold cross-validation on the full dataset. List-like features were averaged, and missing values were imputed with column-wise means. Cross-validated predictions yielded an overall accuracy of 82.1% as shown in Table I, with detailed metrics and a confusion matrix showing the model's ability to distinguish neutral from expressive singing. After cross-validation, the model was trained on the full dataset to extract feature importance scores, revealing which acoustic characteristics—such as pitch, loudness, MFCCs, and spectral features—most strongly influenced model predictions.

B. Multimodal analysis

Since the vocal recordings were collected as part of the witheFlow project, which included both singers and instrumentalists, we adopted a multimodal analysis approach to ensure consistency and enable direct comparison across the dataset. For the physiological signals,

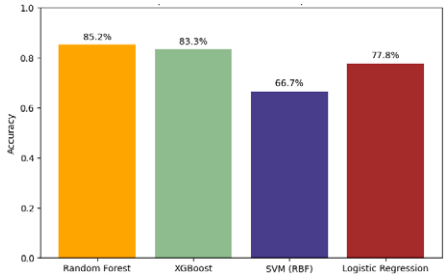


Fig. 2: Comparison of classification performance of four supervised learning models: Random Forest, XGBoost, Support Vector Machine (RBF kernel), and Logistic Regression.

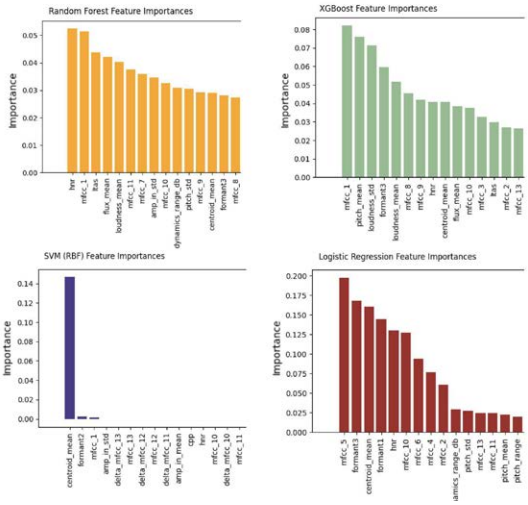


Fig. 3: Feature importance scores for all the classifiers.

features such as mean skin conductance level and heart rate variability were computed over the entire song duration. To identify the most informative features, we combined statistical tests (paired t-tests for audio and physiological modalities, chi-square tests for categorical emotion labels) with machine learning analysis.

The results revealed clear differences between expressive and non-expressive performances. In the audio domain, onset values were higher in expressive performances, whereas pulse clarity was higher in non-expressive ones. Physiological measures indicated greater EEG-based relaxation in performers during expressive performances, and lower heart rates in audience members during non-expressive ones (see Fig. 4). Self-reported emotions further showed that performer–audience alignment emerged only under expressive conditions, consistent with prior research on musical expressivity [9].

For multimodal analysis, we used XGBoost [10], imputing missing values with condition-wise means. Separate models for performers and audience predicted expressive vs. neutral conditions, reaching accuracies with 5-fold cross validation of 85% and 65%, respectively. As shown in Fig. 5, onset-related and biosignal features were most informative, consistent with statistical results.

TABLE I: Random Forest Classification Results (5-Fold Cross-Validation)

Class	Precision	Recall	F1-Score	Support
Neutral	0.83	0.80	0.81	131
Expressive	0.82	0.84	0.83	137
Accuracy			0.82	268
Macro Avg	0.82	0.82	0.82	268
Weighted Avg	0.82	0.82	0.82	268

IV. Discussion

Across models, audio features such as MFCCs, spectral centroid, pitch, and loudness consistently emerged as informative predictors of expressive performance. Random Forest and XGBoost highlighted a broader range of features, with MFCCs and pitch among the most relevant and Logistic Regression emphasized MFCCs and formants, supporting their role in characterizing expressive nuances. Acoustic and behavioral features obtained from the multimodal analysis were consistent with the broader dataset collected within the witheFlow project [9], which included 16 additional instrumentalists recorded under an extended protocol, allowing for direct comparison with the current study. In the expressive condition, both instrumentalists and vocalists tended to use a greater number of notes in their performances. Performers appeared more relaxed under expressive conditions, whereas the audience exhibited higher arousal, reflected in elevated heart rates. Moreover, an alignment of emotional responses between performers and audience was observed only in the expressive condition.

While these findings offer valuable insights, they should be interpreted with caution due to the relatively small sample size, which may limit the generalizability of the results. Moreover, proper stratification is essential to prevent models from learning speaker-specific characteristics rather than the intended patterns. Future studies with larger and more diverse samples are needed to validate and extend these observations.

Ethics Statement: This study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki. All participants provided informed consent prior to their participation, and their anonymity was maintained throughout the study. Ethical approval was obtained from the local ethics committee of the National Technical University of Athens.

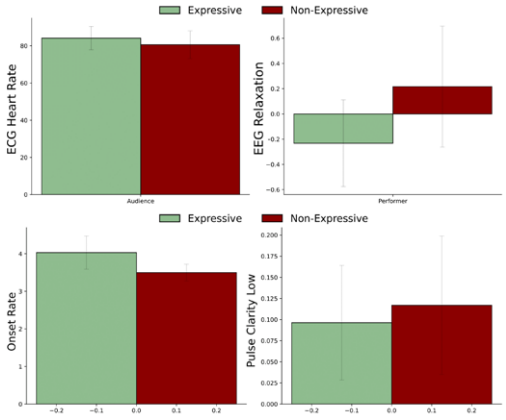


Fig. 4: Audio (bottom) and physiological (top) features that are statistically different between the conditions.

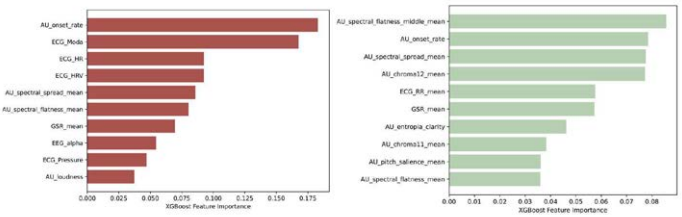


Fig. 5: Top 10 Important Features of the XGBoost Classifier for the Audience (left) and the Performers (right).

Acknowledgments: We thank the volunteer participants for their valuable contribution to this work. The research project is implemented in the framework of H.F.R.I call "Basic research Financing (Horizontal support of all Sciences)" under the National Recovery and Resilience Plan "Greece 2.0" funded by the European Union –NextGenerationEU (H.F.R.I. Project Number: 15111 - Emotional Artificial Intelligence in Music Expression).

References

[1] J. Sundberg, F. M. Lã, and E. Himonides, "Intonation and expressivity: a single case study of classical western singing," *Journal of Voice*, vol. 27, no. 3, pp. 391–e1, 2013.

[2] G. Kotlyar and V. Morozov, "Acoustical correlates of the emotional content of vocalized speech," *Soviet Physics. Acoustics*, vol. 22, no. 3, pp. 208–211, 1976.

[3] J. Sundberg, J. Iwarsson, and H. Hagegård, "A singer's expression of emotions in sung performance," *Vocal fold physiology: Voice quality control*, pp. 217–229, 1995.

[4] H. Siegwart and K. R. Scherer, "Acoustic concomitants of emotional expression in operatic singing: the case of lucia in ardi gli incensi," *Journal of Voice*, vol. 9, no. 3, pp. 249–260, 1995.

[5] J. Sundberg, G. L. Salomao, and K. R. Scherer, "Analyzing emotion expression in singing via flow glottograms, long-term-average spectra, and expert listener evaluation," *Journal of Voice*, vol. 35, no. 1, pp. 52–60, 2021.

[6] P.-O. Jacobsen, H. Strauss, J. Vigl, E. Zangerle, and M. Zentner, "Assessing aesthetic music-evoked emotions in a minute or less: A comparison of the gems-45 and the gems-9," *Musicae Scientiae*, p. 10298649241256252, 2024.

[7] K. Sugimoto, H. Kurashiki, Y. Xu, M. Takemi, and K. Amano, "Electroencephalographic biomarkers of relaxation: A systematic review and meta-analysis," *bioRxiv*, March 2024.

[8] T. K. Sahoo, A. Mahapatra, and N. Ruban, "Stress index calculation and analysis based on heart rate variability of ecg signal with arrhythmia," in *2019 Innovations in Power and Advanced Computing Technologies (i-PACT)*, vol. 1, pp. 1–7, IEEE, 2019.

[9] V. Lyberatos, S. Kantarelis, I. Zioga, C. Anagnostopoulou, G. Stamou, and A. Georgaki, "Music interpretation and emotion perception: A computational and neurophysiological investigation," in *Proceedings of the 22nd Sound and Music Computing Conference*, 2025.

[10] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

TO CREAK OR NOT TO CREAK, THAT IS THE QUESTION

N. Henrich Bernardoni[1], A. Ménard[2], T. Linke[3], A. Katriou [4], M. Girod-Roux[1], I. Atallah [1]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP-UGA, GIPSA-lab, 38000 Grenoble, France
[2] Sillingy, France
[3] Leipzig, Germany
[4] Studio Aliki Katriou Vocals, Opatija, Croatia

Nathalie.Henrich@gipsa-lab.fr

Abstract: **This study questions the notion of creaking in singing. A variety of distorted sounds have been recorded by two singers who are also singing teachers from the international Sing&Scream team. Based on audio and electroglottographic recordings together with endoscopic views, it is shown that creaking is a vocal effect that can be added to sung sounds, independently of the laryngeal mechanism in use. While glottal pulse train and acoustical signal are greatly modified by adding a creaking effect to a sustained sound, the supraglottic adjustments evidenced on endoscopic videos are subtle. Lateral supraglottic constriction seems to be a control mechanism to introduce creaking and increase its degree of irregularity. Pharyngeal constriction has also been evidenced in female singing.**

Keywords: **creaking, pulse, fry, distortion, supraglottic**

I. INTRODUCTION

Creaky voice is a human voice production found in speech and singing (see [1,2] for a review). Its prototypical form has been characterized by low rate of vibration, irregular glottal pulses and long closed phases during glottal cycle. In phonetics, it is classified as a phonation type. Together with modal voice and breathy voice, it is among the most common phonation types, serving as a contrastive cue in many languages [3,4]. In singing, it has barely been studied. Two recent studies on vocal effects in complete vocal technique has addressed creaking as a rough vocal effect used in modern singing [5,6]. From an acoustical and physiological point of view, creaky voice originates at the glottis [1,4,5]. Supraglottic articulations (aryepiglottic constriction, pharyngeal narrowing, tongue root retraction) may often accompany or reinforce it. These adjustments may help to stabilize the low-frequency vibration. They are reflected in the acoustic signal through the appearance of subharmonic components and/or increased noise [1].

The aim of this study is to advance the understanding of creaking in singing, specifically within the context of vocal distortion. Using audio, electroglottographic, and videoendoscopic recordings of transitions between non-creaky and creaky phonation, we seek to understand how this effect is controlled in singing.

II. METHODS

Subjects: Seven singers (3F,4M) participated in the study. They are part of an internationally-recognized singing teacher team (Sing&Scream - singandscream.fr, singandscream.com/colleagues/). In this paper, we present the results from two of them (1F, 1M) who were proficient in producing creaking.

Protocol: The protocol was elaborated collaboratively between a voice scientist and two singing teachers. Singing tasks consisted of sustained sounds and glides, and different types of screams. To explore the physiological properties of creaky voice, a part was dedicated to add or remove creaking from target sounds: a sustained low-pitch note in modal voice (G3), a high-pitch note in falsetto/head voice (G4), and descending-ascending glides from modal to pulse register.

Database: Audio and electroglottographic (EGG) signals were recorded synchronously with an EG-2 Glottal-Enterprise electroglottograph combined to a computer. They were sampled at 48 kHz. Endoscopy videos were recorded simultaneously at an image frame rate of 25 fps and audio sampling rate of 44.1 kHz. They were synchronized in post-processing using the resampled audio signal at 48 kHz.

Data analysis: The database annotation was done manually by singing teachers with Praat software [8]. All post-processing and analyses were carried out using Matlab (R2023a, The MathWorks, Natick, MA, USA).

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

Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

III. RESULTS

A. Creaking in pulse register

Figure 1 presents six samples from a male and a female singer singing in pulse register with and without creaking. Without creaking, each glottal cycle is characterized by a long closed phase and a damped oscillation, as expected for such very-low-pitch sounds (less than 35 Hz for the samples shown here). Endoscopic recordings show a reduced vocal-fold vibrating length and an anterior ventricular narrowing, more pronounced in male than in female singer. These features evidence a production in laryngeal mechanism M0 [9,10]. For creaking productions in pulse register, the singers make a distinction between regular creaking and irregular creaking, both being consciously controlled by them and perceptually distinguishable. In regular creaking samples, a dicrotic vibratory pattern is observed on EGG signal, with two pulses closely spaced in time during one glottal cycle. In irregular creaking, pulses occur at irregular (non-periodic) intervals, with or without dicrotic patterns, and no long closed phase is observed. Glottal length is increased, reflecting a reduction in the antero-posterior supraglottic compression. Even if the image frame rate limits the observation of glottal opening, changes can be observed between the samples, with opening confined to a limited region of the mid-portion of the vocal folds in non-creaking and regular creaking cases, and extending over a larger region when irregular creaking occurs. For the female singer, singing in pulse register with irregular creaking is associated with a posterior glottal chink, similarly to that observed during her production of modal and falsetto registers without creaking.

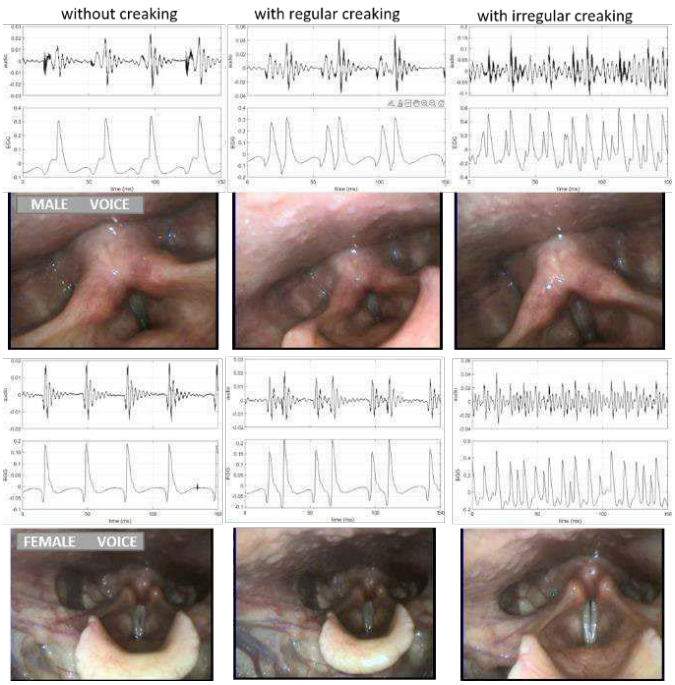


Figure 1: Audio, EGG and endoscopic view of singing samples sung in pulse register by a male singer (top) and a female singer (bottom), without (left panels) and with creaking (middle/right panels for regular/irregular sound productions).

B. Creaking in modal register

For singers mastering vocal distortion, creaking is not only a property related to pulse register. They are able to add creaking sound quality while singing higher pitches, in laryngeal mechanisms M1 or M2. Figures 2 and 3 illustrate the sounds produced by a male and a female singer while singing on the same pitch without or with creaking. All sounds are produced in laryngeal mechanism M1 at the same pitch of 196 Hz (note G3).

The singers are able to intentionally control the creaking effect, producing it in a regular or irregular manner. The main finding in the endoscopic videos is a subtle increase in lateral supraglottic constriction accompanying the transition from no creaking to creaking. The greater the lateral supraglottic constriction, the more irregular the creaking. In the male voice, a convergence of the cuneiform cartilages and an increase in antero-posterior supraglottic constriction are also observed. A regular creaking is characterized for

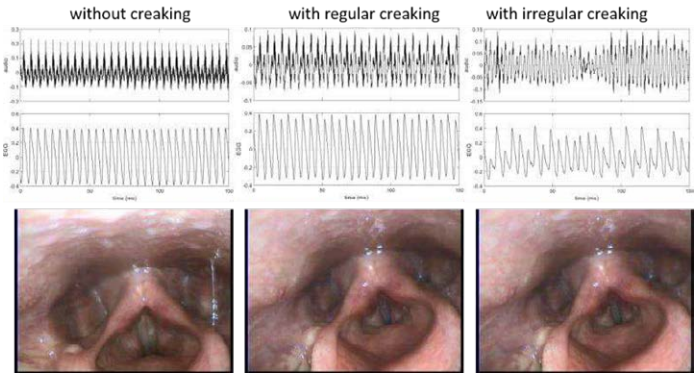


Figure 2 : Audio, EGG and endoscopic view of a sound in M1 sung by a male singer without (left panel) or with creaking (middle panel : regular, right panel : irregular). Pitch G3 (196 Hz).

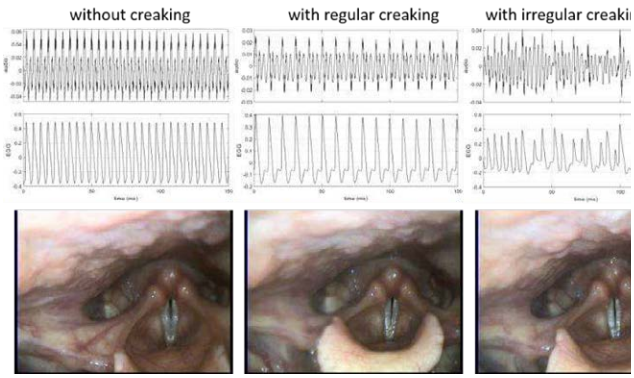


Figure 3 : Audio, EGG and endoscopic view of a sound in M1 sung by a female singer without creaking (left panel) or with creaking (middle panel : regular, right panel : irregular). Pitch G3 (196 Hz).

both singers by an alteration in glottal contact area reflected on EGG signal. Whereas the second glottal contact is slightly attenuated for the male singer, it is almost completely attenuated for the female one.

C. Creaking in falsetto/head register

When singing at higher pitches, and in particular an octave above (392 Hz), both singers can use laryngeal mechanism M2. Singing in falsetto or head register does not prevent them from adding a creaking effect on the sound. Similarly to M1, the singers are able to adjust the regularity of their production, as shown in Figures 4 and 5.

For both singers, the degree of creaking is adjusted by a lateral supraglottic constriction. In the female case (see Figure 5), this constriction is greater in regular than in irregular creaking. A constriction of the pharyngeal space is also evidenced for her creaking, reflecting the contraction of her constrictor muscles.

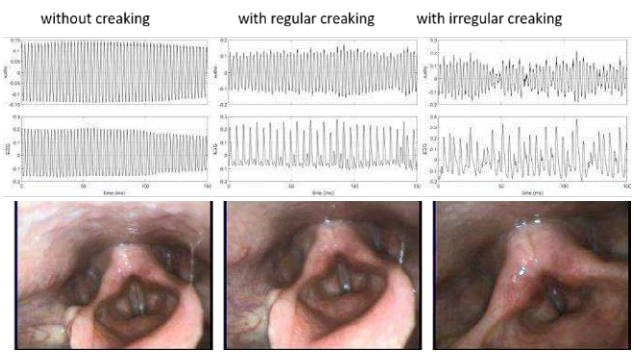


Figure 4 : Audio, EGG and endoscopic view of a sound in M2 sung by a male singer without creaking (left panel) or with creaking (middle panel : regular, right panel : irregular). Pitch G4 (392 Hz).

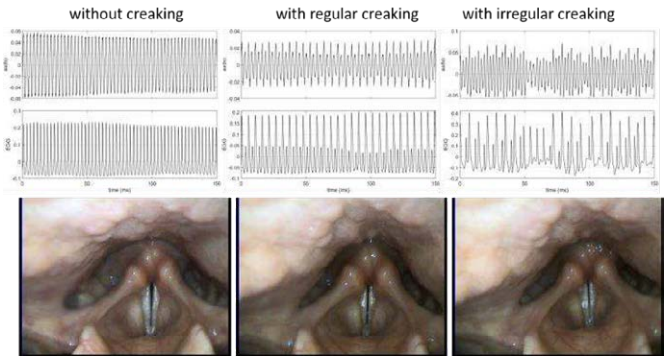


Figure 5 : Audio, EGG and endoscopic view of a sound in M2 sung by a female singer without creaking (left panel) or with creaking (middle panel : regular, right panel : irregular). Pitch G4 (392 Hz).

IV. Discussion

The observations presented in the Results part demonstrate that creaking is an effect that can be added to sung sounds, regardless of the laryngeal mechanism used. The mastering of this effect by both male and female singers has been illustrated.

Why does it creak? Creaking is first of all a perceptual effect added to the sound by the singer, changing a smooth quality into a rougher one. It also reflects a physiological modification in vocal-fold vibratory pattern and glottal configuration, assessed by changes in vocal-fold contact area detectable on EGG signals. In all regular creaking cases, dicrotic vibratory patterns have been evidenced. This feature, which is commonly found in vocal fry [e.g. 11,12] is also mastered here in modal, falsetto and head register singing. Irregular creaking cases demonstrate a disruption of periodicity and predictability in vocal-fold contact area. These observations call for further study with highspeed endoscopy. Such irregular vibratory behavior has been observed in singing, yet most often related to interactions with supraglottic structures. The supraglottic adjustments evidenced here are surprisingly subtle in contrast to the strikingly erratic behavior in measured glottal contact. These observations provide evidence of nonlinear dynamic behavior in singing [13].

How to control creaking? Our data indicate that lateral supraglottic constriction is a primary control mechanism. Other potential control mechanisms may exist, notably aerodynamic ones. They could not be evaluated in the present study, but they undoubtedly need to be considered. Constriction of the lateral pharyngeal wall has been observed with creaking. While evident in female singing, it was not observed in male singing, where the singer has learned to perform both with and without pharyngeal constrictor engagement.

V. Conclusion

In exploring the physiological and acoustical correlates of singing production without and with creaking, we have evidenced that creaking is an effect that can be added to any voiced sound in singing. The laryngeal mechanism in use does not prevent from adding this effect. The regularity of creaking can be deliberately adjusted by the singer. Control mechanisms at the laryngeal level (as evidenced by endoscopic imaging and electroglottographic analysis) are manifested at the ventricular level by increased lateral supraglottic constriction. Additional constrictions have been observed at the pharyngeal level. Another level of control is certainly the degree of subglottal pressure, which is not assessed here but calls for further research.

Acknowledgment

This work is supported by the French National Research Agency in the framework of the "Investissements d'avenir" program (ANR-15-IDEX-02) and ANR AVATARS (ANR-22-CE48-0014).

References

[1] P. Keating, M. Garellek, and J. Kreiman, "Acoustic properties of different kinds of creaky voice," in *Proc. ICPhS*, 2015.

[2] K. Dallaston and G. Docherty, "The quantitative prevalence of creaky voice (vocal fry) in varieties of English: A systematic review of the literature," *PLoS ONE*, vol. 15, no. 3, p. e0229960, 2020.

[3] M. Gordon and P. Ladefoged, "Phonation types: a cross-linguistic overview," *Journal of Phonetics*, vol. 29, no. 4, pp. 383–406, 2001.

[4] M. Garellek, "Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality," *Journal of Phonetics*, vol. 94, p. 101155, 2022.

[5] M. Aaen, J. McGlashan, and C. Sadolin, "Laryngostroboscopic Exploration of Rough Vocal Effects in Singing and their Statistical Recognizability: An Anatomical and Physiological Description and Visual Recognizability Study of Distortion, Growl, Rattle, and Grunt using laryngostroboscopic Imaging and Panel Assessment," *Journal of Voice*, vol. 34, no. 1, p. 162.e5-162.e14, 2020.

[6] M. Aaen, J. McGlashan, N. Christoph, and C. Sadolin, "Extreme Vocal Effects Distortion, Growl, Grunt, Rattle, and Creaking as Measured by Electroglottography and Acoustics in 32 Healthy Professional Singers," *Journal of Voice*, vol. 38, no. 3, p. 795.e21-795.e35, 2024.

[7] J. A. Edmondson and J. H. Esling, "The Valves of the Throat and Their Functioning in Tone, Vocal Register and Stress: Laryngoscopic Case Studies," *Phonology*, vol. 23, no. 2, pp. 157–191, 2006.

[8] P. Boersma and P. Weenink, "Praat: doing phonetics by computer [Computer program]," *Version 6.4.43, retrieved 14 September 2025 from http://www.praat.org*, 2025.

[9] B. Roubeau, N. Henrich, and M. Castellengo, "Laryngeal Vibratory Mechanisms: The Notion of Vocal Register Revisited," *J. of Voice*, vol. 23, no. ., pp. 425–438, 2009.

[10] L. Bailly, N. Henrich Bernardoni, F. Müller, A.-K. Rohlfs, and M. Hess, "Ventricular-fold dynamics in human phonation," *J Speech Lang Hear Res*, vol. 57, no. 4, pp. 1219–1242, 2014.

[11] M. Blomgren, Y. Chen, M. L. Ng, and H. R. Gilbert, "Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers," *The Journal of the Acoustical Society of America*, vol. 103, no. 5, pp. 2649–2658, 1998.

[12] Y. Chen, M. P. Robb, and H. R. Gilbert, "Electroglottographic Evaluation of Gender and Vowel Effects During Modal and Vocal Fry Phonation," *J Speech Lang Hear Res*, vol. 45, no. 5, pp. 821–829, 2002.

[13] J. G. Švec and Z. Zhang, "Application of nonlinear dynamics theory to understanding normal and pathologic voices in humans," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 380, no. 1923, p. 20240018, 2025.

SESSION III
VOICE ANALYSIS AND SYNTHESIS

ABOUT THE EXCESS VARIABILITY OF POPULAR ACOUSTIC FEATURES OF VOCAL JITTER AND SHIMMER

J. Schoentgen[1], A. Kacha[2] , F. Grenez[1]

[1] Université Libre de Bruxelles, Brussels, Belgium
[2] University of Jijel, Jijel, Algeria
jean.schoentgen@icloud.com, grenez.francis@ulb.be, Abdelha.Kacha@ulb.be

Abstract: **The topic concerns the inconsistent definition and excess variability of acoustic features of jitter and shimmer that report perturbations of vocal cycle lengths and amplitudes sampled once per vocal cycle. We examine the frequency bands and gains as well as the correlation with vocal frequency and the intra-corpus variability of several popular jitter and shimmer features. The results indicate that a discrepancy exists between the frequency bands of the features and the actual frequency bands of jitter and shimmer, and that this discrepancy varies with vocal frequency. The perturbations reported by various features in different frequency bands cannot be compared owing to differences in feature gains. The excess volatility attributable to sampling once per vocal cycle tends to be more readily detectable in corpora exhibiting substantial variability in vocal frequency, particularly in features that are not normalized by the average vocal cycle duration.**

Keywords: **vocal jitter, vocal shimmer, frequencies of vocal jitter and shimmer, variable-rate sampling**

I. INTRODUCTION

Vocal jitter and shimmer designate fast (>20Hz) perturbations of the vocal cycle lengths and peak amplitudes. The presentation focusses on the excess variability of acoustic features that describe vocal perturbations when the vocal cycle lengths or peak amplitudes are sampled once per vocal cycle.

The features indeed involve high-pass filtering of the lengths or amplitudes to separate fast perturbations (i.e. jitter and shimmer) from slow perturbations (i.e. drift, tremor and flutter) [1,2]. The frequency bands of the features therefore depend on the vocal frequency because the analog cut-off frequency of a discrete filter depends on the sampling frequency. The feature values evolve for that reason with the vocal frequency due to the requirements of signal processing alone, loose from any physiological causes.

The two following issues are related to the former and must be examined jointly. One issue is that the filters involved in different features do not have unity gain, which alters the reported size of jitter or shimmer and makes it difficult to compare values across different features. A second issue is the mismatch between the frequency bands of the features and the actual frequency bands of jitter and shimmer. This mismatch may weaken the perceptual relevance of the features and cause the actual size of the perturbations to be over or underestimated.

These topics are relevant because over the past fifty years numerous articles have been published that document acoustic perturbations that have been tracked at the rate of the vocal frequency, e.g. [3]. The features reported by the PRAAT and MDVP speech analysis software programs are the most popular. We therefore examine five jitter and six shimmer features obtainable via PRAAT and which are known as $Local_{jit}$, $Local_{abs}$, *RAP*, *PPQ5* and DDP_{jit} as well as $Local_{shim}$, $Local_{dB}$, *APQ3*, *APQ5*, *APQ11* and DDP_{shim} respectively [2]. In the text, PRAAT features are referred to as variable-rate non-unity gain features.

Table 1 shows the six high-pass filters used for extracting jitter and shimmer features in PRAAT. The formulas are copied from the manual [2]. In Table 1, symbol X_i designates the cycle lengths or peak amplitudes and Δ_i the perturbations. The features report in percent the average unsigned perturbation $|\Delta|$ divided by the average cycle length or amplitude.

Exceptions are $Local_{abs}$ and $Local_{dB}$. $Local_{abs}$ provides the average unsigned perturbation in μsec. $Local_{dB}$ reports in *dB* the average unsigned log-ratio of the amplitudes of two consecutive vocal cycles A_i and A_{i+1}. However, the first-order Taylor series expansions of $\log(A_{i+1}/A_i)$ and $1/A_i$ show that $Local_{shim}$ and $Local_{dB}$ are equivalent up to a difference in gain, which is confirmed by inspecting their numerical values.

The high-pass filter gain refers to the magnitude of the filter at half the sampling frequency. The filter cut-off refers to the relative frequency at which the logarithm of the magnitude squared decreases by -3dB. Gain and cut-off are calculated numerically.

We compare variable-rate non-unity gain features (i.e. PRAAT features) to fixed-rate unity-gain features that describe perturbations of vocal cycle lengths and amplitudes that are sampled at a fixed rate. They isolate fast from slow perturbations at a fixed frequency of 20Hz, which is perceptually relevant. The

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

switch from the perception of tremulousness to the perception of roughness is indeed expected to occur in the vicinity of that frequency [4].

The results section presents the distance between variable-rate and fixed-rate features, the intra-corpus variability of the features and their correlation with vocal frequency.

Table 1: Discrete high-pass filters involved in popular jitter and shimmer features as well as their gain at $f/f_s = 0.5$ and their relative cut-off frequency f_c/f_s at -3dB. f_c designates the analog cut-off frequency, f_s the sampling frequency i.e. vocal frequency f_o and X_i the cycle length T_i or peak amplitude A_i.

Features	Filter	Gain	Cut-off
$Local_{jit}$ $Local_{shim}$ $Local_{abs}$	$\Delta_i = X_i - X_{i-1}$	2	0.25
$Local_{dB}$	$\Delta_i = X_i - X_{i-1}$	40/ln(10)	0.25
RAP *APQ3*	$\Delta_i = X_i - (X_{i-1} + X_i + X_{i+1})/3$	4/3	0.32
PPQ5, APQ5	$\Delta_i = X_i - (X_{i-2} + \cdots + X_{i+2})/5$	4/5	0.13
DDP_{jit} DDP_{shim}	$\Delta_i = -X_i + (X_{i-1} - X_i + X_{i+1})$	-4	0.32
APQ11	$\Delta_i = X_i - (X_{i-5} + \cdots + X_{i+5})/11$	10/11	0.06

II. METHODS

A. Corpora

We have four corpora of 2-second fragments of sustained vowel [a] sampled at a rate of 44kHz. All vowel sounds are pseudo-periodic and monophonic. Corpora I and II include 18 male and 18 female speakers respectively. They have been downloaded from the website of the ATIC Research Group, Dept. Ingenieria de Comunicaciones, Universidad de Malaga [5]. Corpora III and IV comprise 33 female teachers recorded pre and post vocal loading. They have been provided by the Speech Therapy Department, Faculty of Psychology, Speech Therapy and Education Sciences, University of Liège, BE.

Since signal processing is the focus of the presentation, corpora I–IV have been aggregated into corpora *Pooled*, *Hi* and *Lo*, each with a distinct average and range of vocal frequencies with the goal to have an f_o −range that is as large as possible. The corpus *Pooled* includes 18 ♂ and 84 ♀ speakers, *Lo* and *Hi* break up *Pooled* according to whether *fo* is smaller or larger than the mid-frequency of the f_o-range of *Pooled.* Table 2 presents, for each corpus, the number of vowel signals, the median and interquartile range of *fo*, and the median and interquartile range of the frequency of jitter. The bottom of Table 2 shows the median analog cut-off frequency for each corpus and for each relative cut-off frequency.

Table 2: Test corpora. Top, left to right: number of vowel signals, median and interquartile range of vocal frequency and jitter frequency. Bottom: median *analog* cut-off frequencies for each corpus and *relative* cut-off frequency.

Tally, median, inter-quartile range	nbr	*fo* med (IQR)	f_{jit} med (IQR)
Pooled	102	196 (49) Hz	61 (17) Hz
Lo (< 187.5 Hz)	39	159 (71) Hz	49 (21) Hz
Hi (> 187.5 Hz)	63	217 (27) Hz	67 (10) Hz

Rel. cut-off freq.	0.06	0.13	0.25	0.32
Pooled	12 Hz	25 Hz	49 Hz	63 Hz
Lo (< 187.5 Hz)	10 Hz	21 Hz	40 Hz	51 Hz
Hi (> 187.5 Hz)	13 Hz	28 Hz	54 Hz	69 Hz

B. Signal processing

The 2-second vowel fragments have been up-sampled four times to improve temporal resolution. The signals have subsequently been low-pass filtered at 1kHz by means of a linear-phase Bessel filter to avoid spurious perturbations owing to the superposition of the first and second formants. Filtering also removes the contribution of the higher formants to vocal shimmer. Filtering does not typically influence genuine jitter.

The positive and negative cycle peaks have been detected via a conventional peak picker. The negative and positive peak amplitudes, along with the intervals between consecutive peaks, have been assigned to two amplitude and two cycle length time series. The length series with the lowest standard deviation, along with its corresponding amplitude series, has been kept. The PRAAT features have been calculated using the formulas in Table 1. The cycle amplitudes or lengths provided by PRAAT have not been retained because PRAAT does not track the actual cycle peaks.

Fixed-rate unity-gain features have been obtained as an alternative. They report vocal perturbations faster than 20Hz whose shape and energy are preserved. The signal processing involves resampling the time series at a constant rate of 800Hz using a cubic spline interpolator, followed by subtracting the average length or amplitude to obtain the raw perturbation time series.

An orthogonal cosine transform breaks up the raw perturbations into jitter or shimmer in the frequency band $20Hz \rightarrow f_o/2$, as well as into drift, tremor, and flutter in the frequency band $0 \rightarrow 20Hz$ [1,4]. Up-sampling, filtering, peak-picking, interpolating and cosine transform routines are included in Python's standard signal processing toolbox.

The fixed-rate unity-gain features (abbr. FR-UG) report, in analogy with PRAAT, the average of the unsigned values of the jitter or shimmer time series, divided by the average cycle length or amplitude.

The frequencies of jitter and shimmer have been empirically estimated by counting the unidirectional zero-crossings of the corresponding time series. The jitter frequencies are summarized in Table 2. The median shimmer frequencies are circa 16% lower.

The sum of the variances of the perturbations below and above 20Hz equals the variance of the raw perturbations because the cosine transform is orthogonal. The perturbations are neither attenuated nor boosted. Also, fast perturbations are separated from slow perturbations at a fixed frequency equal to 20Hz. Fixed-rate unity-gain features are therefore targets to which variable-rate non-unity gain features may be compared.

C. Descriptive statistics

A measurement error is the distance between a measured value and a target value. The results section therefore reports the distance between fixed-rate and variable-rate features. The feature values, however, cannot be compared directly because they pertain to different gains and frequency bands. The values have therefore been replaced by their ranks, which means that features are regarded to be equivalent when they rank speakers identically.

Spearman's *F-distance* has been used to report the difference between the ranks of variable-rate and fixed-rate features. Spearman's F is the normalized sum of the unsigned differences between ranks. It has a value between zero and one [6]. The average difference in number of ranks is equal to $F \times N/2$ when the number N of signals is large (e.g. N>10). The *F-distance* is not affected by transforms of the feature values that preserve their order (e.g. multiplication by different gains).

The results section also reports the relative interquartile range as well as the Pearson correlation with vocal frequency f_o for each corpus and each feature. Indeed, intra-corpus variability and magnitude of the correlation with f_o are expected to be lower for fixed-rate than for variable-rate features when they share roughly the same frequency band.

III. RESULTS

Feature pairs $Local_{shim}$ and $Local_{dB}$, *APQ3* and DDP_{shim} as well as *RAP* and DDP_{jit} only differ by a constant gain (Table 1). The features $Local_{dB}$, DDP_{shim}, and DDP_{jit} are therefore omitted from the Results section.

A. Intra-corpus relative interquartile range and correlation with vocal frequency

Tables 3 and 4 report the Pearson correlation with f_o and the relative interquartile range of five shimmer and five jitter features for three corpora. *FR-UG* designates the fixed-rate features.

Table 3: Correlation with *fo* (top) and relative interquartile range (bottom) of five shimmer features for three corpora. $FR\text{-}UG_{shim}$ is the fixed-rate feature.

	$Local_{sh}$	*APQ3*	*APQ5*	*APQ11*	$FR\text{-}UG_{shim}$
Pooled	-0.59	-0.54	-0.63	-0.63	-0.35
Lo	-0.63	-0.58	-0.63	-0.72	-0.43
Hi	-0.34	-0.36	-0.40	-0.22	-0.10
Pooled	0.69	0.74	0.72	0.62	0.55
Lo	0.61	0.71	0.65	0.56	0.56
Hi	0.59	0.77	0.57	0.55	0.60

Table 4: Correlation with *fo* (top) and relative interquartile range (bottom) of five jitter features for three corpora. $FR\text{-}UG_{jit}$ is the fixed-rate feature.

	$Local_{jit}$	*RAP*	*PPQ5*	$Local_{abs}$	$FR\text{-}UG_{jit}$
Pooled	-0.32	-0.21	-0.38	-0.74	-0.17
Lo	-0.29	-0.14	-0.32	-0.73	-0.02
Hi	-0.20	-0.15	-0.29	-0.40	-0.30
Pooled	0.63	0.58	0.65	0.78	0.60
Lo	0.56	0.59	0.65	0.94	0.56
Hi	0.56	0.57	0.62	0.70	0.65

B. Inter-feature distance

Table 5 shows the *F-distance* of four shimmer and four jitter features for three corpora. Spearman's F is the relative distance between the ranks of features $FR\text{-}UG_{shim}$ or $FR\text{-}UG_{jit}$ and the ranks of the PRAAT features. f_o-rate sampling, normalization and unequal frequency bands are expected to affect *F-distance*. In contrast, the feature-typical gain has no direct influence on *F-distance*, intra-corpus variability or correlation with f_o.

Table 5: *F-distance* between fixed-rate and four shimmer (top) and four jitter (bottom) variable-rate features for three corpora.

	$Local_{shim}$	*APQ3*	*APQ5*	*APQ11*
Pooled	0.19	0.23	0.16	0.17
Lo	0.19	0.20	0.15	0.24
Hi	0.18	0.24	0.17	0.09
	$Local_{jit}$	*RAP*	*PPQ5*	$Local_{abs}$
Pooled	0.12	0.11	0.10	0.27
Lo	0.17	0.11	0.16	0.37
Hi	0.07	0.10	0.05	0.10

IV. DISCUSSION

The absolute jitter feature $Local_{abs}$ is an exception because it reports the average unsigned perturbation in μsec. All other features are normalized and describe perturbations in *percent*.

Differences between variable-rate and fixed-rate features are explained by normalization, unequal

frequency bands and inflated variability, which are discussed separately.

T_o-normalization: Magnitude of f_o-correlation, relative interquartile range and *F-distance* are larger for variable-rate features of shimmer than for variable-rate features of jitter that share the same relative cut-off frequency (Tables 3 to 5). The normalization by the average cycle length T_o indeed reduces the excess variability owing to f_o-rate sampling. This explanation is confirmed by the distinctive properties of $Local_{abs}$, which lacks that normalization (Tables 4 and 5).

As a rule, the purpose of normalization is to cancel incidental differences between signals or corpora. T_o-normalization of variable-rate features is indeed beneficial because of a weakness in signal processing. T_o-normalization of fixed-rate features, however, may disguise a genuine physiological dependence on vocal frequency.

Feature-typical vs actual perturbation frequency bands: The frequencies of jitter and shimmer, which are similar, have been estimated by counting unidirectional zero-crossings in the jitter or shimmer time series after re-sampling at a fixed rate. The frequencies of jitter are summarized in Table 2 (top, right) together with the analog cut-off frequencies of the PRAAT features (bottom).

Table 2 shows that the *median* analog cut-off frequencies of feature *APQ11* are lower than 20Hz. This means that *APQ11* includes flutter and therefore overestimates shimmer. Conversely, the *median* analog cut-off frequencies of *RAP* and *APQ3* are higher than the *median* frequencies of jitter. This means that *RAP* and *APQ3* may miss half of the frequency band of the actual perturbations and therefore underestimate jitter and shimmer. The mismatch between frequency bands is also noticeable in $Local_{jit}$ and $Local_{shim}$ for high-pitched voices.

One may conclude that the frequency bands of PRAAT features do not match the actual frequency bands of jitter or shimmer. This mismatch, however, is not passed on pro rata to the feature values because actual differences in size are masked by unequal feature gains (Table 1). Also, the mismatch is made worse by f_o-rate sampling because a match at low f_o may turn into a mismatch at high f_o or vice versa.

Variable-rate vs fixed-rate features: Intra-corpus variability and magnitude of correlation with f_o are expected to be lower for fixed-rate features because of the absence of the inflation of the variability owing to f_o-rate sampling. This is confirmed for corpora *Pooled* and *Lo*, but not for corpus *Hi* (Tables 3 and 4). Possible explanations are the low intrinsic f_o-variability of corpus *Hi* and the discrepancy, which is largest for high-pitched voices, between the frequency bands of variable-rate and fixed-rate features, together with a mis-representation owing to T_o-normalization, of the correlation with f_o of fixed-rate features (Table 2).

Table 5 shows the relative distance in number of ranks between variable-rate and fixed-rate features. The *F-distance* appears to evolve with the relative cut-off frequency and decreases with T_o-normalization.

Irrespective of the value of the distance between variable-rate and fixed-rate features, the merit of fixed-rate features is that they report the actual size of perturbations sampled at a fixed rate in a frequency band that covers the actual frequency band of jitter and shimmer. The expected benefit is the ease of the numerical or perceptual interpretation.

V. CONCLUSION

The shortcomings of popular variable-rate features consist in the mismatch between the frequency bands of the features and the actual frequency bands of jitter and shimmer. This mismatch evolves with the vocal frequency. Any disparity between frequency bands is not relayed pro rata to the feature values because unequal feature gains mask the actual size of the length or amplitude perturbations. The excess variability *per se* owing to f_o-rate sampling appears to be more readily observable in corpora with a large intrinsic f_o-variability and in features that are not T_o-normalized.

REFERENCES

[1] Buder E. H., and Strand, E. A. (2003). “Quantitative and graphic acoustic analysis of phonatory modulations: the modulogram,” J. Speech, Language and Hearing Res. 46, 475–490.

[2] Boersma, P., Weenink, D. (2025). Praat: doing phonetics by computer, computer program. Version 6.4.43. Retrieved from http://www.praat.org/

[3] Baken R. J. (1987), Clinical measurement of speech and voice, College-Hill Press, 544 pages

[4] Fastl H., and Zwicker, E. (2007). Psychoacoustics, Facts and Models, Chap. Fluctuation Strength (Springer), unpaginated e-book.

[5] ATIC Research Group, U. (2018). ATIC-DB Perceptual Voice Quality Assessment 2, http://www.atic.uma.es/index_atic.html, [Online; accessed 23-January-2018].

[6] Diaconis P. and Graham R. L. (1977), Spearman's Footrule as a Measure of Disarray, J. Royal Stat. Society. Series B, 39(2), pp. 262-268

A GRAPHICAL USER INTERFACE FOR GENERATING SYNTHETIC VOWELS WITH PREDEFINED ACOUSTIC PARAMETERS

D. Gasperini[1], S. Orlandi[2,3], A. Bandini[1,4,5]
[1]Health Science Interdisciplinary Research Center, Scuola Superiore Sant'Anna, Pisa, Italy
[2]Department of Electrical, Electronic, and Information Engineering "Guglielmo Marconi" – DEI, University of Bologna, Bologna, Italy; Health Sciences and Technologies, Interdepartmental Center for Industrial Research (CIRI-SDV), University of Bologna, Italy
[3]IRCCS Istituto delle Scienze Neurologiche di Bologna, Bologna, Italy
[4]The BioRobotics Institute and Department of Excellence in Robotics and AI, Scuola Superiore Sant'Anna, Pisa, Italy
[5]KITE - Toronto Rehabilitation Institute, University Health Network, Toronto, ON, Canada
daniela.gasperini@santannapisa.it; silvia.orlandi9@unibo.it; andrea.bandini@santannapisa.it

***Abstract*: Dysarthric speech is an important biomarker for clinical assessment and diagnostic support in neurological diseases. However, speech recordings are often collected in uncontrolled and noisy environments, such as clinics and home settings. Speech enhancement and denoising tools can help with this challenge, but their effects on important acoustic features, like fundamental frequency (F_0) and the first two formants (F_1, F_2), are not well understood. This uncertainty raises concerns about possible distortions. To address this, a Graphical User Interface (GUI) was developed to create synthetic American English vowels with predefined fundamental frequency and first two formants, providing a controlled ground-truth reference. The fundamental frequency is derived from Gaussian distributions. The first and second formants are sampled using a kernel density estimation technique to ensure physiologically plausible vowels. The glottal source is modulated with jitter and shimmer to mimic variations in speech. The GUI allows flexible control over vowel type, noise condition, signal duration, and modulation parameters. This enables reproducible benchmarking of speech processing algorithms under controlled noise scenarios. Planned extensions include adding support for Italian vowels, higher-order formants, and validating quality indices for signal denoising and speech enhancement algorithms.**

***Keywords:* synthetic speech, vowels, phonetics, GUI**

I. INTRODUCTION

Neurological diseases such as Parkinson's disease (PD) and amyotrophic lateral sclerosis (ALS) commonly lead to dysarthria, a motor speech disorder characterized by impaired speech execution [1], [2]. Acoustic parameters, particularly fundamental frequency (F_0) and formants (F_1, F_2), serve as valuable biomarkers for monitoring disease progression [3].

F_0 reflects the periodic vibration of the vocal folds and differs between males and females. In contrast, F_1 and F_2 correspond to vocal tract resonances during phonation, characterizing different vowels. These formants are strongly correlated with jaw and tongue muscular activity, making them sensitive indicators of neuromotor degeneration [4]. By mapping vowels in the F_1–F_2 plane, it is possible to derive the vowel space area (VSA), which is a critical acoustic biomarker for monitoring neurological diseases due to its sensitivity to articulatory impairments. Patients with PD and ALS typically present a reduced VSA relative to healthy controls [5], [6].

Estimating these parameters in patients with neurological diseases presents significant challenges. The inherent variability in neurological voice signals, especially when affected by dysarthria, necessitates robust analysis methods that can accommodate irregular vocal patterns while maintaining measurement accuracy. Furthermore, audio recordings are typically acquired in noisy environments (e.g., hospitals or at home), where background noise from medical equipment or power lines can corrupt the voice signal [7].

Recent advances in machine learning and deep learning have improved speech signal analysis [8]. For instance, Jolad & Khanai proposed a Competitive Crow Search Algorithm-based Speech Enhancement Generative Adversarial Network (FCCSA-SEGAN) to enhance dysarthric speech, boosting quality and intelligibility across noise conditions [9], while Wang et al. used a Convolutional Neural Network (CNN) to improve the intelligibility of dysarthric speech, achieving over 10% improvement in automatic speech recognition (ASR) and subjective human intelligibility tests [10]. Nonetheless, evaluating the performance of these algorithms remains challenging, as it requires simultaneous consideration of multiple factors. Traditional metrics such as Signal-to-Noise Ratio (SNR) are insufficient because they are too general and can apply to non-speech signals. A comprehensive evaluation should consider various metrics including

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

Perceptual Evaluation of Speech Quality (PESQ), and Short-Time Objective Intelligibility (STOI), as different algorithms excel in different aspects [11].
The analysis of acoustic features is further complicated by the absence of ground-truth clean signals for direct comparison. Although tools such as Praat and BioVoice can estimate F_0 and formants [12], [13], these remain approximations and do not address the absence of reference data. Importantly, vowel segments are especially critical for speech assessment, but among the most challenging signals to denoise. Vowel spectral envelopes are affected by noise, particularly in the mid-frequency band (1–2.7 kHz), which includes the third formant band. This sensitivity makes vowels more susceptible to distortion and harder to recover accurately in noisy conditions [14]. To overcome this limitation, we present a vowel synthesizer that produces controlled vowel signals with known acoustic parameters, thereby enabling the creation of a reliable ground-truth database for benchmarking speech processing algorithms. Building on the work of Orlandi et al. [15], which employed synthetic signals of infant cries to evaluate and compare various analysis methods, the present study employed a public dataset of American English vowels collected by Hillenbrand [16]. This dataset includes acoustic measurements of F_0, F_1, and F_2 for male, female, and child speakers, and was used as a reference for generating synthetic signals. Specifically, we developed a MATLAB-based Graphical User Interface (GUI) that allows users to select vowels within the F_1-F_2 plane, control signal duration, and introduce realistic sources such as power line (50 Hz) and fan noise, and a theoretical one (i.e., white noise). The vowel space is sampled via kernel density estimation (KDE) to constrain synthesized vowels to physiologically plausible regions, resulting in a synthetic database with known ground-truth parameters. This tool provides a flexible framework to compare denoising algorithms and conduct controlled studies on vowel acoustics.

II. METHODS

According to the source–filter theory [17], the glottal acoustic signal is filtered by the resonant properties of the vocal tract, producing the characteristic formants of speech sounds. To parameterize the synthesizer, we relied on the widely used dataset of American English vowels [18], which reports acoustic measures of F_0, F_1, and F_2 for men, women, and children. In this preliminary study, we focused exclusively on adult speakers. To simulate realistic recording conditions, additive noise was included in the model according to:

$$y[k]=s[k]+n[k],\ k = 1,\ldots., N \quad (1)$$

where $y[k]$ is the noisy speech signal, $s[k]$ is the clean synthesized vowel signal, $n[k]$ is one of the three types of additive noise, and N is the number of samples.
Each formant was modeled as a second-order band-pass filter. Variations in speech were simulated by introducing jitter and shimmer, which reflect relevant characteristics for pathological speech [19]. Candidate values of F_1 and F_2 were initially sampled uniformly within the observed ranges of the dataset. A KDE model, fitted on the original data, was then used to evaluate the likelihood of each (F_1, F_2) pair in the vowel space. Only candidates exceeding the 95-th percentile of the KDE distribution were retained, ensuring that synthesized vowels remained within high-probability regions of the acoustic space. From the accepted points, a single pair was randomly selected for synthesis. Corresponding values for F_0 were sampled from Gaussian distributions parameterized by the mean and standard deviation of the respective vowel in the original dataset, with clamping applied to maintain physiological plausibility.

III. RESULTS

Fig. 1 shows the GUI. Users select the vowel to synthesize, the sex of the speaker, set the duration of the signal, and optionally introduce shimmer and jitter.

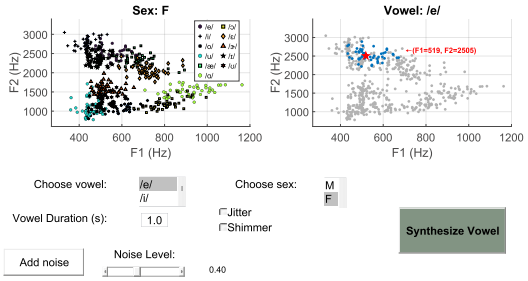


Fig. 1: Graphical User Interface for vowel synthesis.

A noise level slider lets users add background noise from a predefined library. When the "Synthesize Vowel" button is pressed, the GUI selects formant values from the dataset, generates an excitation signal, applies the chosen parameters, and synthesizes the vowel through the resonator model described in the previous section. It then saves both clean and noisy audio and updates interactive plots displaying the vowel space and the selected vowel detail.
Fig. 2 shows F_1-F_2 plane for male speakers in the Hillenbrand dataset. By sampling F_1 and F_2 values using the KDE approach, the synthesized vowels are constrained to high-probability regions, closely reflecting the distributions observed in real speakers. The legend is reported in IPA (International Phonetic

Alphabet), which for American English includes 11 phonemes for vowels *(/i/, /ɪ/, /e/, /ɛ/, /ɝ/, /æ/, /a/, /ɔ/, /o/, /ʊ/, /u/)*.

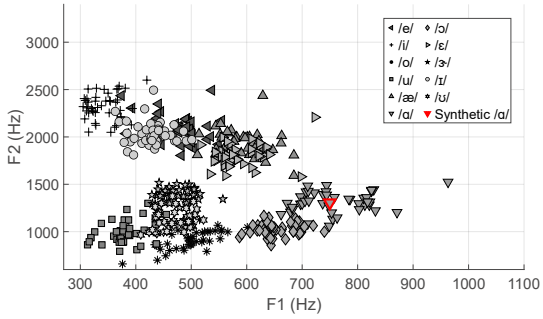


Fig. 2: F_1-F_2 plane for male speakers in the dataset. Each marker represents a different vowel. The red marker indicates the synthetically generated vowel.

Table 1 reports the mean and standard deviation of F_0, F_1, and F_2 values for 48 women and 45 men from the dataset, which were used as reference parameters for generating the synthesized vowels.

Table 1. Mean and standard deviation (in Hz) of fundamental frequency (F_0) and formants (F_1–F_2) for female and male speakers.

Parameter (Hz)	Male	Female
F_0	131 ± 22	220 ± 23
F_1	515 ± 123	602 ± 160
F_2	1538 ± 499	1795 ± 623

IV. Discussion

Neurological diseases often lead to speech impairments such as dysarthria, which compromise communication. Acoustic markers, including formant-related measures, have proven valuable for tracking disease severity and progression. However, their extraction from natural recordings remains challenging, as speech signals are often corrupted by environmental noise, such as background sounds from medical equipment or household environments.

In this paper, we presented an easy tool for generating synthetic vowels with known acoustic parameters to create a synthetic F_1-F_2 plane. In this direction the proposed tool enables controlled testing of denoising algorithms by isolating variables such as SNR and noise type, providing a reliable framework for algorithm evaluation under controlled conditions. Previous works have also explored vowel synthesis for the analysis of pathological voices. In [20], the authors proposed a model that reproduces alterations in acoustic parameters such as jitter and shimmer, using sustained vowels from a database of speakers without speech impairments and individuals affected by various voice disorders. Their method focuses on capturing irregularities in phonation to simulate pathological conditions and evaluate acoustic correlations of vocal disorders. By contrast, our approach does not aim to reproduce pathology-related perturbations but rather to generate clean, parameter-controlled vowels that can be subsequently corrupted with realistic noise sources. More recently, VSpace, a browser-based tool for vowel synthesis, was presented in [21]. It allows exploration of the universal vowel space by selecting formant frequencies within a trapezoid scaled to different speaker ranges. While this tool is valuable for educational and perceptual studies, its design primarily focuses on accessibility and interactive exploration rather than systematic dataset generation. Conversely, in this work we sample the vowel space through a KDE strategy, ensuring that synthesized vowels remain within physiologically plausible regions. This enables the construction of a synthetic database with known ground-truth parameters, specifically tailored for benchmarking denoising and speech enhancement algorithms under controlled conditions. The proposed tool is publicly available on GitHub (https://github.com/Gasp-sh/VowelSynthGUI.git).

V. Conclusion

This work presents a simple vowel synthesizer designed to evaluate algorithms developed for neurological patients with speech impairments. The database is based on real male and female American English speakers and employs a KDE-based approach to generate realistic vowel signals. By allowing precise control over synthesis parameters, including SNR and noise type, the tool provides a reproducible framework for benchmarking of denoising and speech enhancement algorithms under controlled conditions. It is important to note that this framework is currently limited to vowels and the first two formants. While it allows controlled testing of algorithms on synthetic signals, whether the same algorithms perform similarly on natural speech from patients remains to be evaluated. Nevertheless, the tool could represent a promising starting point for controlled evaluations, and future work will extend the synthesis to higher-order formants and additional languages, such as Italian.

References

[1] S. Sapir, «Multiple Factors Are Involved in the Dysarthria Associated With Parkinson's Disease: A Review With Implications for Clinical Practice and Research», *J. Speech Lang. Hear. Res.*, vol.

57, fasc. 4, pp. 1330–1343, ago. 2014, doi: 10.1044/2014_JSLHR-S-13-0039.

[2] B. Tomik e R. J. Guiloff, «Dysarthria in amyotrophic lateral sclerosis: A review», *Amyotroph. Lateral Scler.*, vol. 11, fasc. 1–2, pp. 4–15, gen. 2010, doi: 10.3109/17482960802379004.

[3] P. Gómez-Vilda *et al.*, «Neurological Disease Detection and Monitoring from Voice Production», in *Advances in Nonlinear Speech Processing*, vol. 7015, C. M. Travieso-González e J. B. Alonso-Hernández, A c. di, in Lecture Notes in Computer Science, vol. 7015. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1–8. doi: 10.1007/978-3-642-25020-0_1.

[4] P. Gómez-Vilda *et al.*, «Neuromechanical Modelling of Articulatory Movements from Surface Electromyography and Speech Formants», *Int. J. Neural Syst.*, vol. 29, fasc. 02, p. 1850039, mar. 2019, doi: 10.1142/S0129065718500399.

[5] S. Skodda, W. Grönheit, e U. Schlegel, «Impairment of Vowel Articulation as a Possible Marker of Disease Progression in Parkinson's Disease», *PLOS ONE*, vol. 7, fasc. 2, p. e32132, feb. 2012, doi: 10.1371/journal.pone.0032132.

[6] G. S. Turner, K. Tjaden, e G. Weismer, «The Influence of Speaking Rate on Vowel Space and Speech Intelligibility for Individuals With Amyotrophic Lateral Sclerosis», *J. Speech Lang. Hear. Res.*, vol. 38, fasc. 5, pp. 1001–1013, ott. 1995, doi: 10.1044/jshr.3805.1001.

[7] E. E. Ryherd, K. P. Waye, e L. Ljungkvist, «Characterizing noise and perceived work environment in a neurological intensive care unit», *J. Acoust. Soc. Am.*, vol. 123, fasc. 2, pp. 747–756, feb. 2008, doi: 10.1121/1.2822661.

[8] J. Wang *et al.*, «Automatic prediction of intelligible speaking rate for individuals with ALS from speech acoustic and articulatory samples», *Int. J. Speech Lang. Pathol.*, vol. 20, fasc. 6, pp. 669–679, ott. 2018, doi: 10.1080/17549507.2018.1508499.

[9] B. Jolad e R. Khanai, «An approach for speech enhancement with dysarthric speech recognition using optimization based machine learning frameworks», *Int. J. Speech Technol.*, vol. 26, fasc. 2, pp. 287–305, lug. 2023, doi: 10.1007/s10772-023-10019-y.

[10] S. Wang *et al.*, «Dysarthric Speech Enhancement Based on Convolution Neural Network», in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, lug. 2022, pp. 60–64. doi: 10.1109/EMBC48229.2022.9871531.

[11] Y. Hu e P. C. Loizou, «Subjective comparison and evaluation of speech enhancement algorithms», *Speech Commun.*, vol. 49, fasc. 7–8, pp. 588–601, lug. 2007, doi: 10.1016/j.specom.2006.12.006.

[12] «Praat: Doing Phonetics by Computer», *Ear Hear.*, vol. 32, fasc. 2, p. 266, apr. 2011, doi: 10.1097/AUD.0b013e31821473f7.

[13] M. S. Morelli, S. Orlandi, e C. Manfredi, «BioVoice: A multipurpose tool for voice analysis», *Biomed. Signal Process. Control*, vol. 64, p. 102302, feb. 2021, doi: 10.1016/j.bspc.2020.102302.

[14] G. Parikh e P. C. Loizou, «The influence of noise on vowel and consonant cues», *J. Acoust. Soc. Am.*, vol. 118, fasc. 6, pp. 3874–3888, dic. 2005, doi: 10.1121/1.2118407.

[15] S. Orlandi, A. Bandini, F. F. Fiaschi, e C. Manfredi, «Testing software tools for newborn cry analysis using synthetic signals», *Biomed. Signal Process. Control*, vol. 37, pp. 16–22, ago. 2017, doi: 10.1016/j.bspc.2016.12.012.

[16] J. Hillenbrand, L. A. Getty, M. J. Clark, e K. Wheeler, «Acoustic characteristics of American English vowels», *J. Acoust. Soc. Am.*, vol. 97, fasc. 5, pp. 3099–3111, mag. 1995, doi: 10.1121/1.411872.

[17] I. Tokuda, «The Source–Filter Theory of Speech», in *Oxford Research Encyclopedia of Linguistics*, 2021. doi: 10.1093/acrefore/9780199384655.013.894.

[18] «OSF | A practical guide to calculating vocal tract length and scale-invariant formant patterns». Consultato: 16 settembre 2025. [Online]. Disponibile su: https://osf.io/4c2r9/

[19] H. F. Wertzner, S. Schreiber, e L. Amaro, «Analysis of fundamental frequency, jitter, shimmer and vocal intensity in children with phonological disorders», *Braz. J. Otorhinolaryngol.*, vol. 71, fasc. 5, pp. 582–588, ott. 2015, doi: 10.1016/S1808-8694(15)31261-1.

[20] G. A. Alzamendi, G. Schlotthauer, H. L. Rufiner, e M. E. Torres, «Evaluation of a new model for vowels synthesis with perturbations in acoustic parameters», *Lat. Am. Appl. Res.*, vol. 43, fasc. 3, pp. 225–230, lug. 2013.

[21] M. I. Proctor, «VSpace: A browser-based vowel synthesiser», *J. Acoust. Soc. Am.*, vol. 154, fasc. 4_supplement, p. A203, ott. 2023, doi: 10.1121/10.0023276.

Finite Element Model of Vocal Fold Dynamics with Laryngeal Muscle Activation-Dependent Parameters

Cristobal Ponce[1], Jesús A. Parra[1], Sean D. Peterson[2], Hector Ramirez[1,3], Matías Zañartu[1,3]

[1]Advanced Center for Electrical and Electronic Engineering, Valparaíso, Chile.
[2]Department of Mechanical and Mechatronics Engineering, University of Waterloo, Canada.
[3]Department of Electronic Engineering, Universidad Técnica Federico Santa María, Valparaíso, Chile.
Emails: cristobal.ponces@usm.cl, jesus.parrap@sansano.usm.cl, peterson@uwaterloo.ca, hector.ramireze@usm.cl, matias.zanartu@usm.cl

***Abstract*: This paper presents a finite element model of vocal fold dynamics with muscle activation-dependent parameters. The geometry is parametrized analogously to lumped-parameter formulations, capturing how intrinsic laryngeal muscle activations affect vocal fold length and layer depths. Elastic properties are modeled as polynomial functions of muscle activation, and the polynomial coefficients are identified from data of modal frequencies as functions of activation. The approach provides a simple yet extensible framework that bridges the gap between lumped and highly detailed finite element models, making it well suited for characterizing vocal fold dynamics when activation-dependent data are available.**

***Keywords*: Vocal Folds, Muscle activation, Finite Element Method**

I. INTRODUCTION

Vocal fold (VF) vibration arises from the interplay between tissue mechanics, aerodynamics, and neuromuscular control [1]. A central challenge for modeling is to capture how intrinsic laryngeal muscle (ILM) activation alters tissue stiffness, mass distribution, and geometry, thereby shaping vibratory modes and eigenfrequencies. Ex-vivo and in-silico studies have shown that ILMs regulate both posture and the natural frequency of oscillation [2–4]. Modeling approaches range from lumped-element representations [5], where muscle activity is mapped into effective stiffness or mass coefficients, to highly detailed finite element (FE) models of the VFs [6, 7]. Lumped models are computationally efficient and insightful but lack spatial resolution, while detailed FE models capture anatomy more realistically but at the expense of complexity and computational cost.

Prior studies have incorporated ILM control into FE laryngeal models from MRI data for canine [8] and human larynges [9]. These models have been used to analyze the influence of cricothyroid and thyroarytenoid activation on pre-phonatory posturing and glottic dynamics [10], and investigations of vibration mode changes under biomechanical variations [11]. These high-fidelity models have embedded the laryngeal muscles by modeling the entire tissue and cartilage structure of the larynx, and thus are computationally expensive.

In this work, we propose a method that represents ILM activation in FE models without requiring a full anatomical description of the larynx, while also establishing a connection to existing approaches in lumped-element models. The VFs are represented with a finite element model whose geometry is parameterized analogously to established lumped-element rules, reflecting the antagonistic actions of the cricothyroid (CT) and thyroarytenoid (TA) muscles.

Elastic properties are modeled as polynomial functions of muscle activation, with coefficients identified from data of modal frequencies as functions of activation. This formulation allows activation-dependent variability to be systematically incorporated without requiring detailed tissue-level characterization. The proposed finite element model is constructed using a port-Hamiltonian representation due to its structured formulation and its ability to incorporate additional effects in an energy-consistent manner [12–14]. The present contribution therefore introduces a practical framework that balances interpretability, extensibility, and computational simplicity, paving the way for systematic characterization of activation-dependent vocal fold dynamics.

II. METHODS

A. Lumped-parameter model

The lumped-parameter model is described as the interaction of three coupled masses corresponding to the body and the cover layer of the VF [15]. The main geometric dimensions considered in this model are the VF

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

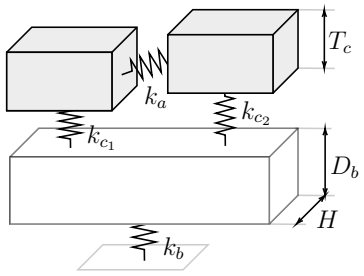


Figure 1: Schematic of lumped model.

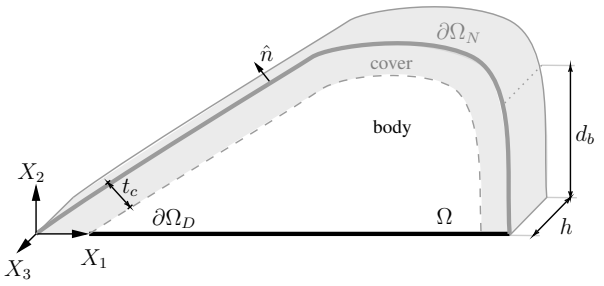


Figure 2: Schematic of distributed model.

length H, the body layer height D_b, and the cover layer thickness T_c. The mechanical behavior of the system is characterized through a set of springs: k_b associated with the body, k_{c_1} and k_{c_2} corresponding to the cover, and k_a representing the interaction between the two, to emulate the mucosal wave, as illustrated in Fig. 1. According to [16], these parameters vary with ILM activation, described by $\alpha = \{\alpha_{\text{CT}}, \alpha_{\text{TA}}\}$, where $\alpha_{\text{CT}}, \alpha_{\text{TA}} \in [0, 1]$ represent the activation levels of the cricothyroid and thyroarytenoid muscles, respectively. This simplified two-muscle representation accounts for the mechanical properties of the VF but not its posture. A more detailed description of the action of the complete set of ILMs in the context of lumped-element models can be found in [17–19]. The specific relationships between muscle activations and the stiffness and geometrical parameters are given in [16] and will be further discussed in Section 3.

B. Finite element model

The VF dynamics are described by the two-dimensional equations of linear elastodynamics written in port-Hamiltonian form [20]. A schematic of the spatial domain Ω is shown in Fig. 2. Then, the evolution of the momentum $p(\mathrm{X}, t)$ and strain $\epsilon(\mathrm{X}, t)$ fields is given by:

$$\begin{bmatrix}\dot{p}(\mathrm{X},t)\\ \dot{\epsilon}(\mathrm{X},t)\end{bmatrix} = \begin{bmatrix}0 & -\mathcal{F}_{\mathrm{X}}^*\\ \mathcal{F}_{\mathrm{X}} & 0\end{bmatrix}\begin{bmatrix}\mathcal{M}(\mathrm{X})^{-1} & 0\\ 0 & \mathcal{K}(\mathrm{X})\end{bmatrix}\begin{bmatrix}p(\mathrm{X},t)\\ \epsilon(\mathrm{X},t)\end{bmatrix}, \quad (1)$$

where the operators are defined as follows:

$$\mathcal{M}(\mathrm{X}) = \begin{bmatrix}\rho(\mathrm{X})h & 0\\ 0 & \rho(\mathrm{X})h\end{bmatrix}, \quad (2)$$

$$\mathcal{K}(\mathrm{X}) = \frac{E(\mathrm{X})h}{1-\nu^2}\begin{bmatrix}1 & \nu & 0\\ \nu & 1 & 0\\ 0 & 0 & \frac{(1-\nu)}{2}\end{bmatrix}, \quad (3)$$

$$\mathcal{F}_{\mathrm{X}} = \begin{bmatrix}\partial_1 & 0\\ 0 & \partial_2\\ \partial_2 & \partial_1\end{bmatrix}, \; -\mathcal{F}_{\mathrm{X}}^* = \begin{bmatrix}\partial_1 & 0 & \partial_2\\ 0 & \partial_2 & \partial_1\end{bmatrix}. \quad (4)$$

Here, $\mathcal{M}(\mathrm{X})$ and $\mathcal{K}(\mathrm{X})$ denote the mass and stiffness density matrices, respectively, $\mathcal{F}_{\mathrm{X}}$ is the differential operator representing kinematics (symmetric gradient), and $\mathcal{F}_{\mathrm{X}}^*$ its formal adjoint, with $\partial_i = \partial/\partial X_i$. The spatially varying density $\rho(\mathrm{X})$ takes the value ρ_b in the body and ρ_c in the cover, while the Young's modulus $E(\mathrm{X})$ equals E_b in the body and E_c in the cover. This PDE model is discretized using finite element procedures [21], which yields the following finite element model:

$$\begin{bmatrix}\dot{\hat{p}}(t)\\ \dot{\hat{r}}(t)\end{bmatrix} = \begin{bmatrix}0 & -I\\ I & 0\end{bmatrix}\begin{bmatrix}\hat{M}^{-1} & 0\\ 0 & \hat{K}\end{bmatrix}\begin{bmatrix}\hat{p}(t)\\ \hat{r}(t)\end{bmatrix}, \quad (5)$$

where $\hat{p}(t)$ and $\hat{r}(t)$ are the vectors of nodal momenta and displacements, respectively. The global matrices $\hat{M}$ and $\hat{K}$ result from the standard assembly and depend on the tissue density, Young's modulus, and geometry as:

$$\hat{M} = \sum_{e=1}^{n_e}(L_r^e)^\top \int_{\Omega^e} N_r^e(\mathrm{X})^\top \mathcal{M}(\mathrm{X}) N_r^e(\mathrm{X})\, d\mathrm{X}\, L_r^e,$$

$$\hat{K} = \sum_{e=1}^{n_e}(L_r^e)^\top \int_{\Omega^e} (\mathcal{F}_{\mathrm{X}} N_r^e(\mathrm{X}))^\top \mathcal{K}(\mathrm{X})(\mathcal{F}_{\mathrm{X}} N_r^e(\mathrm{X}))\, d\mathrm{X}\, L_r^e,$$

where n_e is the total number of elements, Ω^e is the element domain, L_r^e is the assembly matrix, and $N_r^e(\mathrm{X})$ is the shape function.

III. ACTIVATION-DEPENDENT PARAMETERS

A. Proposal

Inspired by the empirical rules in [16] and laryngeal tissue studies [22, 23], this work establishes an analogy between the parameters of lumped-element models and those of the PDE model to describe the action of ILMs. The following equations provide a simplified description of the geometry as a function of muscle activations:

$$h(\alpha) = h_0(1 + \varepsilon(\alpha)), \quad (6)$$

where h_0 represents the resting length of the VFs, and $\varepsilon(\alpha)$ denotes the elongation derived from muscle activation, defined as:

$$\varepsilon(\alpha) = G(R\,\alpha_{\text{CT}} - \alpha_{\text{TA}}) - Y, \quad (7)$$

with constants $G = 0.2$, $R = 3$, and $Y = 0.1$ [16]. This formulation captures the opposing actions of the CT and TA muscles: the CT elongates the VFs through displacement of the cricothyroid joint, while the TA shortens the VF and provides greater space for the body. Consequently, the body depth is expressed as:

$$d_b(\alpha) = \frac{\alpha_{\text{TA}} d_{\text{mus}} + d_{\text{lig}}}{1 + \varepsilon(\alpha)}, \tag{8}$$

where $d_{\text{mus}} = 0.4\,[\text{cm}]$ and $d_{\text{lig}} = 0.2\,[\text{cm}]$ correspond to the depths of the TA muscle and the ligament, respectively. The denominator $(1 + \varepsilon(\alpha))$ accounts for the shortening in this direction due to VF elongation. Lastly, the cover depth is given by:

$$t_c(\alpha) = \frac{\beta(d_{\text{muc}} + d_{\text{lig}})}{1 + \varepsilon(\alpha)}, \tag{9}$$

with $d_{\text{muc}} = 0.2$ [cm] representing the depth of the mucosa, and $\beta = 0.1$ [–] a correction factor. For the elastic behavior, a simplified second-order model for CT and TA activations is used to represent the passive and active components of ILM action. This formulation is motivated by the relationship between stiffness values and the resulting fundamental frequencies (f_o [Hz]) observed in lumped-element models [16, 19]. The Young's moduli of the body and the cover are then proposed as:

$$E_b(\alpha) = \sum_{i=1}^{2} E_{b0}(A_i\,\alpha_{\text{CT}}^i + B_i\,\alpha_{\text{TA}}^i + K_1\,\alpha_{\text{CT}}\alpha_{\text{TA}} + 1),$$

$$E_c(\alpha) = \sum_{i=1}^{2} E_{c0}(C_i\,\alpha_{\text{CT}}^i + D_i\,\alpha_{\text{TA}}^i + K_2\,\alpha_{\text{CT}}\alpha_{\text{TA}} + 1),$$

where the coefficients A_i, B_i, C_i, D_i, K_i are obtained by solving a fitting problem. In this work, these parameters are identified by minimizing the error between the first eigenfrequency predicted by (5) and the fundamental frequency f_o [Hz] obtained from Muscle Activation Plots (MAPs) [19].

Remark 1: The spatial domain and its local finite element subdomains are parameterized by the muscle activation α, i.e., $\Omega = \Omega(\alpha)$ and $\Omega^e = \Omega^e(\alpha)$. Consequently, the assembled mass and stiffness matrices depend implicitly on α through $\Omega^e(\alpha)$, and explicitly through $h(\alpha)$, $E_b(\alpha)$, and $E_c(\alpha)$ via the definitions of the continuous density matrices $\mathcal{M}(\mathtt{X})$ and $\mathcal{K}(\mathtt{X})$.

B. Simulation results

The identification of the parameters A_i, B_i, C_i, D_i, K_i is carried out by minimizing the error between the eigenfrequencies predicted by the FEM model and those reported in [19], whose data are shown graphically in Fig. 3. The assumed known parameters are: $h_0 = 1.25$ [cm], $E_{b0} = 5$ [kPa], $E_{c0} = 2.5$ [kPa], $\rho_b = \rho_c = 1000$ [kg/m^3], and $\nu = 0.4$ [–]. The stochastic global method, Particle Swarm Optimization (PSO), was used in this study. The search domain for the unknown parameters was set to the interval $[-20, 20]$. This range was deemed sufficient given that muscle activations lie within $[0, 1]$ and the maximum fundamental frequency to be replicated, according to the available data, is on the order of 400 [Hz]. A summary of the estimated parameters is presented in Table 1, while the corresponding fits are illustrated in Fig. 4.

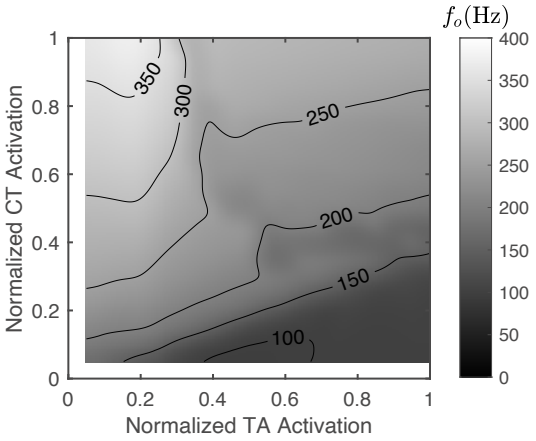


Figure 3: Fundamental frequency f_o from [19].

Table 1: Identified parameters.

Par.	Value	Par.	Value
A_1	18.38	C_1	15.61
A_2	-5.16	C_2	-17.77
B_1	6.97	D_1	16.02
B_2	-4.43	D_2	7.26
K_1	15.40	K_2	17.63

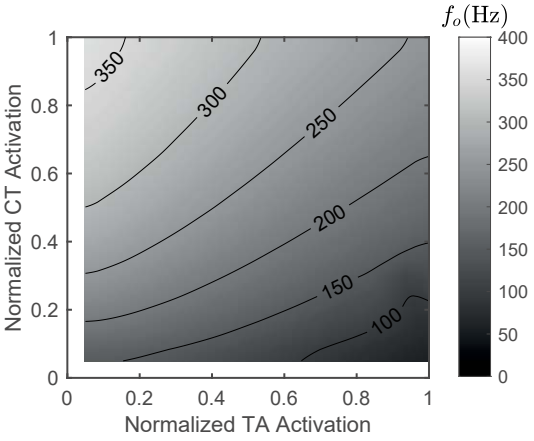


Figure 4: Results using PSO.

C. Discussion

The parameter identification results (Fig. 4) show that the proposed framework successfully reproduces the overall trend of the first natural frequency as a function of muscle activations. A similar trend has also been reported in [6], where a fiber-gel FEM of the vocal folds was employed to map TA and CT activations to postural and acoustic features. The consistency of the observed frequency–activation relationship across

both models reinforces the validity of the present simplified formulation. This agreement is achieved despite using low-order polynomial functions to describe the activation-dependence of Young's moduli and the geometry of the folds. The present study focused primarily on establishing the methodological pipeline linking lumped and distributed modeling frameworks, providing an intermediate level of complexity while maintaining interpretability. For this reason, the current implementation considered second-order polynomials as an initial step, aiming to represent both large- and small-scale effects of activation. Future efforts will explore the influence of higher polynomial orders to capture more complex and possibly subject-specific relationships between ILMs and tissue mechanics, particularly when validating against experimental data or alternative models. It should be noted that the present formulation accounts for the combined action of the CT and TA muscles, which mainly influence VF stiffness and longitudinal strain. Although these components dominate the mechanical contribution to VF dynamics, additional ILMs affecting vocal posture were not included. Extending the framework to incorporate such muscles would require a three-dimensional representation capable of describing postural adjustments. This constitutes an important direction for future work.

IV. CONCLUSION

This work introduced an activation-dependent finite element model of VF dynamics, where both geometry and elastic properties are parameterized by ILM activations. The results demonstrate that low-order polynomial functions suffice to capture the principal trend of the first natural frequency across activation ranges, bridging data from lumped-element formulations with distributed FEM descriptions. The proposed framework provides a practical pathway for incorporating muscle-activation dependence into biomechanical models of VFds, serving as a foundation for connecting finite- and reduced-order modeling domains. Future work will focus on refining the constitutive parameterization, evaluating the effect of higher-order polynomial mappings, and extending the model to include additional ILMs through three-dimensional representations. Validation against experimental measurements will also be pursued to enhance predictive and physiological relevance.

REFERENCES

[1] Z. Zhang. Mechanics of human voice production and control. *The journal of the acoustical society of america*, 140(4):2614–2635, 2016.

[2] I. Titze, E. Luschei, and M. Hirano. Role of the thyroarytenoid muscle in regulation of fundamental frequency. *Journal of Voice*, 3(3):213–224, 1989.

[3] D. Chhetri, J. Neubauer, and D. Berry. Graded activation of the intrinsic laryngeal muscles for vocal fold posturing. *The Journal of the Acoustical Society of America*, 127(4):127–133, 2010.

[4] D. Chhetri and J. Neubauer. Differential roles for the thyroarytenoid and lateral cricoarytenoid muscles in phonation. *The Laryngoscope*, 125(12):2772–2777, 2015.

[5] B. Erath, M. Zanartu, K. Stewart, M. Plesniak, D. Sommer, and S. Peterson. A review of lumped-element models of voiced speech. *Speech Communication*, 55(5):667–690, 2013.

[6] A. Palaparthi, S. Smith, and I. Titze. Mapping thyroarytenoid and cricothyroid activations to postural and acoustic features in a fiber-gel model of the vocal folds. *Applied Sciences*, 9(21):4671, 2019.

[7] W. Jiang, B. Geng, X. Zheng, and Q. Xue. A computational study of the influence of thyroarytenoid and cricothyroid muscle interaction on vocal fold dynamics in an mri-based human laryngeal model. *Biomechanics and Modeling in Mechanobiology*, 23(5): 1801–1813, 2024.

[8] B. Geng, N. Pham, Q. Xue, and X. Zheng. A three-dimensional vocal fold posturing model based on muscle mechanics and magnetic resonance imaging of a canine larynx. *The Journal of the Acoustical Society of America*, 147(4):2597–2608, 2020.

[9] N. Pham, Q. Xue, and X. Zheng. Coupling between a fiber-reinforced model and a hill-based contractile model for passive and active tissue properties of laryngeal muscles: A finite element study. *The Journal of the Acoustical Society of America*, 144(3):EL248–EL253, 2018.

[10] M. Movahhedi, B. Geng, Q. Xue, and X. Zheng. Effects of cricothyroid and thyroarytenoid interaction on voice control: Muscle activity, vocal fold biomechanics, flow, and acoustics. *The Journal of the Acoustical Society of America*, 150(1):29–42, 2021.

[11] B. Geng, M. Movahhedi, Q. Xue, and X. Zheng. Vocal fold vibration mode changes due to cricothyroid and thyroarytenoid muscle interaction in a three-dimensional model of the canine larynx. *The Journal of the Acoustical Society of America*, 150 (2):1176–1187, 2021.

[12] V. Duindam, A. Macchelli, S. Stramigioli, and H. Bruyninckx. *Modeling and control of complex physical systems: the port-Hamiltonian approach*. Springer Science & Business Media, 2009.

[13] T. Hélie and F. Silva. Self-oscillations of a vocal apparatus: a port-Hamiltonian formulation. In *International Conference on Geometric Science of Information*, pages 375–383, 2017.

[14] L. Mora, J. Yuz, H. Ramirez, and Y. Le Gorrec. A port-Hamiltonian fluid-structure interaction model for the vocal folds. *IFAC-PapersOnLine*, 51(3):62–67, 2018.

[15] M. Hirano. Morphological structure of the vocal cord as a vibrator and its variations. *Folia phoniatrica et logopaedica*, 26(2): 89–94, 1974.

[16] I. Titze and B. Story. Rules for controlling low-dimensional vocal fold models with muscle activation. *The Journal of the Acoustical Society of America*, 112(3):1064–1076, 2002.

[17] E. Hunter, I. Titze, and F. Alipour. A three-dimensional model of vocal fold abduction/adduction. *The Journal of the Acoustical Society of America*, 115(4):1747–1759, 2004.

[18] I. Titze and E. Hunter. A two-dimensional biomechanical model of vocal fold posturing. *The Journal of the Acoustical Society of America*, 121(4):2254–2260, 2007.

[19] G. Alzamendi, S. Peterson, B. Erath, R. Hillman, and M. Zañartu. Triangular body-cover model of the vocal folds with coordinated activation of the five intrinsic laryngeal muscles. *The Journal of the Acoustical Society of America*, 151(1):17–30, 2022.

[20] C. Ponce, Y. Wu, Y. Le Gorrec, and H. Ramirez. A systematic methodology for port-Hamiltonian modeling of multidimensional flexible linear mechanical systems. *Applied Mathematical Modelling*, 134:434–451, 2024.

[21] C. Ponce, Y. Wu, Y. Le Gorrec, and H. Ramirez. Structure-preserving discretization of multidimensional linear port-Hamiltonian systems using FEM approaches. In *IEEE 63rd Conference on Decision and Control*, pages 2676–2681, 2024.

[22] F. Alipour-Haghighi and I. Titze. Elastic models of vocal fold tissues. *The Journal of the Acoustical Society of America*, 90(3): 1326–1331, 1991.

[23] Y. Min, I. Titze, and F. Alipour-Haghighi. Stress-strain response of the human vocal ligament. *Annals of Otology, Rhinology & Laryngology*, 104(7):563–569, 1995.

SESSION IB
ANALYSIS OF PATHOLOGICAL VOICE

DESIGNING AN EXPERIMENTAL PROTOCOL FOR ELICITING ALZHEIMER'S DISEASE PHONETIC BIOMARKERS IN RUSSIAN SPEAKERS

E.V. Nikolaeva[1], K.V. Evgrafova[2], V.V. Evdokimova[3], P.A. Skrelin[4]

[1] Saint Petersburg State University/Department of Phonetics, Saint-Petersburg, Russian Federation
[2] Saint Petersburg State University/Department of Phonetics, Saint-Petersburg, Russian Federation
[3] Saint Petersburg State University/Department of Phonetics, Saint-Petersburg, Russian Federation
[4] Saint Petersburg State University/Department of Phonetics, Saint-Petersburg, Russian Federation
elenanikolaeva.prod@gmail.com, k.evgrafova@spbu.ru, v.evdokimova@spbu.ru, p.skrelin@spbu.ru

Abstract: **This article presents the design of an experimental protocol created to identify phonetic biomarkers of Alzheimer's disease (AD) in Russian speakers. Due to the uniqueness of the prosodic system of the Russian language and its lack of study in the context of Alzheimer's disease, the proposed protocol aims to fill this gap. The study is based on a cross-sectional design and includes recording of three types of speech activity: spontaneous conversation, reading a semantically ambiguous poem by Eduard Uspensky, and an ironic poem by Samuil Marshak, which allows for a multifaceted analysis of various levels of speech production. The recording is carried out at the participants' homes using professional equipment to ensure environmental validity and high sound quality. The data analysis plan involves the automated extraction of a wide range of acoustic parameters (tempo, pauses, frequency, and phonation characteristics, etc.) followed by statistical processing and the use of machine learning methods to build classification models. It is expected that the implementation of this protocol will make it possible to identify reliable and reproducible speech biomarkers, which will form the basis for the creation of objective disease screening and monitoring tools in the Russian-speaking population.**

Keywords: **Alzheimer's disease, speech biomarkers, research design, phonetics, neurodegenerative diseases.**

I. INTRODUCTION

The socio-economic consequences of the increasing prevalence of AD highlight the urgent need to create effective tools for early detection of the disease. Modern research on neurodegenerative diseases, including AD, indicates the high diagnostic significance of speech markers, especially in the early stages of cognitive decline. Researchers are actively developing methods for effective acoustic speech analysis to diagnose AD using AI (some of them reach accuracy of 85% or higher) [1], [2], [3], [4]. Speech is considered one of the indicators of cognitive functioning, and the features of speech acoustics and prosody are interpreted as potential predictors of cognitive disorders [6], [7], [8].

It is especially important to take into account the language specificity, since the biomarkers identified for English, German, and other languages do not always turn out to be applicable to the Russian language. The phonetic system of the Russian language, characterized by free verbal stress, a rich inventory of prosodic constructions, and a complex morphological structure, requires special study. In the Russian-speaking scientific community, this trend is scantily represented. There are very little systematic data on the specificity of speech biomarkers in AD in Russian language speakers. This justifies the need for a deeper and methodologically verified study.

The aim of this article is to present and substantiate in detail the design of an experimental study created to identify and verify specific phonetic biomarkers of AD in the speech of native speakers of the Russian language. High-quality design is the basis for obtaining valid, reliable, and reproducible data suitable for subsequent implementation in clinical practice.

We hypothesize that in AD speech impairments extend beyond vocabulary, affecting deeper levels of language organization, which is reflected in the prosodic and rhythmic structure of phrases. Therefore, we offer a unique, innovative approach that allows us to shift the diagnostic focus from isolated lexical deficits to a comprehensive analysis of discourse and higher linguistic functions. This method, being the first to employ poetry for assessing, is aimed at identifying not only what the patients say (content), but how they say it (prosody, rhythm, pauses, etc) in conditions of increased cognitive and linguistic stress. It is expected that this approach will be more sensitive to early and subtle changes in AD than traditional methods.

II. METHODS

Participants: We have developed a cross-sectional study design involving two carefully matched groups: an experimental group of patients with diagnosed AD

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

and a control group of healthy subjects comparable in age, sex, and education level. The inclusion/exclusion criteria were developed to maximize the purity of the sample.

Inclusion in the main group: presence of a diagnosis of AD; native Russian language.

Exceptions (for all groups): somatic symptom disorder; bilingualism; severe uncorrected hearing disorders; refusal of informed consent.

Rationale: The criteria are designed to minimize confounding factors. For example, bilingualism or non-native Russian can make systemic changes to speech that are not related to the cognitive status.

Ethical aspects: All participants (or their legal representatives) sign an informed consent. Researchers are trained in ethics. Confidentiality is ensured by encoding data and using anonymous identifiers.

Conditions: Recording is performed at the patient's home in a quiet room on a professional voice recorder. The recordings will have a sample rate of 32000 Hz and a bitrate of 16 bits.

Rationale: The experimental conditions were chosen in order to ensure the validity and ethics of the study. Speech is recorded at the participants' homes in a quiet room, which minimizes stress and external acoustic interference, as well as guarantees signal quality for acoustic analysis. The duration of the session (15-20 minutes, up to 40 minutes if necessary) allows us to collect a sufficient amount of material without the risk of overloading patients. Creating a comfortable and friendly environment eliminates distortion of speech data and complies with international ethical standards for working with vulnerable groups.

The structure of the speech material was selected to activate various cognitive and speech functions affected by AD:

Part I: Spontaneous speech (answers to questions).

Task: To evaluate the functions of semantic planning, lexical access, and syntactic construction of utterance in real time.

Rationale for the choice of material: Personal open-ended questions ("Describe your favorite place", "What is your dog's personality?") encourage a detailed statement. It is in spontaneous speech that the expected biomarkers are most pronounced: an increase in the duration and number of unvoiced pauses, a decrease in the pace of speech, and an increased ratio of pauses to voiced segments, etc [9].

Part II: Reading of S. Marshak's poem "Visiting the Queen".

Task: To evaluate the ability to understand irony and convey emotional and pragmatic nuances through prosody.

Rationale for the choice of material: An adequate transfer of irony requires complex coordination of cognitive and phonetic processes. Patients with AD are expected to have a "smoothed" prosodic contour, which is expressed in decreased dynamic and frequency ranges, monotony [10], and an inability to convey ironic nuances.

Table 1. Approximate translation of Samuil Marshak's poem

– Gde ty byla sivodnya, kiska?	– Where were you today, kitty?
– U karalevy u angliyskoy.	– The Queen of England.
–Shto ty vidala pri dvore?!	– What have you seen at court?!
–Vidala myshku na kavre!	– I have seen a mouse on a carpet!

Part III: Reading of Eduard Uspensky's poem "Doctors' Advice".

Task: To evaluate the understanding of syntax and morphology in the context of lexical uncertainty.

Rationale for the choice of material: The text was specially chosen because of the high concentration of words of indefinite reference ("someone", "that", "those"). This allows us to test the hypothesis that disorders in AD go beyond vocabulary and affect deep syntactic levels, which will be reflected in the prosody.

Table 2. Approximate translation of Eduard Uspensky's poem's fragment

Adin izvesnyy *koy-kto*	One famous *somebody*
Nam fsem glaza atkryl na *to*,	Has opened our eyes to *one thing*,
Shta *to*, shta my shchitali *tyem*,	That *the thing* that we thought was *that,*
Ano mezh tyem ni *to* safsyem.	Is actually not *that* at all.
I nam para rasstatsa s *nim*	And it's time for us to let *it* go
I zamenit' *jivo drugim*.	And replace *it* with *another one*.
…	…
A tut izvesnyy *koy-kto*	And a famous *somebody*
Vdruk pachimu-ta stal *nikto*.	Suddenly, for some reason, became *nobody*.
A dela akazalos' *ftom*,	And it turned out that,
Shto on ashibsya *koy f chom*,	He was wrong about *something*,
I *koy f kom*, i *koy-gdye*.	And about *some people*,
Karochi gavarya, vizdye.	and about *some places*.
	In short, about everything.

The choice of poems as a stimulating material is a fundamental difference between this design and classical approaches and is justified by the following methodological considerations:

1. Overcoming the limitations of classical naming tests.

The traditional diagnosis of amnesic aphasia often relies on tests with picture descriptions [7], [12]. However, this approach has a significant drawback. It isolates the lexical search from its natural context. A poetic text, especially one such as "Doctors' Advice," exposes the subject to semantic and syntactic ambiguity where meaning is born not from individual words, but from grammatical and prosodic connections between them. This allows us to test not just "word knowledge", but the ability to integrate semantics and the use of contextual cues, functions that are critically impaired in AD.

2. Activation of implicit language knowledge and automatized speech patterns.

The rhythmic and melodic structure of verse is based on deeply ingrained, automated patterns of language [13]. Reading poetry requires certain coordination:

Syntactic forecasting: Waiting for the rhyme and rhythmic pattern of the next line.

Morphological analysis: Understanding the role of words with semantic vagueness through their morphological biomarkers (case endings).

Prosodic design: Correlation of the syntactic structure of a sentence with the prosodic contour.

It is precisely these deep, implicit mechanisms of language that are disrupted in patients with AD. It is expected that they will try to read the verse "word by word", losing the overall rhythm and syntactic integrity, which will objectively manifest itself in acoustic parameters (rhythm disruptions, inadequate pauses, "smoothed" prosodic contour).

3. Provocation of linguistic compensation strategies.

Healthy native speakers, when confronted with Uspensky's text, are expected to unconsciously use compensation strategies: varying pitch, using logical accents to clarify the meaning. In patients with AD, these compensatory mechanisms are expected to be disrupted. Thus, the text will not act as an "incomprehensible set of words", but as a sensitive stress test to identify the deficiency of precisely those higher linguistic functions that are responsible for comprehending a complex utterance.

4. The study of pragmatics and the Theory of Mind.

Violation of the Theory of Mind is a well-known symptom of AD [14]. The patients with AD can lose the ability to understand other people by attributing internal mental states to them. Thus, this assignment allows us to test not only phonetics, but also the pragmatic level of the language which is practically not that much affected by standard tests. Marshak's poem "Visiting the Queen" introduces a component of irony and social intelligence into the experiment. For adequate reading, it is necessary:

– To understand that the speaker (the kitty) has her own mental state, different from the listener.
– To detect the discrepancy between the scale of the event ("at the queen's") and its outcome ("I saw a mouse").
– To convey this discrepancy by voice through complex prosodic means (specific changes in tone, pauses, etc.).

5. Standardization while maintaining validity.

One of the key problems of analyzing spontaneous speech is its variability. A poetic text, unlike a conversation, is a standardized stimulus. All subjects read the same text. This allows us to:

– Compare acoustic parameters (vowel length, formants, pauses, etc.) in absolutely identical phonetic and lexical contexts.
– Offset the impact of individual differences in vocabulary and life experience.

III. RESULTS

The present study is at the stage of developing a methodological protocol. At the moment, the main result is a detailed design of the experiment. It includes the following.

1. A standardized protocol for recording speech data adapted for Russian-speaking patients with AD. The protocol includes three modules: spontaneous speech, reading semantically ambiguous text, and reading text with an ironic component.

2. A system of criteria for the selection of subjects, ensuring the formation of representative groups. Detailed inclusion and exclusion criteria have been developed for the main and control groups. Thus, at the current stage, a comprehensive methodological base has been obtained for the subsequent collection and analysis of speech data aimed at identifying acoustic biomarkers of AD in Russian.

IV. DISCUSSION

The presented research design offers a comprehensive and methodologically sound approach to solving the problem of identifying speech biomarkers of AD in the Russian language. The key advantages of the protocol are:

Balanced design: The combination of spontaneous speech reflecting natural communication and supervised

tasks ensuring standardization allows for a comprehensive picture of speech disorders.

Linguistic validity: The choice of stimulus material is aimed at activating specific cognitive functions (memory, planning, theory of consciousness, syntactic parsing), which allows not only to detect the presence of changes, but also hypothetically associate them with certain neurocognitive deficits.

Methodological rigor: The use of professional equipment, clear selection criteria, a standardized protocol, and a comprehensive analysis plan is aimed at minimizing artifacts and increasing the reliability of future results.

Practical application orientation: A predefined set of analyzed acoustic parameters suitable for automated extraction lays the foundation for the creation of AI screening and monitoring tools in the future.

V. CONCLUSION

The development of a detailed and evidence-based design is a crucial first step in studying speech biomarkers of neurocognitive disorders. The experimental protocol presented in the article provides a standardized, ethical, and methodologically rigorous framework for recording and analyzing speech in Russian-speaking patients with AD. An integrated approach combining the analysis of spontaneous and controlled speech with the use of modern acoustic analysis methods and machine learning is promising for identifying reliable phonetic correlates of cognitive decline. It is expected that the subsequent implementation of this design will provide reproducible results suitable for the development of objective tools to support the diagnosis and monitoring of AD in the Russian-speaking population.

Artificial intelligence (AI) which has been used in research to identify patients diagnosed with AD on the basis of genes, MRI images, and electronic health record data can also be applied to speech analysis. The speech biomarkers obtained with the use of the protocol we suggest is to largely increase the accuracy of AI-based AD diagnosis.

The phonetic properties of AD speech need to be analyzed in a language- and country-specific way. Phonetic analyses of the speech of Russian-speaking patients with AD can contribute into developing AI-based AD-diagnosis systems with high accuracy.

The research perspective is to develop an end-to-end deep learning model based on the created data corpus for the automatic detection of speech biomarkers of AD. The system will allow for non-invasive screening and monitoring of cognitive impairments, and can also be integrated into telemedicine platforms for group population surveys.

REFERENCES

[1] Yang Q, Li X, Ding X, Xu F, Ling Z. Deep learning-based speech analysis for Alzheimer's disease detection: a literature review. Alzheimers Res Ther. 2022 Dec 14;14(1):186.

[2] Petti U, Baker S, Korhonen A. A systematic literature review of automatic Alzheimer's disease detection from speech and language. J Am Med Inform Assoc. 2020;27(11)

[3] Mmadumbu AC, Saeed F, Ghaleb F, Qasem SN. Early detection of Alzheimer's disease using deep learning methods. Alzheimers Dement. 2025;21(5).

[4] Yang X, Hong K, Zhang D, Wang K. Early diagnosis of Alzheimer's Disease based on multi-attention mechanism. PLoS One. 2024;19(9).

[5] Judy Illes, Neurolinguistic features of spontaneous language production dissociate three forms of neurodegenerative disease: Alzheimer's, Huntington's, and Parkinson's, Brain and Language, Volume 37, Issue 4, 1989, Pages 628-642.

[6] Martínez-Nicolás I, Llorente TE, Martínez-Sánchez F, Meilán JJG. Ten Years of Research on Automatic Voice and Speech Analysis of People With Alzheimer's Disease and Mild Cognitive Impairment: A Systematic Review Article. Front Psychol. 2021 Mar 23;12:620251.

[7] Saeedi S, Hetjens S, Grimm MOW, Barsties V Latoszek B. Acoustic Speech Analysis in Alzheimer's Disease: A Systematic Review and Meta-Analysis. J Prev Alzheimers Dis. 2024;11(6).

[8] Meilán JJ, Martínez-Sánchez F, Carro J, López DE, Millian-Morell L, Arana JM. Speech in Alzheimer's disease: can temporal and acoustic parameters discriminate dementia? Dement Geriatr Cogn Disord. 2014;37(5-6):327-34.

[9] Martínez-Sánchez F, Meilán JJG, Carro J, Ivanova O. A Prototype for the Voice Analysis Diagnosis of Alzheimer's Disease. J Alzheimers Dis. 2018;64(2):473-481.

[10] Wang, N., Wen, B., Wu, M., Sun, Y., Shao, Z., Zhou, H., Subbalakshmi, K.P. (2025) Decoding Alzheimer's: Interpretable Visual and Logical Attention in Picture Description Tasks. Proc. Interspeech 2025, 2043-2047.

[11] Ilse Lehiste, Phonetic investigation of metrical structure orally produced poetry, Journal of Phonetics, Volume 18, Issue 2, 1990, Pages 123-133.

[12] Moreau N, Rauzy S, Viallet F, Champagne-Lavau M. Theory of mind in Alzheimer disease: Evidence of authentic impairment during social interaction. Neuropsychology. 2016 Mar;30(3):312-21.

NON-INVASIVE SPEECH ANALYSIS FOR DYSPHAGIA DETECTION IN AMYOTROPHIC LATERAL SCLEROSIS

F. Pierotti[1], D. Gasperini[2], S. Capobianco[3], L. Becattini[4], F. Bianchi[5], A. Nacci[3], A. Santoro[3], B. Fattori[3], G. Siciliano[4], A. Bandini[1,2,6]

[1] The BioRobotics Institute and Department of Excellence in Robotics and AI, Scuola Superiore Sant'Anna, Pisa, Italy
[2] Health Science Interdisciplinary Research Center, Scuola Superiore Sant'Anna, Pisa, Italy
[3] ENT, Audiology and Phoniatrics Unit, Pisa University Hospital, Pisa, Italy
[4] Neurology Unit, Department of Clinical and Experimental Medicine, University of Pisa, Pisa, Italy
[5] Neurology Unit, Department of Neuroscience, Azienda Ospedaliera Universitaria Pisana, Pisa, Italy
[6] KITE Research Institute, University Health Network, Toronto, ON, Canada
francesco.pierotti@santannapisa.it; daniela.gasperini@santannapisa.it; silviacapobianco.md@gmail.com; lu.becattini@gmail.com; francicr86@gmail.com; a.nacci@med.unipi.it; asantoro_dr@hotmail.com; bruno.fattori@unipi.it; gabriele.siciliano@unipi.it; andrea.bandini@santannapisa.it

Abstract: **The acoustic and kinematic analyses of speech represent non-invasive and cost-effective tools to support clinicians in the assessment of bulbar dysfunction, which is particularly challenging in amyotrophic lateral sclerosis (ALS). In this study, we applied multimodal speech analysis to identify suitable acoustic and kinematic biomarkers potentially able to detect bulbar impairment, particularly the presence of dysphagia in its early stages. Our results revealed clear distinctions between dysphagic and non-dysphagic individuals with ALS during connected speech, with the third formant and speech timing metrics being the most significant features ($p < 0.001$). Our findings suggest that a reduced F3 and an increased duration of pauses during connected speech may serve as important biomarkers for detecting the onset of dysphagia in ALS.**

Keywords: **Amyotrophic lateral sclerosis, bulbar impairment, dysphagia assessment, speech biomarkers, audio-video**

I. INTRODUCTION

Amyotrophic lateral sclerosis (ALS) is a rare neurodegenerative disease, with a survival time ranging from 3 to 5 years, depending on several factors, including age, type of onset (spinal or bulbar), and disease progression rate [1]. ALS is characterized by the degeneration of the motor neurons, causing paralysis of the limbs, trunk, and orofacial muscles.

ALS is frequently accompanied by bulbar impairments, including speech and swallowing disorders [2], which are hallmarks of decreased quality of life, shorter survival time, and overall progression [3]. Among bulbar symptoms, dysarthria and dysphagia are the two most prevalent and clinically significant. Specifically, dysarthria is a motor speech disorder caused by loss of articulatory strength and control [4]. Dysphagia is associated with difficulty in safely and efficiently swallowing food and liquids, predominantly affecting the oral and pharyngeal phases of swallowing [5]. Dysphagia reduces the ability to eat normally, increasing the risk of malnutrition, aspiration pneumonia, and death [6]. Therefore, early and accurate detection of dysphagia and its decline is fundamental to gaining insights into disease progression, preventing severe clinical complications, and rapidly accessing tailored therapeutic strategies [7]. Given that the anatomical structures responsible for speech production and swallowing largely overlap [8], dysarthria and dysphagia frequently co-occur in individuals with ALS [9].

The gold standard instrumental assessment methods for swallowing are the fiberoptic endoscopic evaluation of swallowing (FEES) and the videofluoroscopic swallowing study (VFSS) [10]. However, these methods require specialized personnel, in-clinic visits, and can cause discomfort, pain, and even severe complications, such as laryngospasm [11], or require radiation exposure.

In recent years, the acoustic and video analyses of speech production have emerged as promising pathways towards more accessible methods of bulbar function assessment in ALS [12], [13]. Preliminary results have shown that acoustic and kinematic features of speech and orofacial functions may be correlated with bulbar decline in ALS [14], [15]. Recently, a growing number of studies have applied voice analysis for detecting dysphagia in various neurological conditions [16], [17], identifying acoustic parameters that may provide useful insight into dysphagia, including frequency perturbation (e.g., relative average perturbation), amplitude perturbations (e.g., shimmer), and harmonic-to-noise ratio [18]. However, these studies mainly rely on sustained phonations (e.g., vowel /*a*/) or syllable repetitions and do not exploit kinematic analysis of

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

speech. Also, to the best of our knowledge, none of the previous studies investigated dysphagia through speech acoustics in ALS.

We fill this gap by exploiting acoustic and kinematic analysis of speech to objectively describe dysphagia in ALS. Our goal is to identify objective and non-invasive metrics capable of detecting early signs of bulbar impairment, specifically swallowing difficulties.

II. METHODS

A. Data collection

Fifteen individuals with ALS, aged 52-81 years (mean = 64.2 ± 8.5 years, 7 females, 4 bulbar onset, total ALSFRS-r = 39.5 ± 5.3), were recruited at the Neurology Unit of the Pisa University Hospital. The inclusion criteria required participants to have been diagnosed with ALS, with symptom onset within 18 months from the screening visit, without a previous history of speech, swallowing, or oro-facial impairments.

Each participant was asked to perform several speech tasks, including sustained phonation of vowels, diadochokinetic repetitions of /pa/ and /pataka/, a connected speech task involving picture description, and reading aloud a passage , which was proven to provide informative metrics of ALS bulbar dysfunctions [14]. The passage was the Italian phonetically balanced text "*Il deserto*" [19].

Speech signals were recorded at 44.1 kHz using two high-quality microphones (Sennheiser MKE 200 for voice recording and Rode SmartLav+ for environmental noise), while orofacial movements were simultaneously video-recorded via a front-facing webcam (Razer Kiyo 4MP) positioned approximately 50 cm from the participant's face. A custom Python GUI was developed to streamline the recordings, ensuring synchronization between audio and video streams. Data collection was carried out in a quiet room of the Cisanello hospital in Pisa, and participants were asked to stay seated during the test.

Clinical assessments included: ALS functional rating scale revised (ALSFRS-R), Penn Upper Notor Neuron score (PUMNS), Medical Research Council (MRC) scale for the assessment of muscle strength, and instrumental assessment via FEES. In this study, we considered the pooling score [20] obtained from FEES exam, which indicates the level of bolus residue and swallowing dynamics.

B. Audio and video analyses

For each frame of the video recordings, the facial region of the participant was identified using the single-shot scale-invariant face detection model [21]. Subsequently, the 3D coordinates of 68 facial landmarks were extracted using the face alignment network [22], therefore, estimating the positions of the mouth, jawline, nose, eyes, and eyebrows. From these landmarks, we computed a set of features describing the range of motion, speed, symmetry, and shape of different parts of the face [2]. Specifically, we compute cumulative path travelled by the lower lip and by the jaw, the mean value and the range of the mouth area with respect to the rest position, the maximum and minimum mouth speed, lower lip, and jaw, the absolute difference between right and left mouth areas, the correlation between right and left mouth corners movements, the mean value of mouth eccentricity, and the range of mouth eccentricity. All the features were normalized by the intercanthal distance to ensure consistency across frames.

From the audio recordings, we extracted several acoustic features related to phonation, prosody, articulation, speech timings, and intelligibility. These features were computed using Python programming language, by exploiting Parselmouth [23], a Python library implementing Praat functions [24], and MATLAB 2024b. First, we transcribed the audio by using the WhisperX automated speech recognition (ASR) model [25], which provided both the transcription of speech recognized and word-level timestamps, with its own level of confidence. Therefore, we derived several speech timing metrics, such as inter-word interval, speech and articulation rate, articulation entropy, pause percentage, confidence score metrics, and word error rate. These features reflect several aspects of speech timings, velocity, and articulatory difficulty. Additionally, to face the difficult recognition of ASR in the case of a dysarthric speaker, we identified speech activity by applying a voice activity detection (VAD) algorithm [26], and we used these timings to compute metrics of speech/silent timings.

We also utilized Praat functions to extract additional acoustic features, including metrics of fundamental frequency (F0), jitter, shimmer, harmonics-to-noise ratio (HNR), and cepstral coefficients, as well as first, second, and third formants (F1, F2, F3, respectively). We followed the Praat guidelines [24] to impose minimum and maximum pitch frequencies and formants depending on the participant's sex. Furthermore, we derived articulatory features from vowel formant measures. In particular, we identified words with corner vowels from the ASR transcript, and then we aligned them at the phoneme level using the Wav2Vec2 force alignment (FA) algorithm. Then, we extracted F1, F2, and F3 centered at the vowel estimated time. Finally, we computed the vowel space area (VSA), formant centralization ratio (FCR), and the ratio between F2 of /*i*/ and F2 of /*u*/.

Overall, 132 features were extracted, with 106 coming from audio and 26 from video.

C. Statistical analysis

Participants were divided into two groups based on the pooling score with a solid-dry bolus [20]: participants with a pooling score lower than 6 (non-dysphagic group); and participants with a pooling score higher than or equal to 6 (i.e., associated with mild, moderate, or severe dysphagia) [20]. We chose to investigate the solid-dry score, since dysphagia for solids is generally one of the first symptoms of swallowing impairment in ALS, due to reduced tongue propulsion resulting from lingual muscle atrophy [27].

We applied the maximum relevance minimum redundancy (mRMR) algorithm to rank the features based on the outcome obtained with the pooling score. For the statistical analysis, we applied the Shapiro-Wilk test to assess normality of the data distribution, and we applied a t-test or a Mann-Whitney test to find any statistically significant difference for each feature across groups.

III. RESULTS

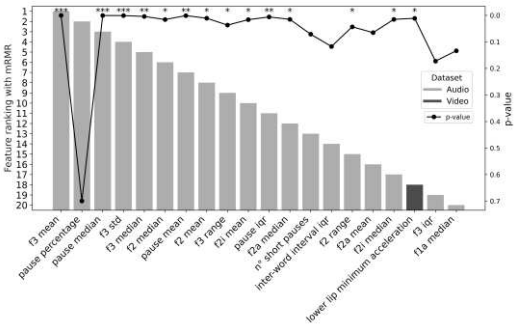


Fig. 1: top-ranked acoustic and kinematic features with the mRMR algorithm and their p-value.

Among all the participants, 14 completed all assessments, whereas one withdrew before the audio-video recording. Based on the pooling score, 5 participants belonged to the first group (no dysphagia, 66.40 ± 10.67 years, 5 females, 2 bulbar onset), with a duration of the recording equal to 131.02 ± 59.45 seconds, while the remaining 9 were included in the second group (62.00 ± 6.90 years, 2 females, 2 bulbar onset), having a recording duration of 110.65 ± 58.77 seconds. Fig. 1 shows the features ranking and the highest statistical difference. Notably, the number of statistically different acoustic features is much higher than the kinematic features, which are significant only for the minimum acceleration of the lower lip. Interestingly, the most important features according to the mRMR ranking are not necessarily those showing the strongest statistical significance between groups.

The results in Fig. 2 showed that highly significant differences exist between dysphagic and non-dysphagic groups for F3 metrics, with individuals with dysphagia showing higher values. In contrast, non-dysphagic participants showed statistically significantly lower values for pause time metrics.

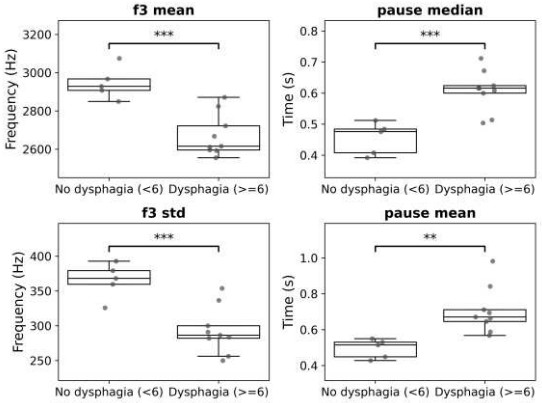


Fig. 2: boxplots of the features with the largest statistically significant differences.

IV. DISCUSSION

In this study, we analyzed audio and video recordings of ALS individuals during a speech task. We aimed to identify any acoustic and kinematic features statistically different between participants with and without dysphagia. The acoustic and kinematic features were directly extracted, considering the entire audio, which differs from all the previous studies focused predominantly on phonation [17]. Metrics of F3 showed the largest statistical difference and importance, with reduced mean values and standard deviations in the group with dysphagia.

This behavior of F3, associated with the position of the soft palate and velopharyngeal closure, aligns with the finding of [28]. Although the association between velopharyngeal deficit and solid-dry dysphagia is not yet fully understood, a reduced F3 could indicate an incomplete velopharyngeal closure, potentially contributing to less efficient swallowing. Moreover, we found pause duration metrics to be statistically different, with increased values in the case of dysphagia. These results reflect the findings of studies on bulbar impairment [29], [30], which highlighted higher pause time and lower speech rate in cases of speech impairment, especially with the bulbar onset and with the disease progression. Therefore, our work contributes to supporting and promoting future research on the identification of features that may reveal difficulties in swallowing before they become clinically obvious.

This study is limited by the small sample size and the restricted set of analyzed speech tasks. Moreover, since the audio was analyzed entirely, some important speech acoustic metrics may be averaged and are not easy to interpret. However, connected speech is a more

comprehensive task, which includes also more complex tasks than simple vowel phonation, allowing the analysis of more complete articulatory movements.

Future work should explore other tasks, speech parameters, more detailed analysis, and a larger dataset. Additionally, a longitudinal study may confirm these findings, allowing for making predictions of the bulbar decline.

V. CONCLUSION

This study highlights that acoustic and kinematic analysis of connected speech can provide valuable biomarkers for the assessment of dysphagia in individuals with ALS. Measures of F3 and pause timings emerged as the most sensitive indicators. These findings aim to support research in the use of speech-based approaches as a tool to support clinicians in the assessment and monitoring of swallowing impairment, offering a more accessible, non-invasive, and cost-effective solution to traditional assessment methods.

ACKNOLEDGEMENTS

The financial support of AriSLA – Fondazione Italiana di ricerca per la SLA is acknowledged (Project MIMOSA - Multimodal Intelligent Methods for Orofacial and Speech Assessment to predict ALS bulbar decline)

REFERENCES

[1] A. Chiò *et al.*, 'Prognostic factors in ALS: A critical review', *Amyotroph. Lateral Scler.*, vol. 10, no. 5–6, pp. 310–323, Jan. 2009, doi: 10.3109/17482960802566824.

[2] A. Bandini, J. R. Green, B. Taati, S. Orlandi, L. Zinman, and Y. Yunusova, 'Automatic Detection of Amyotrophic Lateral Sclerosis (ALS) from Video-Based Analysis of Facial Movements: Speech and Non-Speech Tasks', in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, Xi'an: IEEE, May 2018, pp. 150–157. doi: 10.1109/FG.2018.00031.

[3] S. H. Felgoise, V. Zaccheo, J. Duff, and Z. Simmons, 'Verbal communication impacts quality of life in patients with amyotrophic lateral sclerosis', *Amyotroph. Lateral Scler. Front. Degener.*, vol. 17, no. 3–4, pp. 179–183, May 2016, doi: 10.3109/21678421.2015.1125499.

[4] H. P. Rowe, S. Shellikeri, Y. Yunusova, K. V. Chenausky, and J. R. Green, 'Quantifying articulatory impairments in neurodegenerative motor diseases: A scoping review and meta-analysis of interpretable acoustic features', *Int. J. Speech Lang. Pathol.*, vol. 25, no. 4, pp. 486–499, Jul. 2023, doi: 10.1080/17549507.2022.2089234.

[5] A. Sasegbon and S. Hamdy, 'The anatomy and physiology of normal and abnormal swallowing in oropharyngeal dysphagia', *Neurogastroenterol. Motil.*, vol. 29, no. 11, p. e13100, Nov. 2017, doi: 10.1111/nmo.13100.

[6] P. E. Marik and D. Kaplan, 'Aspiration Pneumonia and Dysphagia in the Elderly', *Chest*, vol. 124, no. 1, pp. 328–336, Jul. 2003, doi: 10.1378/chest.124.1.328.

[7] K. M. Allison, Y. Yunusova, T. F. Campbell, J. Wang, J. D. Berry, and J. R. Green, 'The diagnostic utility of patient-report and speech-language pathologists' ratings for detecting the early onset of bulbar symptoms due to ALS', *Amyotroph. Lateral Scler. Front. Degener.*, vol. 18, no. 5–6, pp. 358–366, Jul. 2017, doi: 10.1080/21678421.2017.1303515.

[8] J. R. Duffy, 'Motor Speech Disorders: Clues to Neurologic Diagnosis'.

[9] B. J. Wang, F. L. Carter, and K. W. Altman, 'Relationship between Dysarthria and Oral-Oropharyngeal Dysphagia: The present evidence', *Ear. Nose. Throat J.*, p. 014556132095164, Oct. 2020, doi: 10.1177/0145561320951647.

[10] D. K.-H. Lai *et al.*, 'Computer-aided screening of aspiration risks in dysphagia with wearable technology: a Systematic Review and meta-analysis on test accuracy', *Front. Bioeng. Biotechnol.*, vol. 11, p. 1205009, Jun. 2023, doi: 10.3389/fbioe.2023.1205009.

[11] A. Nacci, F. Ursino, R. La Vela, F. Matteucci, V. Mallardi, and B. Fattori, 'Fiberoptic endoscopic evaluation of swallowing (FEES): proposal for informed consent', *Acta Otorhinolaryngol. Ital. Organo Uff. Della Soc. Ital. Otorinolaringol. E Chir. Cerv.-facc.*, vol. 28, no. 4, pp. 206–211, Aug. 2008.

[12] J. R. Green *et al.*, 'Bulbar and speech motor assessment in ALS: Challenges and future directions', *Amyotroph. Lateral Scler. Front. Degener.*, vol. 14, no. 7–8, pp. 494–500, Dec. 2013, doi: 10.3109/21678421.2013.817585.

[13] L. E. R. Simmatis, J. Robin, M. J. Spilka, and Y. Yunusova, 'Detecting bulbar amyotrophic lateral sclerosis (ALS) using automatic acoustic analysis', *Biomed. Eng. Online*, vol. 23, no. 1, p. 15, Feb. 2024, doi: 10.1186/s12938-023-01174-z.

[14] M. Neumann, H. Kothare, and V. Ramanarayanan, 'Multimodal speech biomarkers for remote monitoring of ALS disease progression', *Comput. Biol. Med.*, vol. 180, p. 108949, Sep. 2024, doi: 10.1016/j.compbiomed.2024.108949.

[15] Y. Yunusova, J. R. Green, M. J. Lindstrom, L. J. Ball, G. L. Pattee, and L. Zinman, 'Kinematics of disease progression in bulbar ALS', *J. Commun. Disord.*, vol. 43, no. 1, pp. 6–20, Jan. 2010, doi: 10.1016/j.jcomdis.2009.07.003.

[16] J. S. Ryu, S. R. Park, and K. H. Choi, 'Prediction of Laryngeal Aspiration Using Voice Analysis':, *Am. J. Phys. Med. Rehabil.*, vol. 83, no. 10, pp. 753–757, Oct. 2004, doi: 10.1097/01.PHM.0000140798.97706.A5.

[17] K. W. Dos Santos *et al.*, 'Using Voice Change as an Indicator of Dysphagia: A Systematic Review', *Dysphagia*, vol. 37, no. 4, pp. 736–748, Aug. 2022, doi: 10.1007/s00455-021-10319-y.

[18] I. Hwang, J.-M. Kim, J. S. Ryu, and K. Lee, 'Voice-Based Dysphagia Detection: Leveraging Self-Supervised Speech Representation', in *Interspeech 2025*, ISCA, pp. 5683–5687. doi: 10.21437/Interspeech.2025-761.

[19] A. Romano, U. Cesari, M. Mignano, O. Schindler, and I. Vernero, 'LA QUALITÀ DELLA VOCE', presented at the Atti dell'VIII Convegno dell'associazione Italiana di Scienze della Voce, Roma, Jan. 2012, pp. 1–34.

[20] D. Farneti *et al.*, 'The Pooling-score (P-score): inter- and intra-rater reliability in endoscopic assessment of the severity of dysphagia', *Acta Otorhinolaryngol. Ital. Organo Uff. Della Soc. Ital. Otorinolaringol. E Chir. Cerv.-facc.*, vol. 34, no. 2, pp. 105–110, Apr. 2014.

[21] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, 'S^3FD: Single Shot Scale-Invariant Face Detector', in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice: IEEE, Oct. 2017, pp. 192–201. doi: 10.1109/ICCV.2017.30.

[22] A. Bulat and G. Tzimiropoulos, 'How Far are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230,000 3D Facial Landmarks)', in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice: IEEE, Oct. 2017, pp. 1021–1030. doi: 10.1109/ICCV.2017.116.

[23] Y. Jadoul, B. Thompson, and B. De Boer, 'Introducing Parselmouth: A Python interface to Praat', *J. Phon.*, vol. 71, pp. 1–15, Nov. 2018, doi: 10.1016/j.wocn.2018.07.001.

[24] Boersma, Paul, 'Praat, a system for doing phonetics by computer' *Glot International,* 2001, 5:9/10, 341-345..

[25] M. Bain, J. Huh, T. Han, and A. Zisserman, 'WhisperX: Time-Accurate Speech Transcription of Long-Form Audio', in *INTERSPEECH 2023*, ISCA, Aug. 2023, pp. 4489–4493. doi: 10.21437/Interspeech.2023-78.

[26] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung, 'A statistical model-based voice activity detection', *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999, doi: 10.1109/97.736233.

[27] D. C. Wolf, 'Dysphagia', in *Clinical Methods: The History, Physical, and Laboratory Examinations*, 3rd ed., H. K. Walker, W. D. Hall, and J. W. Hurst, Eds., Boston: Butterworths, 1990. Accessed: Sep. 20, 2025. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK408/

[28] Y. Liang, F. A. Numan, K. Li, and G. Liao, 'Spectrum analysis of Chinese vowels formant in patients with tongue carcinoma underwent hemiglossectomy', *Int. J. Clin. Exp. Med.*, vol. 8, no. 2, pp. 2867–2873, 2015.

[29] P. Rong *et al.*, 'Predicting Speech Intelligibility Decline in Amyotrophic Lateral Sclerosis Based on the Deterioration of Individual Speech Subsystems', *PLOS ONE*, vol. 11, no. 5, p. e0154971, May 2016, doi: 10.1371/journal.pone.0154971.

[30] M. Eshghi *et al.*, 'Rate of speech decline in individuals with amyotrophic lateral sclerosis', *Sci. Rep.*, vol. 12, no. 1, p. 15713, Sep. 2022, doi: 10.1038/s41598-022-19651-1.

FEASIBILITY AND CLINICAL SIGNIFICANCE OF AN UPDATED MULTIPARAMETRIC VOICE ASSESSMENT PROTOCOL BASED ON SPEECH DIADOCHOKINETIC PARAMETERS IN PATIENTS WITH PARKINSON DISEASE

L. Franz[1], R. Cenedese[1], C. Birca[1], M. Kob[2], G. Baracca[1], C. de Filippis[1], G. Marioni[1]

[1] Phoniatrics and Audiology Unit, Department of Neuroscience DNS, University of Padova, Treviso, Italy
[2] Erich Thienhaus Institute, Detmold University of Music, Detmold, Germany

leonardo.franz@unipd.it, roberta.cenedese@unipd.it, giovanna.baracca@gmail.com, cristina.birca@unipd.it, malte.kob@hfm-detmold.de, cosimo.defilippis@unipd.it, gino.marioni@unipd.it

Abstract

***Objective*: The aim of this study was to apply motor speech diadochokinetic assessment in voice evaluation of patients with Parkinson's disease (PD).**

***Methods*: Six patients with PD were evaluated with a voice assessment protocol, including the self-assessment questionnaire, the perceptual GRB Scale, the INFVo rating scale for substitution voices, the acoustic analysis of jitter, shimmer, glottal-to-noise excitation ratio (GNE), and the motor speech diado-chokinetic parameters (DDK, DDK standard deviation, DDK jitter, and mean syllable length (MSL) in syllables with anterior, middle, and posterior articulation).**

The UPDRS scale for PD severity was measured as well.

***Results*: the median values of DDK, DDK SD, and DDK Jitter were 3.99 S/s (IQR: 2.60-4.44 S/s), 2.38 S/s (IQR: 2.12-3.39 S/s), and 2.56% (IQR: 2.15-2.72 %), respectively. In the considered PD patients, DDK was significantly higher than the normality range, p=0.0277.**

DDK standard deviation directly correlated with the same perceptive parameter (rho=0.88, p=0.0198), as well as with the tremor (rho= 0.971, p=0.0012), and the intelligibility scores (rho= 0.899, p=0.0149).

***Conclusions*: despite the limited sample size, these preliminary results suggest a promising role of motor speech diadochokinetic parameters in the clinical assessment of patients with PD. However, further studies on larger cohorts are necessary.**

Keywords: **Parkinson disease, Acoustic Voice Analysis, Motor Speech Diadochokinetics, Articulation**

I. INTRODUCTION

In Parkinson's disease (PD), about 89% of patients present with speech disorders, often as the first clinical manifestation of motor dysfunction [1]. Sometimes they are present as early as 5 years before the actual diagnosis of the disease [2] and speech and voice deficits worsen as PD progresses [3,4].

The voice of patients with PD is described as "hypokinetic dysarthria" [5], with changes in the quality, volume and pitch of the voice, including hypophonia and reduced loudness and pitch control [1,6].

The use of for laryngeal diadochokinetic parameters (LDDK) to evaluate the neuromuscular impairment of the phono-articulatory tract has been rising interest in phoniatric literature [7,8]. LDDK requires rapid abduction and adduction of the vocal folds by way of the arytenoid cartilage movements. The LDDK can be measured by commercially available acoustic software packages, which provide semiautomated means to calculate the rate and regularity of syllable repetitions. Although they are programmed for oral speech diadochokinesia, they can measure efficiently the LDDK as well [7,8].

The aim of this study was to propose a novel voice assessment protocol for PD patients, based on motor speech diadochokinetic assessment as an indicator of fluency and intelligibility.

A secondary aim was to determine whether the Motor speech diadochokinetic parameters could be associated with perceptive outcomes in patients with PD.

II. METHODS

A series of six male patients with PD (median age: 70.5 years; range: 61-80 years), referring to the Phoniatrics and Audiology Unit of the University of Padova/AULSS2 Marca Trevigiana, was considered in this preliminary evaluation.

Each patient underwent a multiparametric voice assessment, as well as a complete laryngo-stroboscopic evaluation.

The UPDRS scale was employed as well, to score PD symptom severity [9].

The voice assessment protocol is described below.

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

Perceptual evaluation
Both patients were asked to read the first paragraph of the Italian text "Il deserto". The digital recording was made with a sampling frequency of 44,100 Hz, in a quiet room with less than 40 dB of background noise. The assessment was performed by 1 otolaryngologist and 2 phoniatricians. The considered perceptive parameters were global grade of dysphonia (G), roughness (R), breathiness (B), asthenicity (A), voice tremor (t), strain (S) and instability (I) from the Extended GRBAS Scale (GS) [10]. Moreover, intelligibility (Int) and fluency (Fl) from the INFVo Scale for substitution voicing assessment [11] were considered as well. All the parameters were scored from 0 to 10 on a visuo-analogic scale, from the absence to the most severe impairment of the voice.

Self-evaluation parameters
The voice handicap index 10 (VHI-10) questionnaire [12] was employed for subjective evaluation. It consisted of 10 items, which are to be scored from 0 (healthy voice condition) to 4 (most severe voice condition) by the patient himself.

Acoustic analysis
Acoustic analysis was based on the vowel /a:/. The voice sound signal was captured and recorded by means of Lingwaves device (Wevosys), at a sampled rate of 44,100 Hz, in a quiet room with less than 40 dB of background noise. The most stable second of the sustained emission of the vowel /a:/ at a comfortable pitch and level was selected and analyzed, with the extraction of the jitter%, shimmer% and glottal to noise excitation ratio (GNE).

Motor speech diadochokinetic assessment (MSDA)
The voice signal was captured and recorded by means of Lingwaves device (Wevosys), using the MSDA protocol, at a sampled rate of 44.100 Hz, in a quiet room with less than 40 dB of background noise. The six patients were asked to repeat the English word "buttercup" (a sequence of bʌ/tʌ/kʌ) as many times as possible after a deep inspiration. The following parameters were captured: diadochokinetic index (DDK) and the DDK standard deviation (DDK SD) as a measure of the regularity of the spoken words, both in syllables per second (S/s).
The mean syllable length (MSL, expressed in ms) was calculated for those with anterior, middle and posterior articulation.
The mean energy slope across the syllables was also considered.

Statistical analysis
Continuous variables were summarized by median and interquartile range. Categorical variables were described as count and percentage in each category.
Pairwise comparison of continuous variables was performed using a two-sided sign test.
The correlation between continuous variables was investigated using the Spearman's rank correlation model. For all employed tests, statistical significance was set at p-value <0.05. Statistical analyses were performed using Stata 16.1 (College Station, TX, USA).

III. RESULTS

The median values (IQR) of DDK, DDK SD, and DDK Jitter were 3.99 S/s (2.60-4.44 S/s) [normal values ≥5.70], 2.38 S/s (2.12-3.39 S/s) [normal values ≤2.90], and 2.56% (2.15-2.72 %) [normal values ≤3.00], respectively.
In the considered patients, DDK was significantly higher than its normality range (p=0.0277).
DDK values showed a trend towards an inverse correlation with the asthenia score (rho=-0.79, p=0.059, see also Fig. 1).

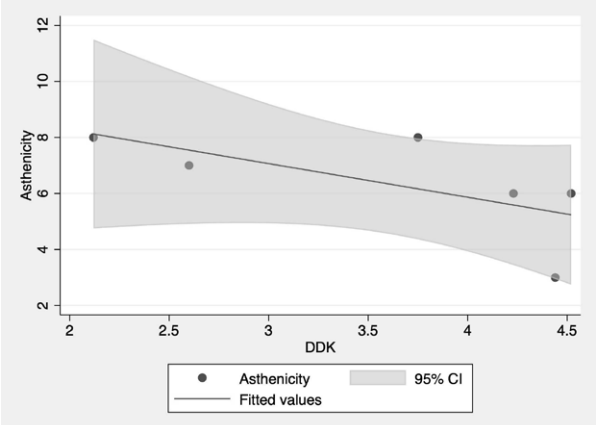


Fig. 1 Inverse correlation between DDK values and asthenia score.

On the other hand, DDK SD directly correlated with the asthenia (rho=0.88, p=0.0198; see also Fig. 2), as well as with the tremor (rho= 0.971, p=0.0012; see also Fig. 3), and the intelligibility scores (rho= 0.899, p=0.0149; see also Fig. 4).

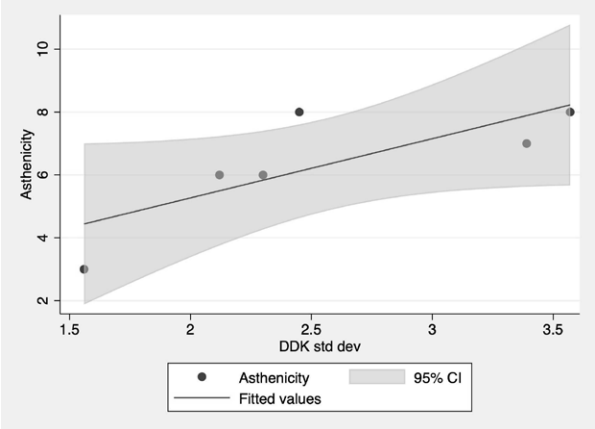


Fig. 2 Direct correlation between DDK SD values and asthenia score.

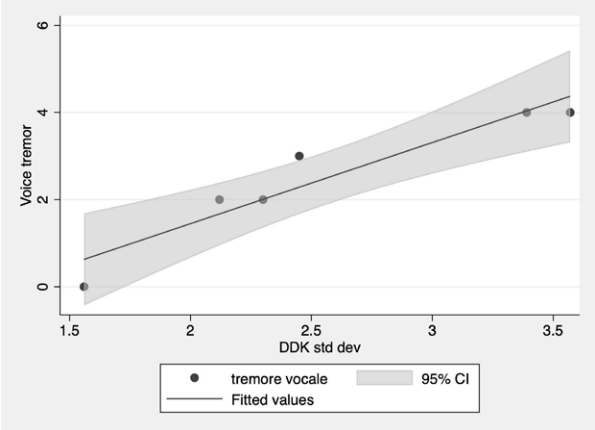


Fig. 3 Direct correlation between DDK SD values and tremor score.

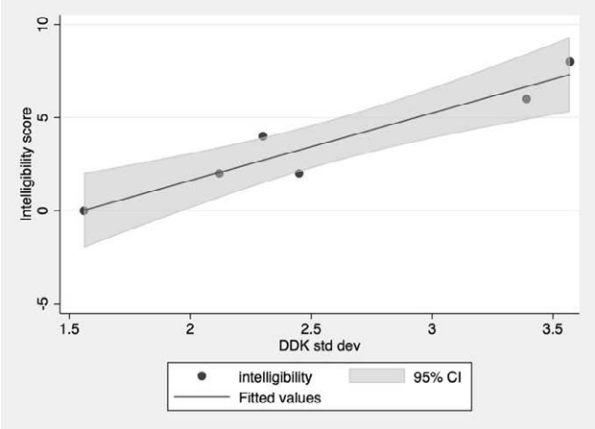


Fig. 4 Direct correlation between DDK SD values and impaired intelligibility score.

Moreover, the energy slope appeared to be correlated with increasing roughness score (rho=0.9411, p=0.0051; see also Fig. 5).

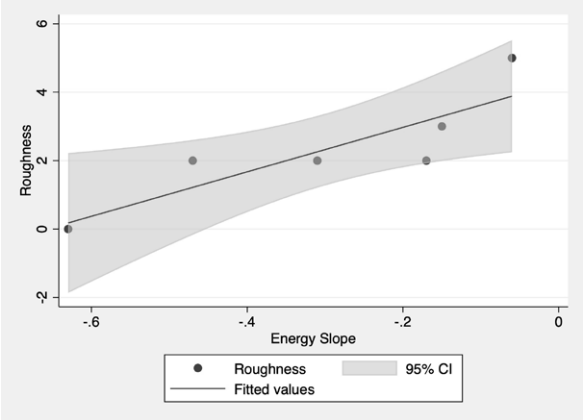


Fig. 5 Direct correlation between energy slope and roughness score.

Also, the MSL appeared to be corelated with perceptual parameters: increasing length of anterior syllables correlated with vocal strain (rho=0.9411, p=0.0051), while MSL of middle syllables correlated with increasing asthenia scores (rho=0.8827, p=0.0198).

IV. Discussion

The possibility to quantitatively asses voice and articulation parameters in patients with neuro-degenerative disease is rising increasing interest, also as an objective clinimetric tool [13].

This study's results seem to indicate a substantial agreement between motor speech diadochokinetic parameters and perceptive voice characteristics in patients with PD.

In particular, the DDK score (referring to the syllable count per time unit) appeared to inversely correlate with the asthenia score, which may be a perceptive voice correlate of bradykinesias and weakness of expiratory, laryngeal, pharyngeal and tongue muscles.

This seems to be conceptually in lie with the observation by Hähnel and colleagues [14], regarding a reduced syllable count in patients with increased UPDRS clinical severity score (and therefore a more marked motor impairment).

Interestingly, the articulatory variability, which might be quantitatively represented by the DDK SD, seemed to be related not only with asthenia (as a speech correlate of motor impairment), but also with impaired intelligibility, and voice roughness. Beside highlighting a substantial consistency between motor impairment correlates, articulatory disorganization, and reduced communication effectiveness, these results seem to show also a possible relationship between diadochokinetic and voice harshness in PD.

Both those aspects may reflect a pathophysiological substrate similar to the one described in PD and

Parkinson-like diseases by previous studies [14, 15], which however focused on different parameters to quantify articulatory and voice outcomes.
Finally, regarding MSL, Hähnel et al. [14] found that prolonged syllable duration related with clinical severity of disease. As a further characterization, in our study, MSL was stratified by syllable type, finding that variations in the duration of anterior and middle syllables might reflect differences in terms of perceptive voice quality.
The main limitation of this study resides in the small sample size and the absence of a control group, suggesting caution in generalizing the results. Moreover, the considered series included only male patients, thus limiting the possibility of evaluating potential gender-related effects on LDDK.
Another possible limitation regards the fact that differences in dopaminergic drug response or in ON/OFF state may affect speech characteristics. However, to date, evidence from the literature do not clearly indicate whether speech analysis should be conducted in the ON or OFF state [14,16].
On the other hand, the main strength of this study resides in the substantial homogeneity of the included cases, in terms of disease, demographics, and evaluation methods. The assessment protocol allowed a quantitative analysis not only of the DDK, but also of other parameters (including DDK SD and DDK jitter, as well as the energy slope, and the MSL for each syllable type), which allowed to precisely characterize speech and voice features of the considered patients.

V. Conclusion

These preliminary results seem to suggest a promising role of motor speech diadochokinetic parameters in the clinical assessment of patients with PD. To characterize the diagnostic yield of this approach compared to the traditional acoustic analysis, further studies on larger series are needed.

REFERENCES

[1] Liu V, Smith D, Yip H. Prevalence and Treatment of Dysphonia in Parkinson's Disease: A Cross-Sectional National Database Study. Laryngoscope Investig Otolaryngol. 2025 May 14;10(3):e70149. doi: 10.1002/lio2.70149. PMID: 40370339; PMCID: PMC12076596.

[2] B. Harel, M. Cannizzaro, PJ Snyder, 2Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: a longitudinal case study" Brain Cogn. 2004;56:24–29. https://doi.org/ 10.1016/j.bandc.2004.05.002.

[3] RJ Holmes, JM Oates, DJ Phyland, AJ Hughes, "Voice characteristics in the progression of Parkinson's disease", Int J Lang Commun Disord. 2000;35:407–418. https://doi.org/10.1080/136828200410654.

[4] J Rusz, T Tykalová, M Novotný, et al. "Distinct patterns of speech disorder in early-onset and late-onset de-novo Parkinson's disease" Npj Parkinson's Dis. 2021;7:98.

[5] EM Critchley, "Speech disorders of Parkinsonism: a review", J Neurol Neurosurg Psychiatry 1981;44(9):751–8.

[6] A. Ma, KK Lau, D. Thyagarajan "Voice changes in Parkinson's disease: What are they telling us?" J Clin Neurosci.2020Feb;72:1-7. doi: 10.1016/j.jocn.2019.12.029. Epub 2020 Jan 14. PMID: 31952969.

[7] L. Lombard, N.P. Solomon, "Laryngeal Diadochokinesis Across the Adult Lifespan", J. Voice, vol. 34, pp. 651-656, 2020.

[8] T. Louzada, R. Beraldinelle, G. Berretin-Felix, et al "Oral and vocal fold diadochokinesis in dysphonic women", J. Appl. Oral Sci, vol. 19, pp. 567– 572, 2011.

[9] P. Martinez-Martin, C. Rodriguez-Blazquez, M. Alvarez-Sanchez, et al. "Expanded and Independent Validation of the Movement Disorder Society–Unified Parkinson's Disease Rating Scale (MDS-UPDRS)", Journal of Neurology, vol. 260, pp. 228–236, 2013.

[10] A. Ricci-Maccarini et al., "Validity, Reliability and Reproducibility of the "Extended GRBAS Scale," A Comprehensive Perceptual Evaluation of Dysphonia", Journal of Voice, Volume 39, Issue 2, 393 – 402

[11] M.B.J. Moerman, J. Martens, M. Van der Borgt, M, "Perceptual evaluation of sub-stitution voices: Development and evaluation of the (I)INFVo rating scale", Eur. Arch. Otorhinolaryngol, vol. 263, pp. 183–187, 2006.

[12] C.A. Rosen, A.S. Lee, J. Osborne, et al, "Development and validation of the voice handicap index-10", Laryngoscope, vol. 114, pp. 1549-1556, 2004.

[13] U. De Silva, S. Madanian, S. Olsen, JM Templeton

, C. Poellabauer, SL Schneider, A Narayanan, R. Rubaiat, "Clinical Decision Support Using Speech Signal Analysis: Systematic Scoping Review of Neurological Disorders". J Med Internet Res. 2025 Jan 13;27:e63004. doi: 10.2196/63004

[14] T. Hähnel, A. Nemitz, K. Schön, L. Berger, A. Vogel, D. Gruber, N. Schnalke, S. Bräuer, BH Falkenburger, F. Gandor, "Speech Differences between Multiple System Atrophy and Parkinson's Disease2 Mov Disord Clin Pract. 2025 May 3. doi: 10.1002/mdc3.70094.

[15] J. Rusz, T. Tykalová, G. Salerno, S. Bancone, J. Scarpelli, MT Pellecchia, "Distinctive speech signature in cerebellar and parkinsonian subtypes of multiple system atrophy" J Neurol. 2019 Jun;266(6):1394-1404. doi: 10.1007/s00415-019-09271-7.

[16] R. Norel, C. Agurto, S. Heisig, JJ Rice, H. Zhang, R. Ostrand, PW Wacnik, BK Ho, VL Ramos, GA Cecchi, "Speech-based characterization of dopamine replacement therapy in people with Parkinson's disease" NPJ Parkinsons Dis. 2020 Jun 12;6:12. doi: 10.1038/s41531-020-0113-5

CNN-BASED SCREENING OF NEONATAL PATHOLOGIES USING INFANT CRY AS A BIOMARKER

M. A. Ruiz-Diaz[1,2], C. A. Reyes-Garcia [1], H. Perez-Espinosa[1]

[1] Department of Biomedical Sciences and Technologies, National Institute of Astrophysics, Optics and Electronics (INAOE), Puebla, México.
[2] Faculty of Computer Science, Meritorious Autonomous University of Puebla (BUAP), México.
antonia.ruiz@inaoep.mx, kargaxxi@inaoep.mx, humbertop@inaoep.mx

Abstract: **Infant cry analysis is an emerging field in biomedical engineering for assessing neonatal health. It offers a non-invasive, simple, and low-cost method to detect physical and physiological conditions. This work explores its potential as a reliable tool for early pathology detection, supporting timely medical intervention. A deep learning approach based on a convolutional neural network (CNN) was developed using spectrograms and Mel Frequency Cepstral Coefficients (MFCC) as input features. Two strategies were evaluated: direct use of crying signals from the database and preprocessing by removing silent segments before MFCC extraction. Unlike binary schemes, this study addressed a multiclass classification problem, categorizing crying signals into asphyxia, deafness, hyperbilirubinemia, hypothyroidism, and healthy infants. The CNN achieved high classification accuracy per class: 0.94 for asphyxia, 1.00 for deafness, 0.92 for hyperbilirubinemia, 0.95 for hypothyroidism, and 1.00 for healthy infants. Model performance was assessed with accuracy, sensitivity, and specificity across multiple classes. These results show that CNNs can achieve strong performance without extensive preprocessing.**

Keywords: **Neonatal screening, signal processing, classification, MFCC, deep learning.**

I. INTRODUCTION

Immediate care for newborns with any pathology or birth defect should be a priority, as well as the prevention and detection of hereditary and congenital conditions and diseases [1]. When a baby is born, a specific medical protocol must be applied to assess its health status. Neonatal screenings are essential for the early detection and treatment of conditions that can affect the development and quality of life of the baby. Infant cry is a sequence of motor performances and associated acoustic manifestations including vocalization, constrictive silence, coughing, choking, interruptions, or various combinations of such performance [2]. In the pediatric literature, it has been proposed that infant crying is a reflection of complex neurophysiological functions and that the analysis of infant crying can be used to evaluate the health status of the infant [3]. Several research works have shown that infant crying contains information about the baby, and through it, it's possible to identify emotions [4,5], sensations and physiological states [6,7,8], health condition [9,10,11], gender [12], diseases (anomalies) [13,14], first cry [13], premature vs term [13,15], etc.

Throughout infant crying research, the use of traditional machine learning classifiers such as KNN, SVM and GMM has been explored in tasks of recognition and classification of crying signals, in recent years have been developed neural network architectures such as CNN and RNN have been applied.

This work presents the classification of five types of crying: perinatal asphyxia, hyperbilirubinemia, hypothyroidism, deafness, and healthy showing that cry analysis can be considered as a screening tool in newborns to detect certain pathologies that allow medical doctors to direct babies to specialized medical services and timely treatment.

II. MATERIALS AND METHODS

A. Databases

For the development of this research, three databases belonging by the National Institute of Astrophysics, Optics and Electronics were used.

BabyChillanto database is a collection of Mexican samples, comprises cry samples from 6 babies with asphyxia, 6 with deafness, 5 healthy babies, 23 babies with pain label, and 33 babies with the hunger label. It is important to note that the last two classes correspond to healthy babies under an episode of pain or hunger during the recording.

BabyChillanto database II is a complement to the Baby Chillanto database, which incorporates two new classes: hyperbilirubinemia and hypothyroidism. The new classes contain samples from 9 babies with hyperbilirubinemia and 47 with hypothyroidism.

Recordings from BabyChillanto I and II were collected and labeled by doctors at the National Rehabilitation Institute (INR) of Mexico, from infants aged 2 days to 6 months.

The last database used corresponds to Mexican babies from indigenous communities in the state of

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

Guerrero, Mexico. Database contains samples of 35 babies and information on the diagnosed pathologies of the babies.

Gender and gestational age of the infants were not reported, so their influence on cry characteristics could not be assessed.

The methodology carried out in this research consists of the following stages: signal preprocessing, feature extraction and classification.

In this research, we focus on the use of Mel-frequency cepstral coefficients (MFCC) for acoustic characterization due to their extensive application in infant cry classification studies, such as those conducted by [16,17], which have shown promising results. Additionally, our previous work [18] employed a convolutional neural network for infant cry classification, utilizing spectrogram and MFCC image representations as input features.

B. Signal pre-processing

For the signal pre-processing stage, two approaches were carried out. a) The first consisted of taking the original crying signals and dividing them into segments of 1-second duration. b) Developed from experimental tests, the second approach was to remove the silent segments from the original samples, concatenate the audible segments to form a single cry unit, and divide the compacted sample into 1-second-long segments.

Python was used to perform all preprocessing steps. For each sample in the database, the split_on_silence function from the Pydub library was used to obtain a list of non-silent segments with the following parameters: min_silence_len = 500, which defines the minimum duration (in milliseconds) of a silence section, and silence_thresh = -40, wich sets the silence threshold in decibels relative to full scale (bBFS). All non-silent segments in the list were concatenated to form a single cry unit. Each resulting cry unit was then divided into individual 1-second segments using a custom function implemented with the AudioSegment class from Pydub. Prior to feature extraction, all cry recordings were resampled to a uniform sampling frequency of 22.05Hz using the librosa library to ensure consistency across datasets with potentially different original sampling rates.

C. Feature extraction

We decided to use MFCC due to some of their main attributes: MFCC capture sound features in a way that emulate how humans perceive pitch and intensity and are effective representing the acoustic content of the signal. In addition, they allow for relatively fast and easy processing and analysis due to their ability to reduce dimensionality and retain most of the important information in the cry wave, making them easy to implement in real systems.

At this stage, the Python library Librosa was used, a tool specifically designed for the analysis of music and audio. The librosa.feature.mfcc function was applied to calculate Mel-frequency cepstral coefficients from an audio spectrogram. For each 1-second segment, 40 MFCC coefficients were extracted using the *librosa* library with default parameter values (n_fft = 2048, hop_length = 512). The n_fft parameter specifies the number of FFT points used to compute the frequency spectrum. By default, the window length (win_length) is set equal to n_fft, corresponding in this case to 2048 samples (approximately 93 ms at a sampling rate of 22.05 kHz). The hop_length parameter defines the number of samples between successive analysis frames (512 samples in this configuration. These coefficients provide a compact representation of the short-term spectral characteristics of the audio signal, which are commonly used in speech and cry analysis tasks.

D. Classification

In this work, the use of a convolutional neural network that receives as parameters the MFCC characteristics extracted from the signal is proposed, having as output parameters the classes: healthy, asphyxia, deafness, hyperbilirubinemia and hypothyroidism. The convolutional neural network architecture used in this work was proposed in [19] for the classification of crying signals into: asphyxia, deaf, healthy, hunger, and pain. Table 1 provides an overview of the CNN architecture applied in our experiments.

Table 1. Convolutional Neural Network (CNN) Architecture

Layer (type)	Output Shape	Param#
conv2d (Conv2D)	(None, 39, 43, 16)	80
max_pooling2d (MaxPooling2D)	(None, 19, 21, 16)	0
dropout (Dropout)	(None, 19, 21, 16)	0
conv2d_1 (Conv2D)	(None, 18, 20, 32)	2,080
max_pooling2d_1 (MaxPooling2D)	(None, 9, 10, 32)	0
dropout_1 (Dropout)	(None, 9, 10, 32)	0
conv2d_2 (Conv2D)	(None, 8, 9, 64)	8,256
max_pooling2d_2 (MaxPooling2D)	(None, 4, 4, 64)	0
dropout_2 (Dropout)	(None, 4, 4, 64)	0
conv2d_3 (Conv2D)	(None, 3, 3, 128)	32,896
max_pooling2d_3 (MaxPooling2D)	(None, 1, 1, 128	0

III. RESULTS

For all experiments, the databases were divided into training, validation, and test sets. The division was performed by individual, ensuring that all samples from the same individual were assigned to a single set and not shared across sets. This approach guarantees that the model is evaluated on unseen individuals, preventing overestimation of performance. The distribution of 1-second segments across the training, validation, and test sets is shown in Table 2.

Table 2. Distribution of 1-second segments per class in the training (Tr), validation (Val), and test (Te) sets across experiments 1–6. Class 1: Asphyxia, Class 2: Deaf, Class 3: Hyperbilirubinemia, Class 4: Hypothyroidism, Class 5: Healthy.

	Experiments								
	1			2			3		
Class	Tr	Val	Te	Tr	Val	Te	Tr	Val	Te
1	177	113	56	85	113	56	177	113	56
2	588	235	155	381	68	67	588	235	155
3	299	72	56	288	72	56	299	72	56
4	573	199	163	573	199	163	573	199	163
5	392	53	92	377	53	91	941	96	104
	Experiments								
	4			5			6		
Class	Tr	Val	Te	Tr	Val	Te	Tr	Val	Te
1	85	113	56	1,435	339	346	1,371	339	254
2	381	68	67	588	235	155	381	235	67
3	288	72	56	712	66	87	673	66	87
4	573	199	163	573	199	163	573	199	163
5	923	96	104	941	96	104	923	96	104

To evaluate the impact of dataset composition on model performance, a series of dataset-level experiments were conducted. Unlike a traditional ablation study that modifies the model architecture, these experiments explore how variations in the training data affect classification results. Experiment 1: Uses the original samples from the Baby Chillanto II database for the five classes (asphyxia, deafness, hyperbilirubinemia, hypothyroidism, and healthy). Experiment 2: Same samples as Experiment 1, but silent segments were removed from each sample. Experiment 3: Expands the healthy class by including samples labeled as pain or hunger from Baby Chillanto II. These represent healthy babies experiencing non-pathological episodes. The resulting healthy set grew from 5 to 60 individuals, with a total of 1,141 1-second segments. Experiment 4: Same as Experiment 3, but silent segments were removed.

In Experiment 5, to increase the number of samples from our Baby Chillanto II database, we incorporated additional samples from the database collected in the northern mountain region of Guerrero, Mexico, as follows:

Asphyxia set: In this work, the samples from the Sierra of Guerrero database labeled with the RDS pathology (Respiratory Distress Syndrome) are used for the training and validation phases of the CNN and the Asphyxia samples from the BabyChillanto database are used taken for the testing phase.

Hyperbilirubinemia set: In the Sierra of Guerrero database, there are three baby samples labeled with hyperbilirubinemia pathology, which were added to the class, which have a total of 12 individuals and 865 of 1-second segments.

The last experiment consists of taking the samples from experiment 5 and eliminating the silent segments. Table 3 presents the class-wise accuracy results obtained for the five-class classification task across all experiments. Each row corresponds to a specific class, while each column represents a different experiment.

Table 3. Class-wise accuracy across six experiments for the five-class classification task. Class 1: Asphyxia, Class 2: Deaf, Class 3: Hyperbilirubinemia, Class 4: Hypothyroidism, Class 5: Healthy.

	Experiments					
Class	1	2	3	4	5	6
1	0.45	0.01	0.00	0.01	0.94	0.92
2	0.92	0.81	0.91	0.81	1.00	0.93
3	1.00	1.00	0.99	0.99	0.92	0.97
4	0.94	0.99	0.93	0.99	0.95	0.96
5	0.15	0.00	0.95	1.00	1.00	0.93

IV. Discussion and Conclusion

This work shows crying signals can be very useful for detecting specific pathologies with high precision and can be the basis for introducing a crying-based screening system.

The best results of the convolutional neural network performance can be seen in Table 3, experiment 5, with 0.94 for the asphyxia class, 1.00 for the deaf class, 0.92 for the hyperbilirubinemia class, 0.95 for the hypothyroidism class, and 1.00 for the healthy class.

Table 4 shows sensitivity and specificity metrics, in experiment 1, the sensitivity for the Asphyxia and Healthy classes was very low, indicating a high risk of false negatives. In experiment 2, despite achieving high specificity across all classes, the extremely low sensitivity for Asphyxia persists, showing that the model continues to produce false negatives for this critical class. In experiment 3 and 4 a sensitivity of 0.00 and 0.01 was observed for asphyxia, indicating that the model is completely ineffective at detecting Asphyxia cases. Experiment 5 shows the best overall performance, especially for critical classes, both sensitivity and specificity improved significantly, with virtually no false negatives or false positives for the Asphyxia class. Finally, Experiment 6 also represents a well-performing and balanced model, minimizing both false positives and false negatives across nearly all classes.

Considerable fluctuations in class-wise accuracies can be observed across experiments and classes. These variations are primarily due to differences in the number of training samples per class. Classes with fewer samples tend to show lower or more variable accuracy, as the network has less data to learn discriminative features. It can also be observed that increasing the number of samples positively impacts the performance of the convolutional neural network, leading to higher and more stable accuracies in classes with more data.

Regarding the approach of eliminating silent segments from the signal, it can be seen in Table 3, the difference in precision between keeping them in the samples or eliminating them is not relevant; we could even state that in some cases, such as deafness, spaces of silence represent an important characteristic of the

class. An important point to note is that there are no restrictions on the sample acquisition process. The samples that make up the final database come from different sources and different acquisition protocols, and the proposed convolutional neural network is able to classify with high precision. Results are very encouraging, this gives a guideline to think that a system could be developed that does not require complex, advanced or high cost technology, to be accessible to medical rural units with limited resources.

Table 4. Sensitivity and specificity results by class.

	Experiment 1		Experiment 2	
Class	Sensitivity-Specificity		Sensitivity-Specificity	
Asphyxia	0.45	0.92	0.01	0.98
Deaf	0.92	0.99	0.81	1.00
Hyperbilirubinemia	1.00	0.90	1.00	0.63
Hypothyroidism	0.94	0.99	0.99	0.99
Healthy	0.15	0.96	0.0	0.98
General	0.69	0.95	0.56	0.92
	Experiment 3		Experiment 4	
Class	Sensitivity-Specificity		Sensitivity-Specificity	
Asphyxia	0.0	1.00	0.01	1.00
Deaf	0.91	1.00	0.81	1.00
Hyperbilirubinemia	0.99	0.96	0.99	0.95
Hypothyroidism	0.93	0.99	0.99	1.00
Healthy	0.95	0.81	1.00	0.77
General	0.76	0.95	0.76	0.94
	Experiment 5		Experiment 6	
Class	Sensitivity-Specificity		Sensitivity-Specificity	
Asphyxia	0.94	1.00	0.92	1.00
Deaf	1.00	1.00	0.93	1.00
Hyperbilirubinemia	0.92	0.97	0.97	0.96
Hypothyroidism	0.95	1.00	0.96	0.99
Healthy	1.00	0.98	0.93	0.98
General	0.96	0.99	0.94	0.99

Our objective is to continue developing a screening system based on infant crying by creating a new database of samples, which will include demographic information such as gender and gestational age, as well as clinical scores including Apgar and Silverman. Future work will also focus on incorporating qualitative and disease-specific features to enhance the system's diagnostic capabilities.

REFERENCES

[1] Ley general de salud, Cámara de Diputados del H. Congreso de la Unión, (2023).

[2] R. Prescott, Infant cry sound; developmental features, The Journal of the Acoustical Society of America 57 (1975) 1186–1191. doi:10.1121/1.380577.

[3] H. A. Patil, “Cry Baby”: Using Spectrographic Analysis to Assess Neonatal Health Status from an Infant’s Cry, Springer US, 2010, pp. 323–348. doi:10.1007/978-1-4419-5951-5.

[4] K. S. Alishamol, T. T. Fousiya, K. J. Babu, M. Sooryadas, L. Mary, System for infant cry emotion recognition using dnn, IEEE, 2020, pp. 867–872. doi:10.1109/ICSSIT48917.2020.9214198.

[5] S. Yamamoto, Y. Yoshitomi, M. Tabuse, K. Kushida, T. Asada, Recognition of a baby’s emotional cry towards robotics baby caregiver, International Journal of Advance Robotic Systems 10 (2013) 86. doi:10.5772/55406.

[6] C.-Y. Chang, L.-Y. Tsai, A CNN-Based Method for Infant Cry Detection and Recognition, 2019, pp. 786–792. doi:10.1007/978-3-030-15035- 8 76.

[7] R. J. Rosen, D. Tagore, T. J. Iyer, N. Ruban, A. N. J. Raj, Infant mood prediction and emotion classification with different intelligent models, IEEE, 2021, doi:10.1109/INDICON52576.2021.9691601.

[8] K. Teeravajanadet, N. Siwilai, K. Thanaselanggul, N. Ponsiricharoenphan, S. Tungjitkusolmun, P. Phasukkit, An infant cry recognition based on convolutional neural network method, IEEE, 2019, pp. 1–4. doi:10.1109/BMEiCON47515.2019.8990191.

[9] H. F. Alaie, L. Abou-Abbas, C. Tadj, Cry-based infant pathology classification using gmms, Speech Communication 77 (2016) 28–52. doi:10.1016/j.specom.2015.12.001.

[10] M. Hariharan, S. Yaacob, S. A. Awang, Pathological infant cry analysis using wavelet packet transform and probabilistic neural network, Expert Systems with Applications 38 (2011) 15377–15382. doi:10.1016/j.eswa.2011.06.025.

[11] H. A. Patil, A. T. Patil, A. Kachhi, Constant q cepstral coefficients for classification of normal vs. pathological infant cry, IEEE, 2022, pp. 7392–7396. doi:10.1109/ICASSP43922.2022.9746946.

[12] C. Ji, Y. Jiao, M. Chen, Y. Pan, Infant Cry Classification Based-On Feature Fusion and Mel-Spectrogram Decomposition with CNNs, 2022, pp. 126–134. doi:10.1007/978-3-031-23504-7 10.

[13] O. Wasz-Höckert, K. Michelsson, J. Lind, Twenty-Five Years of Scandinavian Cry Research, Springer US, 1985, pp. 83–104. doi.org/10.1007/978-1-4613-2381-5_4

[14] O. F. R. Galaviz, C. A. R. Garcia, Infant Cry Classification to Identify Hypoacoustics and Asphyxia with Neural Networks, 2004, pp. 69–78. doi:10.1007/978-3-540-24694-7_8.

[15] S. Orlandi, C. A. R. Garcia, A. Bandini, G. Donzelli, C. Manfredi, Application of pattern recognition techniques to the classification of full-term and preterm infant cry, Journal of Voice 30 (2016) 656–663. doi:10.1016/j.jvoice.2015.08.007.

[16] S. Bano, K. RaviKumar, Decoding baby talk: A novel approach for normal infant cry signal classification, IEEE, 2015, pp. 1–4. doi:10.1109/ICSNS.2015.7292392.

[17] G. Z. Felipe, R. L. Aguiar, Y. M. G. Costa, C. N. Silla, S. Brahnam, L. Nanni, S. McMurtrey, Identification of infants’ cry motivation using spectrograms, IEEE, 2019, pp. 181–186. doi:10.1109/IWSSIP.2019.8787318.

[18] C. A. Reyes-Garcia, Valencia-Hernandez, O. F. I. A. Reyes-Galaviz, Infant cry for pathologies classification using a deep learning approach, Firenze University Press, 2023, pp. 57–60. doi:10.36253/979-12-215- 0146-9.

SPECIAL SESSION II
HISTORICAL ASPECTS OF VOICE RECORDINGS AND ANALYSIS

Organized by P.H. Dejonckere

HUMANITY'S FIRST VOICE RECORDINGS

P.H. DeJonckere [1]

[1] Federal Agency for Occupational Risks, Brussels, Belgium
ph.dejonckere@outlook.com

Abstract: **The first functional sound recording device was developed by Léon Scott de Martinville in 1857. In March 25th of that year, he obtained patent No. 31470 for a method of drawing or writing from sound. Scott's "phonautograph" consisted of an acoustic horn connected to a diaphragm (membrane) made of rubber, to which was attached a stylus made of a boar's hair. This stylus continuously traced the oscillations on a strip of smoke-blackened paper wrapped around a rotating cylinder. At the time, it was not possible to reproduce the sounds, but this device played a crucial role in developments that followed two decades later. Modern technology has made these phonautograms audible and in 2015, UNESCO inscribed the invention of the phonautograph on the International Memory of the World Register, highlighting the cultural and scientific value of these unique sound archives, celebrated as the heritage of all mankind.**

Keywords: **First voice recording, phonautograph, Scott de Martinville.**

I. INTRODUCTION

On October 9, 2015, the UNESCO (United Nations Educational, Scientific and Cultural Organization) inducted Edouard-Léon Scott de Martinville's phonautograms and manuscripts onto its prestigious 'Memory of the World Register', highlighting the cultural and scientific value of these unique sound archives. These recordings have been recognized as the first recordings of the human voice and celebrated as the heritage of all mankind [1]

This was the result of a research project conducted by David Giovannini and Patrick Feaster [1]. On March 28th, 2008, the research team publicly presented for the first time at Stanford University the world's oldest playable voice recordings which had been made in 1857 – 1860 by Scott de Martinville (Fig. 1), but never heard or made audible since that time, even by their creator. This predates Edison's invention of the phonograph, two decades later.

Fig. 1 Edouard Léon Scott de Martinville 1817-1879)

It was also an innovation in archiving and digital restoration methods. Scott's device captured sound vibrations as squiggly lines on soot-blackened paper, but unlike Edison's phonograph, it wasn't designed to be played back - only studied visually.

II. THE GREAT PRECURSORS

Capturing, making visible and preserving sound emissions, particularly vocal ones, has always fascinated mankind. Leonardo da Vinci (1452–1519) is believed to have observed the behavior of particles on vibrating surfaces, noting how they formed distinct patterns in response to sound. Though no formal publication exists, references to such observations appear in his notebooks, particularly in the *Codex Atlanticus* and *Codex Madrid II*, which contain sketches and reflections on acoustics and mechanical resonance [2]. So did Galileo Galilei (1564-1642), who described in his *Discorsi e dimostrazioni matematiche intorno a due nuove scienze* (1638), how bristles or particles placed on the sounding board of musical instruments would remain motionless in certain areas, intuitively identifying nodal points of vibration [3].

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

Robert Hooke (1635–1703) conducted a pivotal experiment on July 8, 1680, where he ran a violin bow along the edge of a glass plate covered with flour. He observed the emergence of nodal patterns - a precursor to Chladni figures - marking one of the earliest physical visualizations of standing wave phenomena.

Ernst Chladni (1756 - 1827) exhaustively investigated the patterns formed by these lines, what are now called *Chladni figures* [4].

In 1807, the physicist Thomas Young described a method to graphically record the vibrations of a tuning fork (invented in 1711 by John Shore, G. F. Händel's favorite trumpeter), using a stylus to trace the fork's oscillatory movement on a soot-covered rotating cylinder. This device, which he called a "*vibrograph*", is an early example of graphically representing sound vibrations and is considered a precursor to sound recording devices.

III. THE BREAKTHROUGH OF SCOTT

However, in all these experiments, what was made visible was the vibration of an object (plate, tuning fork, rod, etc.). A decisive breakthrough was achieved by Scott by making visible and recording ***air vibrations*** (particularly those produced by ***vocal emissions***). Scott actually was an editor and typographer of manuscripts at a scientific publishing house in Paris. One day in 1853 or 1854 he was pouring over a text on human physiology of the ear when he envisioned an amazing new possibility : If photography could capture fleeting images with lenses modeled on the eye, might not a replica of the ear similarly capture spoken words? Scott's earliest ambition was actually rooted in a stenographic ambition i.e. ***to write as fast as one speaks***. Trained in stenography and frustrated by its limitations, he envisioned a device that could inscribe the spoken word directly onto a physical medium, bypassing the need for manual shorthand. He imagined a mechanical ear that could transcribe speech automatically. His 1849 treatise *Histoire de la sténographie* reflects this preoccupation, offering both a historical critique of stenographic systems and a speculative proposal for a "natural" form of writing—one that would allow sound itself to become script [5].

His first attempt at building a machine to transcribe sound was made in either 1853 or 1854 (his own notes and letters to the French Academy give contradictory dates).

He imagined using an acoustic horn (as sound collector : an idea of Athanasius Kircher) and a rubber membrane (mimicking the eardrum) capable of vibrating in correspondence with sounds in the air. To the membrane was attached a small stylus (mimicking the middle ear ossicles) - in fact a boar's bristle, about 1 cm in length - which had the required flexibility and low inertia. The acoustic signal was inscribed on a rotating cylinder wrapped with smoke-blackened paper and cranked manually (Fig. 2 & 3). He called the device "phonautograph": It captured sound vibrations as squiggly lines on the soot-covered paper, making the phonautograph a rudimentary precursor to the oscilloscope, but unlike Edison's phonograph, it was not designed to play back...

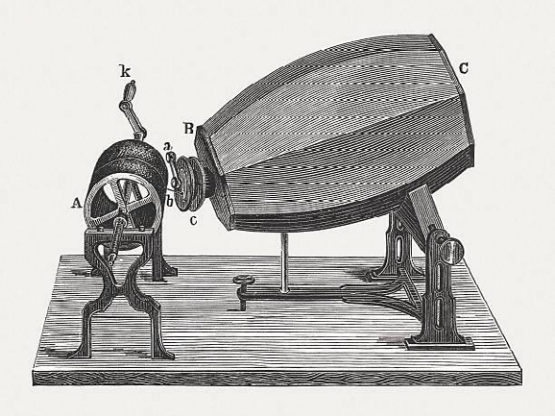


Fig. 2 Scott de Martinville's phonautograph (1857)

Scott deposited a sealed envelope (containing text, drawings and recordings) at the Academy of Sciences in Paris on 26 January 1857 and filed a patent application on 25 March 1857 [1,6,7].

On 9 April 1860, he recorded 'Au clair de la lune', the oldest preserved (and now made audible) recording of the human voice. At the time, these phonautograms could not be reproduced as sound**.**

Scott's carefully documented experiments were logged upon receipt and reported in contemporaneous publications [8,9].

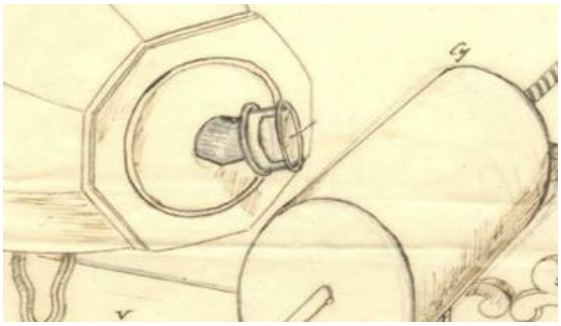


Fig. 3 Enlargement of Scott's original drawing, showing how, at the end of the acoustic horn, the membrane to which the boar's hair is attached is slightly offset from the axis of the horn, in order to allow for optimal vibration inscription on the rotating cylinder.

Scott's started around 1859 a collaboration with the German-born acoustician Rudolph Koenig. Koenig, renowned for his precision instruments, helped refine and commercialize the phonautograph.

Scott's phonautograph was manufactured and marketed from 1859 as a laboratory instrument for sound analysis in the nascent field of acoustical study. While Scott was primarily interested in capturing the nuances of human speech, Koenig saw the device as a tool for scientific analysis. Scott did believe the calligraphy inscribed in the soot - ***"the words that wrote themselves"*** - embodied a form of "*natural stenography*" that would someday be read as easily as a stenographer deciphered his own jottings.

Their partnership led to the production of several improved models, though diverging goals eventually caused a rift between them. Scott's phonautograms lay silent and forgotten in the venerable French institution for 150 years. In addition, an album of phonautograms presented to Professor Henri Victor Regnault has been in the possession of the Institute of France since its accession of Regnault's papers upon his death in 1878 [1].

IV THE REVIVAL OF SCOTT'S RECORDINGS

Though Scott didn't intend for his phonautograph recordings to be reproduced, he must have understood the possibility. In 2008 sound researchers (David Giovannoni, Patrick Feaster, Meagan Hennessey and Richard Martin) reproduced these inscriptions by optically scanning the sheets and digitally reconstructing the waveforms held within.

The very first draft of the phonautograph was recorded onto glass coated in soot (as opposed to paper coated in soot like Scott would later use), and was only capable of capturing a second or two of sound. This short recording of a guitar is, however, not one of the listenable ones. It was recorded in 1853/54 and features a friend of Scott's, Adolphe Giacomelli, strumming a guitar. This is the oldest recording of any instrument, but the sound content is too poor to be reproduced intelligibly, and the guitar is not recognizable.

On March 28th, 2008, the world's oldest playable - digitally reconstructed - voice recordings were presented publicly for the first time at Stanford University, during the annual convention of the Association of Recorded Sound Collections [1].

Signal processing of Scott's phonautograms [1]

(Fig. 4]

(1) The processing started with capturing the original soot-inscribed paper scrolls in extremely high optical resolution. These detailed scans preserved every subtle ink trace and blemish, creating digital masters suitable for further processing.

(2) Then, two complementary techniques were used in combination, because a phonautogram is not simply an oscillogram : It's a mechanical trace made by a boar's bristle responding to sound vibrations, drawn on soot-covered paper. A boar bristle is more flexible than a stylus: it can vibrate laterally and vertically, producing variations in thickness in the trace. The thickness of the soot line does also carry information, not because Scott intended it, but because of how the stylus interacted with the surface. The two complementary approaches are:

2.1. : The Virtual Stylus (VS) Method **(**Developed by Lawrence Berkeley National Laboratory) treats the phonautogram like a record groove. A digital "stylus" follows the center of the wavy line as if it were a needle on a vinyl record, and the trajectory is vectorisated.

2.2. The Variable Width (VW) Method (devised by Patrick Feaster) converts the width (thickness) of the soot line into a light passage, similar to how optical soundtracks work in film. The visual width of the soot line is converted into a varying-width light passage, and the light modulation is decoded and used to reconstruct an audio signal. This graphic conversion bypasses many groove-tracking pitfalls.

(3) Time calibration

Several phonautograms include tuning fork traces (e.g., 250 Hz) to help correct for speed fluctuations caused by hand-cranked recording. Scott de Martinville's most precise phonautograms include a simultaneous tuning-fork trace vibrating at a known frequency. Feaster used this fork trace as a built-in pilot tone, measuring its period to establish exact playback speed. He then manually adjusted the time axis of recordings like "Au Clair de la Lune" to correct speed miscalculations.

(4) Sound reconstruction

Once the waveform had been reconstructed, it was converted into an audio signal. The software can apply adaptive filters to reduce noise while preserving the harmonic details and correct distortions like smear and drop out [1] .

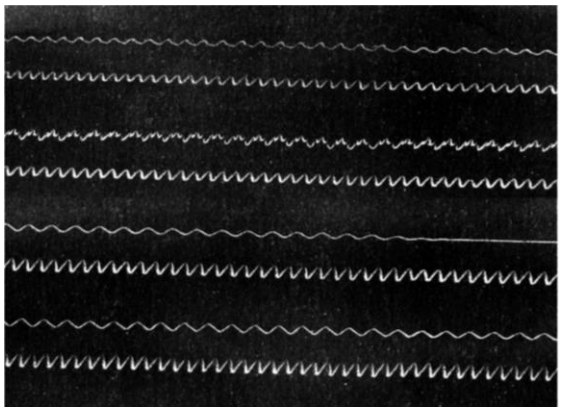


Fig.4 Contrast-enhanced detail of paired tracings from Scott 1860 phonautogram No. 5. The bottom trace of each pair was made by a tuning fork; the top trace is the recording of airborne sounds: a voice singing “Au Clair de la Lune.”

Scott’s Phonautogram No. 5 (Fig. 4) is the most interesting ; it’s also the earliest dated sheet. After processing of the signal, it is possible to clearly hear “Au Clair de la Lune” as recorded in Paris on 9 April 1860. The voice is probably that of Scott de Martinville himself.

V. THE NEXT STEP

On November 29th 1877 Thomas Edison designed the "tinfoil phonograph" and entrusted John Kruesi, one of his top laboratory mechanics, with the task to build a prototype. It is not known whether Edison was aware of Scott's experiments, but it is entirely possible that he learned about the phonautograph through Koenig's catalogues. Edison’s machine consisted of a cylindrical drum wrapped in tinfoil (paraffin paper in the first experiments) and mounted on a threaded axle. A small horn with a diaphragm at its end was connected to a stylus that engraved a groove related to the vibrational pattern of the diaphragm. For playback the mouthpiece was replaced with a "reproducer" that used a more sensitive diaphragm. Edison recited "Mary had a little lamb" into the mouthpiece for the first demonstration.

VI. CONCLUSION

Scott de Martinville realized in 1857 the first vocalizations captured from the air by a machine, and inscribed onto a permanent medium. At the time, it was not possible to reproduce the sounds, but this device played a crucial role in developments that followed two decades later. Modern technology has made these original phonautograms audible.

VII. REFERENCES

[1] https://firstsounds.org (David Giovannoni, Patrick Feaster, Meagan Hennessey and Richard Martin)

[2] L. da Vinci, *The Madrid Codices I & II*. Translated and edited by Ladislao Reti. New York: McGraw-Hill, 1974 & L. da Vinci, L, *Codex Atlanticus* (Facsimile ed.). Giunti-Barbera & Johnson Reprint Corporation 1979.

[3] G. Galilei, *Discorsi e dimostrazioni matematiche* (1638) Leiden Elzevier 1638. New York The MacMillan Company 1914 p. 101.

[4] E. F. F. Chladni, *Entdeckungen über die Theorie des Klanges.* Weidmanns Erben und Reich Leipzig (1787)

[5] E. L. Scott de Martinville, *Histoire de la sténographie*. Charles Tondeur Paris (1849).

[6] S. Benoit, D. Blouin, J.Y. Dupont, G. Emptoz, *Chronique d’une invention : Le phonautographe d’Édouard-Léon Scott de Martinville (1817-1879) et les cercles parisiens de la science et de la technique*. Documents pour l’histoire des techniques : 17 (1er Semestre 2009) 69-89.

https://doi.org/10.4000/dht.502

[7] E. L. Scott de Martinville, « Inscription automatique des sons de l’air au moyen d’une oreille artificielle » note aux *Comptes Rendus de l’Académie des sciences*, 1861, tome LIII, pp. 108-111.

[8] L. Figuier, « *Essai d’une fixation graphique des sons, par M. Léon Scott* », L’Année scientifique et industrielle, Paris, Hachette, 3e année, 1858, tome I, pp. 62-69 (citation pp. 65-66).

[9] F. Moigno, « Phonautographe et fixation graphique de la voix, par M. Édouard-Léon Scott », *Cosmos*, tome 14, premier semestre 1859, pp. 314-320; id., « Phonotography ; or, the graphic fixing of the voice », *Photographic news*, London, 15 avril 1859, pp. 62-64.

ENRICO CARUSO – A VOCAL PROFILE BASED ON HISTORICAL RECORDINGS

B. Richter[1]

[1] Freiburg Institute for Musicians' Medicine, University of Music Freiburg, Medical Centre and Faculty of Medicine – University of Freiburg, Freiburg Centre for Music Research and Teaching, Freiburg, Germany
bernhard.richter@uniklinik-freiburg.de

Abstract: **Enrico Caruso made almost 500 recordings between 1902 and 1920. He recorded many of the arias and songs from his repertoire several times. This enables us to compare recordings from different periods of his career and to create a vocal profile that covers almost his entire active artistic career. From a voice doctors' perspective, this is interesting in two respects. On the one hand, from the perspective of *vocal physiology*, we can understand Caruso's technique, which was considered novel by contemporary critics. On the other hand, from a phoniatric point of view, Caruso's voice can be assessed to determine whether his numerous health problems, including two operations on his vocal cords, are audible as vocal damage, as contemporary critics who heard him live described such acoustically perceptible limitations. Since Caruso's original recordings have several technical limitations, the analysis was carried out by an expert rating with a special focus on vocal technique and the parameters of hoarseness and dysodia.**

Keywords: Enrico Caruso, Vocal profile, Historical recordings, Voice physiology perspective

I. INTRODUCTION

Enrico Caruso (1873–1921) was already a celebrated singer in his late twenties and retained this status until his early death at the age of 48. Caruso gained his fame not only on stage, but also through the gramophone. The gramophone was the only medium for distributing music available during Caruso's lifetime, although Caruso did take part in an early experiment in radio history in 1910 [1]. As the sound quality on the receiving devices was unsatisfactory, the experiment was not continued. The first public radio broadcast in Europe was not made until December 24, 1921, from the Eiffel Tower – around four months after the singer's death. Leo Blech (1871–1958), general music director at the Berlin Court Opera since 1906, who often worked with Caruso as a conductor from 1907 onwards, exclaimed after Caruso's death: "How fortunate that we at least have gramophone records of him." [2].

Caruso made his first recordings for Emile Berliner's company at the age of 29 on April 11, 1902, in Milan. The possibility of making sound recordings literally "astonished" the first listeners. They did not know whether the reproduction of the human voice was a magic trick – a bluff – or reality. Thomas Mann gives us a very detailed description of how the sound recordings may have affected listeners at the time in his novel „Der Zauberberg" (The Magic Mountain), published in 1924 [3]. In a separate chapter entitled "Fülle des Wohllauts" (The Fullness of Melody), he conjures up the effect of a gramophone on the inhabitants of the Berghof, namely Hans Castorp, the sad "hero" of the novel, in the minds and ears of his readers. In doing so, he also creates an unforgettable literary monument to Caruso—whom he does not name, but whom he quite clearly characterizes as "the world-famous tenor voice that was so often featured in the albums."

The fact that Caruso made almost 500 recordings between 1902 and 1920 makes it possible to create a vocal profile of Caruso from the perspective of his artistic career. From a medical perspective, this is helpful in two ways. On the one hand, from a *voice physiology* perspective, it allows us to understand Caruso's technique, which was perceived as new by contemporary critics [4], by listening to it. On the other hand, from a *phoniatric* perspective, Caruso's voice can be assessed to determine whether the numerous health problems he had, especially in the second half of his career, are audible as vocal damage.

II. METHODS

Six pieces of music will be examined as examples in the lecture:

1. Ballata of the Duke of Mantua, *Questa o quella*, from the first act of Giuseppe Verdi's *Rigoletto.* Three recordings: one from 1902, one from 1904, the other from 1908.
2. Aria of the painter Mario Cavaradossi, *E lucevan le stelle*, from the third act of Giacomo Puccini's *Tosca.* Two recordings: one from 1904, the other from 1909.

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

3. Italian art songs: *Ideale* and *Luna d'estate* by the same composer, Paolo Tosti and *I' m'arricordo 'e Napule* by Giuseppe Gioè. Three recordings: *Ideale* from 1906, *Luna d'estate* from 1916, *I' m'arricordo 'e Napule* from 1920.
4. Aria of Éléazar *Rachel, quand du Seigneur* from the fourth act of Fromental Halévy's *La Juive.* Single recording from 1920.

The analysis was carried out by an expert rating of the author, since Caruso's original recordings have several technical limitations which hinder acoustic analysis using measurement techniques. The focus was on assessing vocal technique and whether the voice sounded hoarse or appeared to be limited in its performance in terms of dysodia.

III. RESULTS

Voice physiology perspective

Caruso recorded the Ballata of the Duke of Mantua, *Questa o quella*, from the first act of Giuseppe Verdi's *Rigoletto* several times. In one of the early recording sessions, which took place on April 11, 1902, he was accompanied on the piano by Salvatore Cottone. We hear a lyrical, youthful, rather bright-colored voice that can be carried seamlessly in full voice function up to B♭ 4 in the final cadence. The vocal technique is already flawless, but on the first A♭4 in the first verse, on the vowel /o/ in the word "cedo," we hear a very open vowel coloration that tends almost toward an open /a/. As a secondary finding in terms of *phoniatrics*, Caruso can be heard clearing his throat distinctly in the interlude to the second verse. The reason for this can already be guessed at the end of the first verse, as the typical sound of mucus in the vocal folds can be heard on the G4 on the vowel /o/ in the word 'Forse'. In 1902, this "clearing of the throat" did not seem to be a reason to stop the recording, nor did the pianist's minor " misplays" in the prelude, interlude, and postlude.

In a recording from February 1, 1904, again with piano accompaniment, the vocal color in the upper register is already more balanced and improved, and the vocal technique also seems to have stabilized and to be functioning even better. No physiological or medical voice problems can be heard.

In a recording made on March 16, 1908, this time with orchestra, the voice has matured, but is not significantly darker, the vocal color is completely balanced, and the technique is mastered effortlessly. In the final cadence, the vocal 'attack' on the high notes up to B♭4 is even more powerful than in the earlier recordings, and the high notes are held more effortlessly and for longer than in the two previous recordings. There are no vocal flaws to be heard.

Caruso also recorded the aria of the painter Mario Cavaradossi, *E lucevan le stelle*, from the third act of Giacomo Puccini's *Tosca* frequently since 1902. For comparison, two versions from 1904 and 1909 are used here.

In the earlier recording, made on February 1, 1904, he was accompanied on the piano. Here, as in the second *Rigoletto* recording made on the same day, we hear a seamless vocal function up to A4. At A4, he achieves a technically flawless diminuendo. The open vocal coloration in the upper register is no longer audible. Veristic sobs are incorporated at the beginning of the word "speranza" and at the very end before "la vita" in the second stanza.

In the second recording used for comparison, made on November 6, 1909, this time with orchestral accompaniment, the voice has matured further, but is not significantly darker; the vocal color is completely balanced, and the technique is mastered effortlessly up to the high B4. In the first verse, the diminuendo to B4 is not performed as clearly. In the second verse, the vibrato seems to be slightly faster towards the end than in the first verse. At the end, veristic sobs are again incorporated. No physiological or medical voice problems can be heard.

Caruso usually recorded Italian art songs only once, so only two different compositions from this genre can be used for comparison. They were recorded over a period of almost a decade, but are by the same composer, Paolo Tosti, and have a similar expressive character and comparable vocal requirements in terms of range and tessitura. Both have an A♭4 as their highest note. These are the songs *Ideale* and *Luna d'estate.* Both recordings were made with an orchestra. *Ideale* was recorded on December 30, 1906, and *Luna d'estate* on February 5, 1916.

In both recordings, the voice is effortlessly formed throughout the entire range by a full voice function. The vocal coloration is very balanced. The high notes are reached effortlessly. No physiological or medical voice problems can be heard in either recording.

The same applies to the Neapolitan folk songs recorded on September 14, 1920, such as Giuseppe Gioè's *I' m'arricordo 'e Napule*. Here, the highest note is a G4. The voice sounds fresh and effortless, and in keeping with the character of the piece, it is light and flexible,

sounding not like grand opera but like a simple folk song.

Caruso recorded Éléazar's aria *Rachel, quand du Seigneur* from the fourth act of Fromental Halévy's *La Juive* only once, on September 14, 1920, in the same recording session as the Neapolitan folk song *I' m'arricordo 'e Napule* already discussed, with orchestral accompaniment. The sound of his voice is fully mature, somewhat darker than in the previous opera recordings – but not artificially darkened – the vocal coloration is perfectly balanced in the French text, and the technique is mastered effortlessly up to the high B4. The high notes are sung in a more dramatic vocal style and with significantly more power in the final passage, without however being forced. In the ascents to the high notes, the voice is placed very far forward via brighter vowels in the passaggio. In the second verse, in which Éléazar quotes his daughter's words, the voice is much brighter and more lyrical at the beginning of the phrase than at the beginning. Slightly pronounced veristic broken word endings (bonheur) and veristic implied sobs are occasionally incorporated at the beginning of individual words (grâce). No physiological or medical voice problems can be heard.

Phoniatric perspective

Over the course of the versions of *Questa o quella*, an increasingly veristic vocalization can be heard in some places, with note endings torn off by excess pressure and slight vocal sobs and sighs. Caruso also used similar effects in *E lucevan le stelle*, and particularly pronounced ones in *Rachel, quand du Seigneur*. Jens Malte Fischer even says that Caruso "gasped" between phrases in this aria [5]. These noises are not heard in the recordings of songs and folk songs made at the same time. Caruso was therefore able to control these "effects" vocally; they were not an expression of inability.

According to Christian Springer, Caruso underwent two operations on his vocal cords, in 1907 and 1909, performed by Professor Temistocle della Vedova from Milan. Neither in the recordings made before the operations, on December 30, 1906 (*Ideale)*, nor after the first operation, on March 16, 1908 (*Questa o quella)*, is there any evidence of hoarseness or dysodia. Nor is there any indication of this in the recording from November 6, 1909 (*E lucevan le stele)*, after the second operation.

Even the later recordings after 1916 show no audible voice problems. The slightly increased vibrato of the 1909 recording is no longer present in the later recordings. Generalized vocal judgments—one could almost speak of condemnations—such as those made by singing teacher George Conelli, who reported on his impressions of Caruso's performances in Paris in 1909 and Milan in 1916, cannot be verified in the recordings. He wrote: "I was deeply disappointed. His voice seemed heavy, guttural, and considerably strained to me. In the third act of Aida and the last act of Manon Lescaut, it sounded rough, and in Pagliacci, only his overwhelming talent and his grasp of the character of the role saved him. Later, in Rigoletto, I heard his voice break on the high B because he forced it in the cadenza to "La donna è mobile." Still later, in Milan in 1916, his voice was so damaged that in Pagliacci it was practically indistinguishable from that of the baritone." [6]. It is, of course, possible that Caruso was vocally indisposed in each of the performances in question, but from a voice doctors' perspective and based on the available recordings, no systematic decline in Caruso's voice or vocal technique can be diagnosed over the span of his career.

IV. Discussion

The phenomenon that Caruso used "different" voices depending on the repertoire had already been noticed by critics of his time. One critic wrote: "The singer always claimed that he had his 'different voices' in special drawers. One contained his Aida voice, another the one he needed for Martha, a third the precious instrument with which he sang in Bohème, and so on throughout his entire repertoire." [7].

Since no vocal limitations can be heard in the recordings before and after the phonosurgical operations, it is understandable and logical that Caruso sued his doctor for breach of confidentiality in his dispute with Professor della Vedova, as the latter informed the press during the "crisis of 1911" – without examining Caruso – that he probably had "vocal cord nodules" again and would have to be operated on by him, even though this crisis was probably more psychological than physical in nature.

V. Conclusion

In summary, in the vocal-medical-auditory assessment, we hear a technically very well-controlled and healthy voice in the observation period from the first to the last recordings, with an age-appropriate vocal maturation process from initially rather lyrical to later somewhat more dramatic – without, however, losing the ability to perform some passages in operas or songs lyrically. There are no audible signs of systematic fatigue or wear and tear in his voice.

REFERENCES

[1] P. Fryer, *The opera singer and the silent film* (note 159), Jefferson. McFarland & Co Inc, 2005, pp. 181-183.

[2] L. Blech, „Der Gesangskünstler Enrico Caruso“, in: Emil Ledner – Erinnerungen an Caruso“. Hannover: Paul Steegemann Verlag 1922, p. 3.
[3] T. Mann, „Der Zauberberg“. Berlin: S. Fischer Verlag, 1924.
[4] B. Gentili, „The birth of ‘modern’ vocalism: The paradigmatic case of Enrico Caruso“, *Journal of the Royal Musical Association,* vol. 146/2, pp. 425-453, 2021. doi:10.1017/rma.2021.11
[5] Quoted from C. Springer, *Caruso – Tenor der Moderne*. Wien: Holzhausen Verlag, 2002, p. 299.
[6] J. M. Fischer, „Grosse Stimmen: von Enrico Caruso bis Jessye Norman“, Stuttgart: Metzler, 1993, p. 14.
[7] Quoted from: J. M. Fischer, „Grosse Stimmen: von Enrico Caruso bis Jessye Norman“, Stuttgart: Metzler, 1993, p. 1.

A SURVEY ON VIBRATO PARAMETERS IN HISTORIC OPERA SINGERS

I. Ferrante[1]

[1] Dipartimento di Fisica dell'Università di Pisa e INFN sezione di Pisa, Italy
isidoro.ferrante@unipi.it

Abstract: **In this work I report on the findings of a survey of vibrato parameters in three different large datasets of old recordings of classical western music. Each dataset is designed with different criteria and different purposes. The trend of decreasing rate and increasing rate with recording time is confirmed also in new data. Moreover, a small but significant difference of -0.3±0.07 Hz in vibrato rate between male and female singers is found, male vibrating slower, and an overall decrease of vibrato rate correlated with age in five famous tenors.**

Keywords: **Vibrato, Opera, Modulation, Singers**

I. INTRODUCTION

Analysis of vibrato in old studio recordings has been widely used in the past since the dawn of voice studies. Examples are the pioneering study by Seashore, Prame, Titze, Sundberg [1] [2] [3] [4] [5] [6] but also many modern ones, for example [7] [8]. The use of commercial, or live recordings avoids the artificial setting of a university laboratory and allows to study the performance practice of the historical period or cultural environment in which the recordings have been made. Moreover, due to the availability of large music collections, large datasets can easily be collected. On the downside, there is no access on the singer, so no direct measurement is possible, by inspection or by EMG, and the information about sound production mechanism is not easily accessible. The variables used in those studies are usually vibrato rate extent (as in this paper) and jitter and shimmer (like in [9]), even if nonlinear variable have been tried with interesting result [10]. In those two last papers the difference between vibrato in opera and jazz have also been studied. The validity of the method has been checked by Glasner and Johnson [11] who studied the effects of recordings system on vibrato parameters: while vibrato rate is not affected, as long as playing speed is correct, surprisingly the added surface noise in cylinder recordings increases vibrato extensions, albeit of a very small amount.

In a previous study of the same author [7] vibrato rate and extent have been studied in a set of 105 sopranos singing "Vissi d'arte " from Tosca.

Main results were the decrease of the measured vibrato rate in dependence of the year of recording, together with an increase of vibrato extent. Rate and extend showed also a strong anticorrelation. In this study those findings have been extended to different dataset, investigating also the dependence on sex, register and singer's age

II. METHODS

A. Datasets

Three different datasets have been used:

1. Soprano dataset

This dataset extends to the present date the one analyzed in previous work. It consists in 76 tones from Tosca's aria *Vissi d'arte*, precisely the highest note on the word "signor". The tones have been extracted from YouTube recordings (and thus in live performances) from 2000 to 2024. No selection on performers has been applied, but only on recording quality.

2. Record of Singing (RS)

It consists of a single tone from each track of the well know CD collection "The record of Singing" , Voll 1 and 2. The first volume covers the 78 rpm era, while the second one goes from the dawn of LP's to year 2000. After cleaning for those few tracks for which pitch reconstruction gave bad results, the sample consists of 191 male singers and 218 female ones, each of them represented by a single tone. Male singers were tenors (95 tracks) and baritones or basses (84 tracks), plus 10 falsettists and a single castrato (Moreschi). Female singers were 179 sopranos and 39 altos and mezzos.

3. Tenor dataset

In this set, recordings of five famous tenors, whose activity is well documented, four of which with a remarkable long recording career have been chosen. The tenors are:

- Enrico Caruso – one tone for each tracks from the 12 CD RCA set "The complete Caruso"– 231 tones total.
- Beniamino Gigli – A set of 84 random excerpts from an EMI 7 CD coffret, containing

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

operatic studio recording uniformly distributed in time from1928 to 1955

• Giacomo Lauri Volpi – 38 random excerpts from a 5 CD TIMA club coffret uniformly distributed in time from 1922 to 1957. Recordings made at 80 have been excluded.

• Jussy Björling - 94 random excerpt from a 4 CD EMI set containing studio recordings from 1930 to 1959.

• Placido Domingo – single tones from live and studio recordings – 69 excerpts from 1962 to last exhibitions in tenor repertoire in 2014.

B. *Analysis method*

Tones in each recording have been selected individually after spectrogram inspection, choosing the ones in which there was at least one harmonic clear and free from orchestral sound. Pitch reconstruction has not changed since previous paper [7], but vibrato parameters have been calculated by use of the Hilbert transform of the pitch profile: the phase of the Hilbert transform is the instantaneous rate, while the modulus is equal to the instantaneous amplitude. The phase and the amplitude are then mediated, and the amplitude is converted in cents giving thus the vibrato extension.

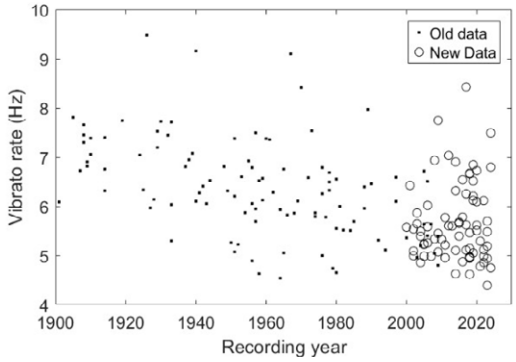


Fig. 1 Vibrato rate vs. recording year. Points are the data already published, while circle are the new ones. The trend is confirmed

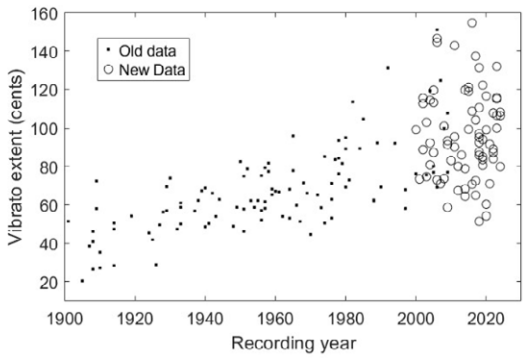


Fig. 2 Vibrato extent as function of recording year. Again, the trend is confirmed. New data, however, show a greater spread around the mean.

III. RESULTS

A. *Soprano dataset*

In Fig. 1 you can see vibrato rate and extent as function of recording year, superimposed with the results from previous study.

One can see that the general trend of an increase in extent and a decrease in rate is confirmed, even if a few performers show a higher rate even in recent performances. One can also observe that the new method of calculating rate and extent through Hilbert transform is consistent with the old method.

The minimum rate of 4.4 Hz has been observed in a 2024 recording by Sonja Yoncheva, while the higher extent (155 c) has been found in a 2009 recording by Karita Mattila.

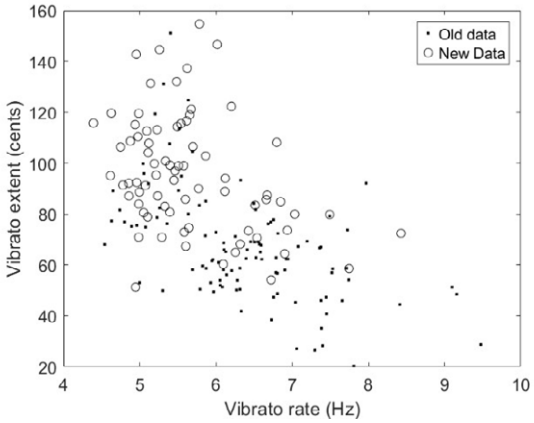


Fig. 3 Vibrato rate and extent are anticorrelated. New data can be superimposed to old ones.

Also, the anticorrelation between rate and extent observed in the old sample is confirmed, and extended to larger extent values.

B. *Record of Singing sample*

The heterogeneity of this sample allows to investigate differences in singer sex or register or singing style. On the other hand, mixing very different singers can bring confusing results.

Maximum vibrato rate has been measured for females in Eleen Beech Yaw (9.83 Hz) and for males in Sir Charles Santley (9.16 Hz), while minimum values have been found in Rudolph Schock (4.8 Hz) and Sir Geraint Evans (4.87 Hz)

Maximum extent has been found in Diana Damrau (155 c) and Tatiana Troyanos (138 c), while minimum values have been measured in Montserrat Figueras (12 c) and Friedrich Schorr (13 c).

Aggregated results are shown in Table 1.

Table 1Mean values for the full sample, for male and female subsamples, and for tenors and low male voice.

	Rate (hz)	Extent (c)
Full Sample	**6.30±0.04**	**64.8±1.3**
Males	6.1±0.05	60.6±1.5
Females	6.5±0.05	68.1±2.0
M-F	*-0.33±0.07*	*-7.5±2.5*
Tenors	6.1±0.07	57.5±1.7
Baritones and basses	6.2±0.09	65.4±2.7
T-BB	*-0.09±0.11*	*-8.0±3.2*

Females seem to have a slightly faster vibrato with higher extent: the difference is significant for rate ((p=0.0002) and much less for extent (p = 0.054) . An inspection of the distributions finds a clear shift in rate, while extent histogram seems to show a bimodal distribution for female vibrato extent of unknown origin. (see Fig. 4)

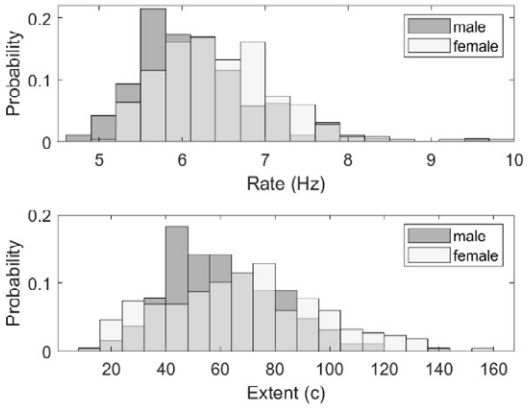


Fig. 4 Vibrato rate and extent distribution in RS dataset. The distributions look different, although this difference is not much significant.

A small but significant (P=0.012) difference in extent can be found between tenors and baritones and basses. The number of mezzos and contraltos was too low to allow comparison with sopranos. With the help of RS sample we can see that the historic trend is confirmed also in this more heterogeneous data collection, as can be seen in Fig. 5: but one can also observe, starting from the 60' onward, the onset of period baroque singing, with small or no vibrato at all.
Also the anticorrelation between rate and extent is confirmed.

A. Tenor dataset

This data have been used to verify the effect of aging in vibrato parameters. For example in figure Fig. 7 one can see that vibrato rate shows a slow constant decline through che career of almost all singers in the sample, with the exception of Bjorling, who remains almost stable, and of Lauri Volpi, whos dramatic decrease of vibrato rate is worth investigating in more detail.

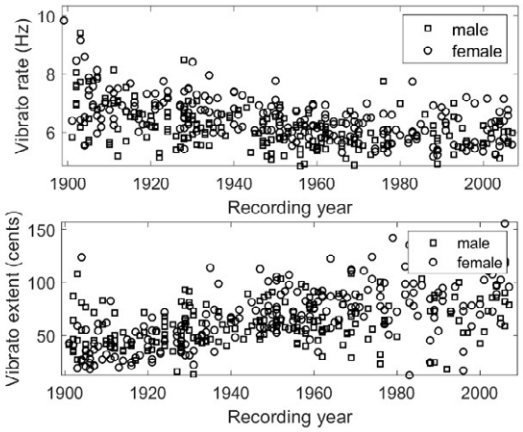


Fig. 5 Rate and extent as function of recording year in RS dataset. Once again, one observe the trends noted in the soprano dataset.

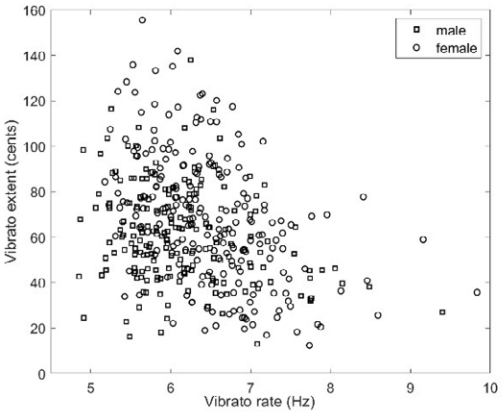


Fig. 6 Correlation between extension and rate in RS dataset.

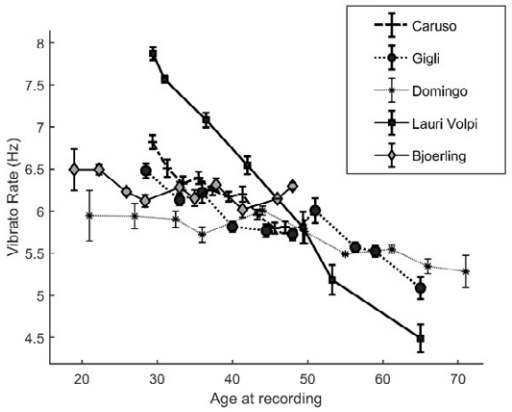


Fig. 7 Vibrato rate vs. age for the five tenors in the sample. A slow decrease can be seen, but the rate decrease for Lauri Volpi is striking.

On the other hand, vibrato extent, as can be seen in Fig. 8 does not change very much throughout singer's career. In other studies [12] no effect on vibrato rate due to age has been observed, but data were limited in number and in time span.

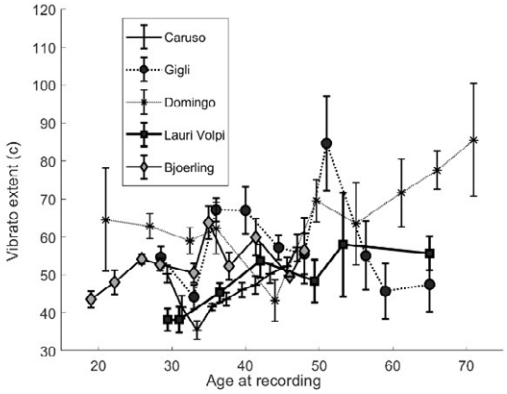


Fig. 8 Vibrato extent vs. age for the five tenors in the sample. Extent seems to be constant during the singer career

IV. Discussion

The three datasets used in this study, even if constructed with different criteria, show consistent results. First of all, the historic trends observed in the past, and well known by collectors of old recordings, can now be measured and quantified. Also the intuitive anticorrelation between vibrato and extent (which can be expressed as: something which oscillates slowly can go further) is demonstrated. Another well-known fact, i.e. that older singers have a slower vibrato has got experimental evidence. Lacks, anyway, from those data, any indication of vibrato production mechanism. There is also the need of a better statistical analysis, which could show hidden correlations between vibrato parameters and mean F0 or emission volume. Also jitter and shimmer, which have been left unobserved, could give other useful indication.

V. Conclusion

Analysis of vibrato in commercial, live and historic recordings has a large potential due to the amount of data that we have at our fingerprints. Unfortunately part of this potential is spoiled because tone selection in a track is made by hand, and this can be painfully slow and can bring to unknown biases. However, a careful choice of the criteria used in selecting dataset can help to discover differences due to sex, register, style, age or fatigue. New variables could also be chosen, and other effects can be investigated, like the amplitude modulation and the explanations given by Sundberg [13]. The development of new methods of data selection and of statistical analysis are thus welcome.

REFERENCES

[1] C. E. Seashore, Psychology of music, New York: McGraw Hill, 1938.

[2] E. Prame, «"Measurement of the vibrato rate of ten singers",» *J. Acoust. Soc. Am 96,* p. 1979–1984, 1994.

[3] E. Prame, «Vibrato extent and intonation in professional western lyric singing,» *J. Acoustic Soc. Am,* vol. 102, pp. 616-621, 1997.

[4] A. Timberlake, I. Titze and A. Keidar, "Vibrato characteristics of tenors singing high C's," in *Transcripts of the 13th symposium Care of the Professional Voice, Part 1*, 1984.

[5] J. Bretos and J. Sundberg, "Vibrato extent and intonation in professional western lyric singing.," *Journal of voice,* vol. 17, pp. 343-352, 2003.

[6] J. Bretos and J. Sundberg, "Measurements of vibrato parameters in long sustained crescendo notes as sung by ten sopranos," *Journal of voice,* vol. 17, no. 3, pp. 343-352, 2003.

[7] I. Ferrante, "Vibrato rate and extent in soprano voice: A survey on one century of singing.," *Journal of the Acoustical Society of America,* vol. 130, no. 2, pp. 1683-1688, 2011.

[8] T. Nestorova, E. Brander, B. Gingras and C. Herbst, "Vocal Vibrato Characteristics in Historical and Contemporary Opera, Operetta, and Schlager," *Journal of voice,* vol. 39, no. 4, pp. 1133.e21-1133.e34, 2025.

[9] C. Manfredi, D. Barbagallo, G. Baracca, S. Orlandi, A. Bandini and P. Dejonckere, "Automatic Assessment of Acoustic Parameters of the Singing Voice: Application to Professional Western Operatic and Jazz Singers," *Journal of voice,* vol. 29, no. 4, pp. 517.e1-9, 2015.

[10] G. A. Martinez and H. Daffern, "Complexity of Vocal Vibrato in Opera and Jazz Recordings: Insights From Entropy and Recurrence Analyses," *Journal of voice,* 2023.

[11] J. D. Glasner and A. M. Johnson, "Effects of Historical Recording Technology on Vibrato in Modern-Day Opera Singers," *Journal of Voice,* vol. 36, no. 4, pp. 464-478, 2020.

[12] A. Sobolewska, P. Claros, C. Pujol, A. Claros-Pujol and A. Claros, "Ageing of professional opera singer's voice – preliminary findings," *Otolaryngol Pol.,* vol. 73, no. 4, pp. 29-34, 2019.

[13] J. Sundberg, The science of the singing voice, Northern Illinois University Press, 1987.

SPECIAL SESSION III
UPDATES ON VOICE RANGE PROFILE MEASUREMENTS

Organized by M. Kob and G. Baracca

DETAILED ANALYSIS OF VOICE RANGE PROFILES

Malte Kob[1], Giovanna Baracca[2]

[1] Erich Thienhaus Institute, Detmold University of Music, Germany
[2] University of Padua, Italy
malte.kob@hfm-detmold.de, giovanna.baracca@gmail.com

Abstract: **Voice performance can be characterized with respect to dynamics and tonal range by measurement of a voice range profile (VRP). Usually two series, a soft and a loud scale of the sustained phoneme /a:/ from the lowest to the highest possible pitch are recorded and plotted in a diagram from which the extreme values are extracted (lowest, highest, softest, loudest sound). We propose a more detailed analysis of the voice ranges by extraction of seven markers from the diagram, allowing an assessment of the development of the edges of the voice range during education or therapy.**
Keywords: **Voice Range Profile, Singing Voice, Voice development**

I. INTRODUCTION

Voice range profiles have been extensively used to characterise the current status of the speaking voice or of the singing voice. The VRP is a standardized representation [1] of vocal limits in singers and have also been used for clinical applications [2]. It involves singing the lowest and highest volumes in a slow, continuous sequence of tones, usually on the vowel /a:/. The profiles recorded in this process show the control over the voice per pitch and can therefore not only be used to track the development of voice ranges during the education but also be used to display different voice registers or even the passaggio. Various extensions have been proposed that allow the visual representation of additional voice features within the VRP, e.g. Voice Maps [3]. We introduce a method to have a more detailed view on the initial parameters, the fundamental frequency and the voice level, at the edges of the VRPs.

II. METHODS

A. Voice range profile measurement

The Lingwaves system (Wevosys) was used for the VRP measurement in all cases. This system consists of a level-calibrated measurement microphone positioned 30 cm in front of the mouth and converts the audio signal into digital wav format at a sampling rate of 22,050 Hz.
The VRP is evaluated using the Lingwaves software. Custom Octave code was developed to provide the additional information of the detailed VRP (DVRP).

B. DVRP Analysis

The vph file that is saved after completion of the VRP measurement is processed using a script that is run with Octave. First, the raw data is imported and parsed for the various pre-calculated SPL and pitch ranges as well as for the matrix of SPL values. From the matrix the piano and forte curves are extracted using the *min* and *max* functions.
For the detailed analysis the VRP data for the piano and the forte curves is segmented into an uneven number of ranges, i.e. 3, 5, 7, or 9, that span from the lowest recorded pitch to the highest of the piano and forte curves.
Since the measured SPL values do not correspond to all possible frequencies in the matrix, the data are cleaned in the piano and the forte curve from values that deviate beyond a threshold value from the average SPL in in each segment. The average values, together with the data of the segments, are exported in an excel sheet for further evaluation.

C. Studies

The DVRP method was evaluated in two studies given in Table 1.

Tab. 1: Overview of studies

A	Voice transition female to male
B	Longitudinal study of voice pedagogues

We recorded all VRP data of study A in a sound-proof cabin (Studiobricks One Plus VO) which is usually used for dubbing or voice recordings. In study B, the first two measurements took place in classrooms with medium reverberation time, the last in the cabin as in study A.
The subjects had passed a logopedic screening and showed no symptoms of voice or language disorders.
All subjects were anonymized using a three-digit ID, which were used for intra-individual investigations. Here, we describe two case reports from these studies.

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

III. RESULTS

The detailed analysis of the voice range profiles accessed in study A during the transition of a trans person (female→male, start at age 25) is given in the graphs of Fig. 1.

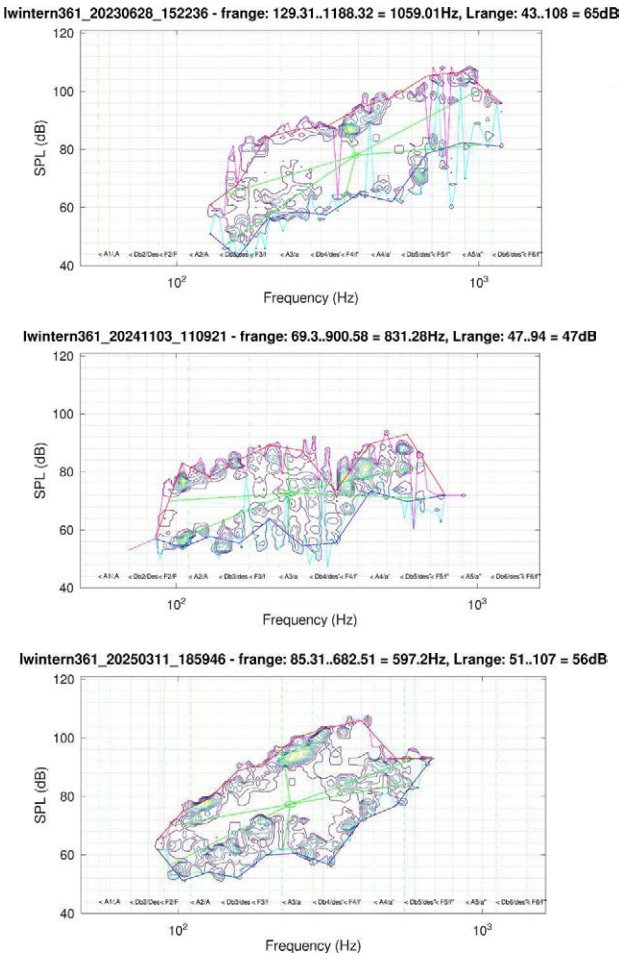


Fig. 1: DVRPs of a trans person at the beginning (top), after 267 days (centre), and after 622 days (bottom) of transition

In each plot the raw VRP is shown as a contour plot, the piano and forte curves in light blue (p) and magenta (f) before cleaning, and the final curves are shown in dark blue and red. For better readability, musical notes were added to the horizonal axis and vertical green lines at the positions of the notes "A" and "Db" for each octave.
The six segments of the DVRP allow a detailed analysis of the edges of the VRPs, the end of the green lines from the green circle in the centre of gravity indicate the location of the averages of each of these segments.
Another example is given in Fig. 2. Here the development of a female teacher is documented with respect to her voice status before (age 21, top), after four years of education (centre) as well as another 10 years later in her job as teacher (bottom).

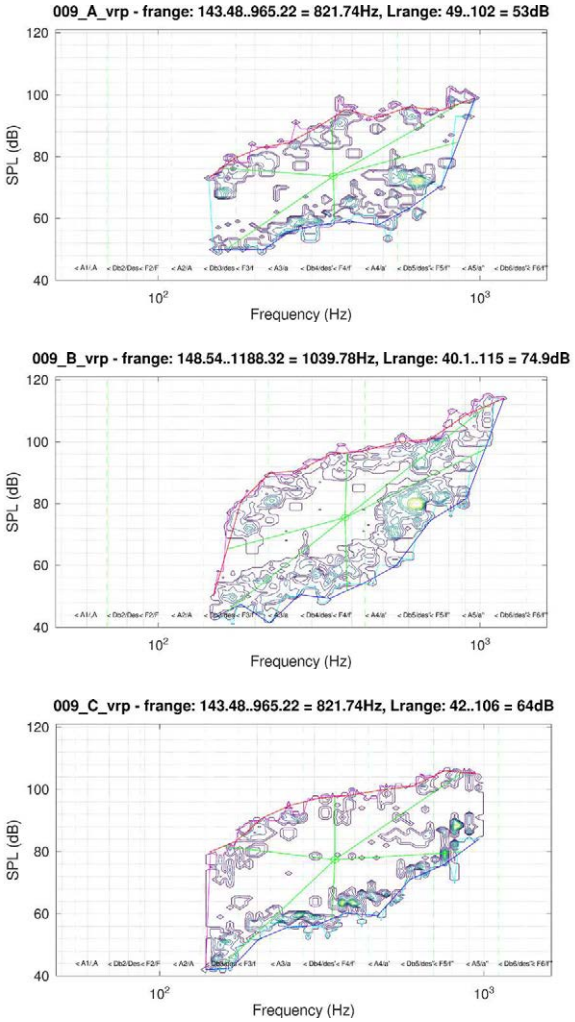


Fig. 2: DVRP of a female teacher before (top), after four years (centre) and 10 years after her education

The voice development during the education is clearly seen between in the overall extension of the VRP. It can also be observed that the VRP seems to be reduced after the education. However, precise details in the development can be better traced by observation of the edges.
In Figs. 3&4, for both studies the six standard voice range parameters that are usually used in VRP programmes, are displayed first: overall minimum and maximum fundamental frequency (f_{min}/f_{max}) and voice level (L_{min}/L_{max}) as well as their differences Δf and ΔL.
In Fig. 3 from these standard parameters a clear decrease of both extreme fundamental frequencies can be observed, with the tonal range varying strongly during the treatment between -7 and +6 semitones, in average rather unchanged. Regarding the dynamics, an increase of the minimum voice level and a decrease of the maximum voice level can be observed, resulting in an overall decreased voice dynamics of ca. -8 dB in average, ranging from -5 to -18 dB.

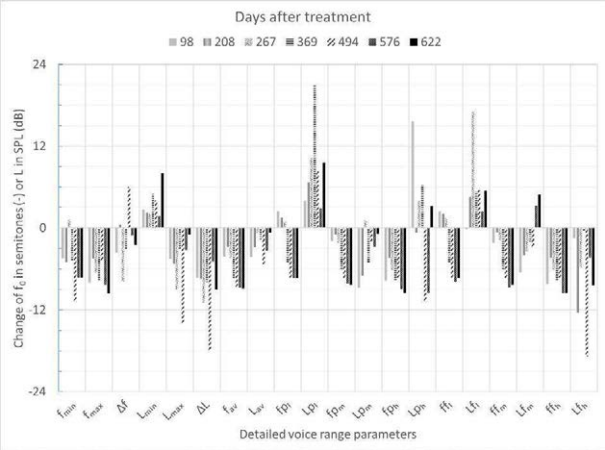


Fig. 4: Development of detailed voice ranges in study A

The average frequency (f_{av}) and average level (L_{av}) of the six segments are then represented as the next 2 parameters; these were indicated as a green circle in the Figs. 1&2. The analysis of the average fundamental frequency f_{av} shows a clearly decreasing slope from -3 to -6 semitones, and L_{av} – compared to the standard level differences – indicates a much smaller variation of voice level from < -1 dB to -5 dB, recovering at the end of the observation period.

The detailed analysis of the six segments is then shown in the following 12 parameters with labels indicating the level (L) or frequency (f) of the piano (p) or forte (f) curve with a subscript indicating the low (l), mid (m) or high (h) range of the VRP. As an example, "Lp_m" is the level of the piano curve in the medium segment.

In the six segments the direction of development of the voice ranges can be observed in detail: almost all ranges exhibit a clear trend of development during the duration of the study.

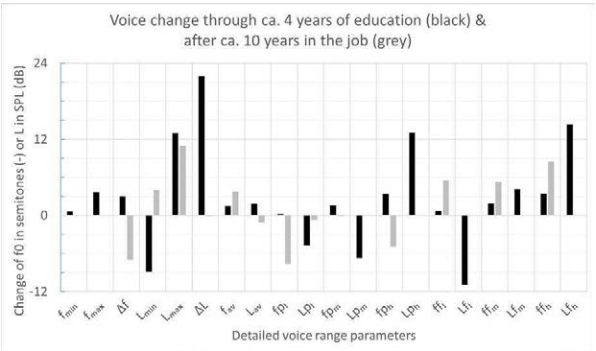


Fig. 3: Example of DVRP development in study B

In Fig. 4 the development of the teacher of study B is shown more in detail. Whereas the standard evaluation suggests that the tonal range has only slightly increased after the education and has even be reduced in the job, the forte curve has both increased its frequency range in the medium and high frequency range while in the job.

IV. Discussion

The DVPR allows a more precise view on the edges of the VPR. The averaging of several measurement points can improve the reliability of the comparison of VRP parameters. While the standard values can still be useful for a global assessment of voice development, the detailed edge values provide insight into the specific nature of this development – for example, in the soft–low register or the forte mid-range of the voice. The configuration of the parameters for removal of outliers have been adjusted to give satisfactory results for most of the recorded VRPs. However, if many artefacts are present, a manual cleaning in the lingwaves software is needed to obtain satisfactory results.

V. Conclusion

The detailed voice range profile can help to better document quantitatively the development of the singer's capability at the "corners" of the voice range profile. Since the details are extracted from averaged regions, they are rather stable against outliers during the measurements that occur frequently even with experienced instructors and singers.

The example of the voice range development of a trans person shows that the detailed analysis reveals clearly the development in the corners of the VRP and confirms that e.g. the tonal range has shifted significantly downwards while the dynamics are not reduced.

The detailed analysis of a voice teacher reveals in which pitch ranges the voice performance was improved even after years in the job.

ACKNOWLEDGEMENTS

We are grateful to the subjects who participated in this study and Jeremias Thiele for the support with the VRP measurements.

REFERENCES

[1] Schutte H.K. & Seidner W.: Recommendation by the Union of European Phoniatricians (UEP): Standardizing voice area measurement/phonetography. Folia Phoniatr. Logop. 1983; 35(6):286-288. doi:10.1159/000265703

[2] Schneider-Stickler, B. & Bigenzahn W.: Stimm-diagnostik, 2nd ed., Springer, 2013, pp.106-114.

[3] Ternström S. & Pabon P.: Voice Maps as a tool for understanding and dealing with variability in the voice. Applied Sciences. 2022; 12(22):11353. https://doi.org/10.3390/app122211353

[4] Source code of the DVRP on gibhub.com: https://github.com/voiceresearch/analysistools

MULTIPARAMETRIC VOICE ASSESSMENT OF ADDUCTOR SPASMODIC DYSPHONIA PRE- AND POST-TREATMENT WITH BOTULINUM TOXIN INJECTION: A PROPOSAL FOR IMPLEMENTATION

G. Baracca[1], C. Birca[1], M. Kob[2], R. Cenedese[1], G. Marioni[1], C. de Filippis[1], L. Franz[1]

[1] Phoniatrics and Audiology Unit, Department of Neuroscience DNS, University of Padova, Treviso, Italy
[2] Erich Thienhaus Institute, Detmold University of Music, Detmold, Germany
giovanna.baracca@gmail.com, cristina.birca@unipd.it, malte.kob@hfm-detmold.de, roberta.cenedese@unipd.it, gino.marioni@unipd.it, cosimo.defilippis@unipd.it, leonardo.franz@unipd.it

Abstract: **The aim of this study was to propose an updated panel of acoustic parameters to measure the effectiveness of Botulin toxin laryngeal injections for adductor spasmodic dysphonia (AdSD), including the motor speech diadochokinetic assessment.**
Study Design. This is a case report study
Methods**: Two cases diagnosed with adductor spasmodic dysphonia were evaluated at baseline, and one and three months after Botulinum toxin injection in the thyroarytenoid muscle. The voice protocol assessment included the self-assessment questionnaire, the traditional perceptual GRB Scale and the INFVo rating scale for substitution voices, the acoustic analysis of jitter%, shimmer%, glottal-to-noise excitation ratio, and the motor speech diadochokinetic parameters.**
Results**: In both clinical cases, the motor speech diadochokinetic parameters showed an improvement one month after the treatment compared to the baseline, and a subsequent decrease 3 months after treatment, in agreement with the trends showed by the self-assessment questionnaire. Conversely, the traditional perceptual and acoustic parameters did not demonstrate to be consistent in both cases with the self-assessment results.**
Conclusions**: The results suggested that motor speech diadochokinetics parameters could be relevant to assess the outcome of Botulinum toxin treatment for AdSD**
Keywords: **Adductor spasmodic dysphonia, Acoustic voice analysis, Motor speech diadochokinetic assessment**

I. INTRODUCTION

Laryngeal dystonia is a neurological voice disorder described as a task-dependent focal action-induction dystonia which affects laryngeal motor control [1,2]. It is classified in literature as spasmodic dysphonia, a low prevalence condition (1/100000), characterized by the presence of laryngeal spasms [3]. Two main different types of laryngeal dysphonia have been described: the most common adductor spasmodic dysphonia (AdSD), affecting 90% of cases and involving the overadduction of the vocal folds [4,5], with a female predominance and an average age of onset at 45 years, and the rarer abductor spasmodic dysphonia (AbSD), affecting vocal fold abductor muscles [6]. AdSD is caused by a spasm of the adductor muscles during phonation, occurring mainly during vowel emissions. Perceptual characteristics of AdSD include laryngeal spasms at the beginning of phonation, strained, groaning, staccato, effortful voice, intermittent voice breaks [5]. The causes of SD are still debated, and the assessment of this condition is based on clinical and perceptual voice evaluation, even if several acoustic and aerodynamic approaches have been proposed [7-10]. Moreover, machine learning strategies has been utilized in the last years to implement objectivation in the clinical setting [9,11-13]. However, a universally accepted consensus on the assessment of SD severity and treatment effectiveness still needs to be reached.

The effects of this voice impairment can have a profound impact on quality of life [14]: both the reduction of the intelligibility and the increase of vocal fatigue reduce the verbal capacity and frequently lead to devastating consequences in the efficiency of communication. AdSD is not exhaustively evaluated through the standardized basic multidimensional protocol proposed and revised by the European Laryngological Society [15,16]: it requests specific acoustic parameters for the running speech and the intelligibility [7,17]. One of the parameters employed to evaluate patients with neurogenic disorders of speech production is the oral speech diadochokinetic index (DDK), a fast repetition of syllables as in the sequence bʌ/tʌ/kʌ, in which the labial-coronal-dorsal succession is accompanied by rapid laryngeal adjustments given that the stop consonants are unvoiced and the vowel is voiced [18].

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

The DDK is successfully measured by commercially available acoustic software packages which provide semiautomated means to calculate rate and regularity of syllable repetitions. Although they are programmed for oral speech DDK, they can efficiently measure the laryngeal DDK (LDDK) as well [19,20]. However, the use of for LDDK to evaluate the neuromuscular impairment of the laryngeal mechanism is currently limited mainly to the dysarthria setting [19,20].
The aim of this study was to propose an updated voice evaluation protocol to measure the effectiveness of Botulin toxin laryngeal injections for AdSD, including the motor speech diadochokinetic assessment as an indicator of fluency and intelligibility.

II. METHODS

A. Clinical cases
Case 1. A retired 81-year-old woman came to the Phoniatric and Audiology Unit of the University of Padua/AULSS2 Marca Trevigiana with a diagnosis of AdSD to investigate the opportunity of a voice treatment. She complained about reduced voice intensity, vocal fatigue, and reduction of speech fluency with tremor and voice breaks.
Case 2. A retired 73-year-old man reported history of dysphonia for 20 years, with a progressive increase of vocal weakness, decreased vocal loudness and fluency impairment. Vocal tremor was also associated. He underwent a neurological consult, receiving diagnosis of vocal dystonia in association with high frequency tremor of the head. A diagnosis of AdSD was therefore posed, and the patient underwent a treatment with laryngeal injection of Botulinum toxin at our department.
In both cases the videolaryngostroboscopic examination was performed to confirm the diagnoses.

B. Treatment
Percutaneous injection of 2,5 to 5 International Units of botulin toxin A (BoTox) in the thyroarytenoid muscles, via cricoarytenoid approach under electromyographic control, was performed bilaterally (with a 7-day interval between each injected side). All procedures were performed by the same surgeon (LF).

C. Voice Assessment
The multiparametric voice assessment was performed before, and 1 and 3 months after BoTox injection.
The protocol consisted of the following aspects.

Perceptual evaluation. Both patients were asked to read the first paragraph of the Italian text “Il deserto”, which is phonetically balanced. The digital recording was made with a sampling frequency of 44,100 Hz, in a quiet room with less than 40 dB of background noise. The assessment was performed by 1 otolaryngologist and 2 phoniatricians. The analysed parameters for the GRB Scale were: global grade of dysphonia (G), roughness (R) and breathiness (B) from the GRBAS Scale [21] (Hirano), while the parameters dedicated to SD were intelligibility (Int), fluency (Fl), voicing (Voic), spasmodicity (Sp) from the INFVo Scale, defined for substitution voicing assessment by Moerman et al [22]. All the parameters were scored from 0 to 10 on a visuo-analogic scale, from the absence to the most severe impairment of the voice.
Self-evaluation parameters by use of the voice handicap index 10 (VHI-10) questionnaire [23]. It consisted of 10 items, which are to be scored from 0 (healthy voice condition) to 4 (most severe voice condition) by the patient himself.
Acoustic analysis of the vowel /a:/. The voice sound signal was captured and recorded by means of a Lingwaves device (Wevosys), at a sampled rate of 44,100 Hz, in a quiet room with less than 40 dB of background noise. The most stable second of the sustained emission of the vowel /a:/ at a comfortable pitch and level was selected and analyzed, with the extraction of the jitter%, shimmer% and glottal to noise excitation ratio (GNE). The patients were asked to repeat the recordings three times, and the most stable one was selected for analysis.
Motor speech diadochokinetic assessment (MSDA). The voice signal was captured and recorded by means of a Lingwaves device, using the MSDA protocol, at a sampled rate of 44.100 Hz, in a quiet room with less than 40 dB of background noise. The two patients were asked to repeat the English word “buttercup” (a sequence of bʌ/tʌ/kʌ) as many times as possible after a deep inspiration. The following parameters were captured: diadochokinetic index (DDK) and the DDK standard deviation as a measure of the regularity of the spoken words (DDK SD), both in syllables per second (syl/s).

III. RESULTS

The perceptual and self-evaluation parameters, collected before, one and three months after the treatments, are shown in table 1a (first case) and 2a (second case). In the first case all the perceptual parameters showed an improvement one month after the injection and some of them a slight reduction 3 months after. The VHI-10 was decreased 1 month after treatment (from 28 to 19 in the first patient, from 24 to 8 in the second patient), then raising again to 27 in the first patient and to 20 in the second one, along with the reduction of BoTox pharmacological effect. Concerning the acoustic parameters (jitter %, shimmer % and GNE), their values at the baseline, and one and three months after treatment are shown in figures 1a (first case) and 1b (second case). The jitter % showed a progressive decrease in the first case, from 3.57 to final 1.73, while

it increased and then reduced its value in the second case (3.97-5.83-4.87). The GNE seemed to be quite stable before and after treatment. The trends of the MSDA values, including the DDK and its standard deviation are described in figures 2a and 2b. In both cases they improved 1 month after treatment and showed a tendency to return close to the initial value after 3 months.

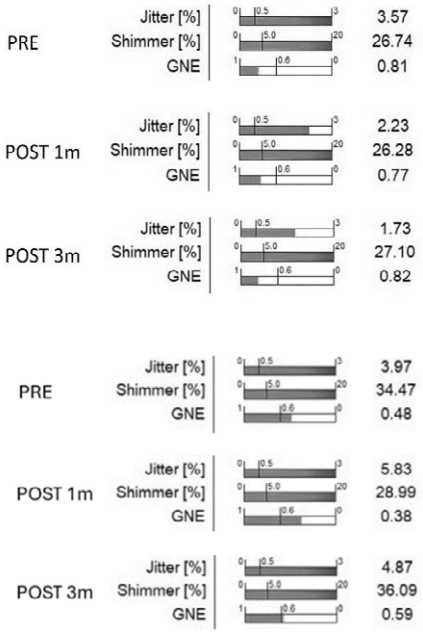


Fig 1a (top) and 1b (bottom): results of the acoustic parameters in case 1 and case 2, before, 1 and 3 months after the Botulinum toxin injection

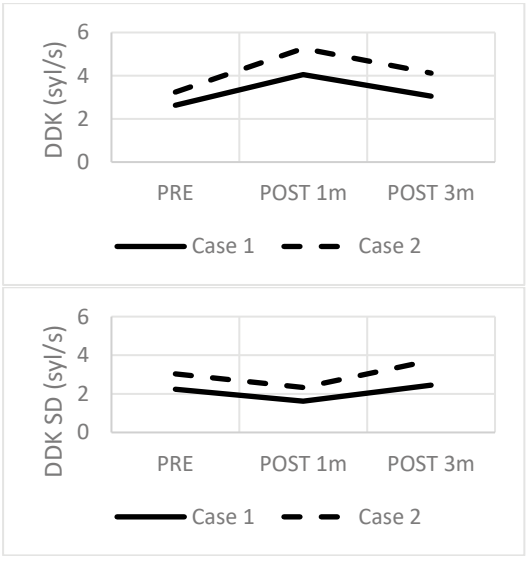


Fig 2a and 2b: results of the MSDA: DDK (top) and DDK SD (bottom) in case 1 and case 2, before, 1 and 3 months after the Botulinum toxin injection

Time	G	R	B	Int	Fl	Voic	Sp	VHI 10
PRE	6	3	5	4	8	1	5	24
POST 1 m	6	3	3	3	4	3	3	8
POST 2 m	6	3	4	3	5	2	4	20

Time	G	R	B	Int	Fl	Voic	Sp	VHI 10
PRE	8	6	6	2	8	5	8	28
POST 1 m	2	4	2	1	4	2	4	19
POST 3 m	4	3	5	1	6	5	6	27

Tab 1a (top) and 1b (bottom): results of the perceptual and self-evaluation parameters in case 1 and case 2, before, 1 and 3 months after the Botulinum toxin injection.

IV. Discussion

The main purpose of this work was to describe the option to implement the acoustic analysis to better evaluate the outcome of the BoTox treatment for AdSD, through the assessment of DDK and its standard deviation as a measure of communication impairment. There is not a strict consensus about the protocol to determine the severity of the AdSD. Indeed, the traditional voice multiparametric assessment for dysphonia established by the ELS is not considered so suited for this pathological condition [7]. AdSD is mainly associated with irregularity of the fundamental frequency, increased noise energy, deterioration of harmonic structure and increased speaking and articulation times [24,25]. Dejonckere et al. [7] proposed a new perspective to establish the outcome of the Botulinum toxin treatment by means of both voice perceptual assessment and analysis program AMPEX (Auditory Model Based Pitch Extractor). One of the most interesting aspects of this study was the three-dimensional assessment approach of AdSD: perceptual rating with traditional and dedicated parameters, objective acoustic assessment using a specific program, and self-reported quality of life. Focusing on the severe consequences in terms of fluency and intelligibility in patients affected by AdSD, the protocol proposed in this study included the evaluation of the functional conditions affecting the communication skills in the assessment of the outcome of the botulinum toxin injection. We conducted a multidimensional evaluation of the voice one month after the treatment, when the optimal effect of the botulinum toxin is expected, and again three months after the treatment, when its effect typically begins to diminish. We combined the self-evaluation questionnaire, the perceptual parameters considered by Dejonckere et al. [7], and the traditional acoustic parameters with the MSDA. MSDA results showed, in both clinical cases, an improvement one month after the treatment compared to the values before

the injection, and a decrease of the values 3 months after the treatment, in agreement with the trends of VHI values. These results could be explained by the well-known temporary effect of Botulin toxin injection, which is strong during the first months and then starts fading. Interestingly, the MSDA parameters showed in both cases a specular trend compared to the self-assessment questionnaire. This tendency seemed to be less evident for the acoustic traditional parameters such as jitter%, shimmer% and GNE and the perceptual ones. These results suggest that the assessment of adSD and its treatment should include functional parameters representative of the communication impairment.

V. Conclusion

Describing these two cases report, we found that the MSDA parameters might be relevant to assess the outcome of the Botulinum toxin treatment for AdSD, more than the traditional acoustic and perceptual parameters. More data are necessary to quantify the reliability of this voice assessment proposal.

REFERENCES

[1] C. Ludlow, “Spasmodic dysphonia: a laryngeal control disorder specific to Speech”, J. Neurosci, vol. 31, pp. 793–797, 2011.

[2] T.K. Meyer, “The treatment of laryngeal dystonia (spasmodic dysphonia) with botulinum toxin injections”, Oper. Tech. Otolaryngol, vol. 23, pp. 96–101, 2012.

[3] J.M. Hintze, C. Ludlow, C, S. Bansberg et al, “Spasmodic Dysphonia: A Review. Part 1: Pathogenic Factors”, Otolaryngol. Head Neck. Surg, vol.157, pp. 551–557, 2017.

[4] N. Roy, “Differential diagnosis of muscle tension dysphonia and spasmodic Dysphonia”, Curr. Opin. Otolaryngo.l Head Neck. Surg, vol. 18, pp. 165–170, 2010.

[5] A. Blitzer, “Spasmodic dysphonia and botulinum toxin: experience from the largest treatment series”, Eur. J. Neurol, vol. 17(suppl. 1), pp. 28–30, 2010.

[6] H.A. Jinnah, A. Berardelli, C. Comella, et al, “The focal dystonias: Current views and challenges for future research”, Mov. Disord, vol. 28, pp. 926–943, 2013.

[7] P.H. Dejonckere, K.J. Neumann, M.B.J. Moerman, et al, “Tridimensional assessment of adductor spasmodic dysphonia pre- and post-treatment with Botulinum toxin”, Eur. Arch. Oto-Rhino-Laryngol, vol. 269, pp. 1195–1203, 2011.

[8] G. Cantarella, A. Berlusconi, B. Maraschi, B, et al, “Botulinum toxin injection and airflow stability in spasmodic dysphonia”, Otolaryngol. Head Neck Surg, vol. 134, pp. 419–423, 2006.

[9] A. Suppa, F. Asci, G. Saggio, et al, “Voice analysis in adductor spasmodic dysphonia: Objective diagnosis and response to botulinum toxin”, Park. Relat. Disord, vol. 73, pp. 23–30, 2020.

[10] N. Roy, A.M. Ma, S.N. Awan, “Automated acoustic analysis of task dependency in adductor spasmodic dysphonia versus muscle tension dysphonia”, Laryngoscope, vol.124, pp. 718–724, 2013.

[11] G. Costantini, P. Di Leo, F. Asci, et al, “Machine Learning based Voice Analysis in Spasmodic Dysphonia: An Investigation of Most Relevant Features from Specific Vocal Tasks”, in Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2021), Vienna, 2021, vol. 4, pp. 103–113.

[12] M.E. Powell, M.R. Cancio, D. Young, D, et al, “Decoding phonation with artificial intelligence (D e P AI): Proof of concept”, Laryngoscope Investig. Otolaryngol, vol. 4, pp. 328–334, 2019.

[13] F. Calà, L. Frassineti, C.Manfredi C, et al, “Machine Learning Assessment of Spasmodic Dysphonia Based on Acoustical and Perceptual Parameters”, in Bioengineering, Basel, 2023 vol 28;10(4), pp. 426.

[14] M. Branden, M. Johns, A. Klein A, et al, “Assessing the effectiveness of botulin toxin injections for adductor spasmodic dysphonia: clinician and patient perception”, J. Voice, vol. 24, pp. 242–248, 2008.

[15] P.H. Dejonckere, P. Bradley P, P. Clemente P, et al, “A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS)”, Eur. Arch. Otorhinolaryngol, vol. 258(2), pp. 77-82, 2001.

[16] J.R. Lechien, A, Geneid A, J.E. Bohlender, et al, “Consensus for voice quality assessment in clinical practice: guidelines of the European Laryngological Society and Union of the European Phoniatricians”, Eur. Arch. Otorhinolaryngol, vol. 280(12), pp. 5459-5473, 2023

[17] M.P. Cannito, M. Doiuchi, T. Murry, et al, Perceptual structure of adductor spasmodic dysphonia and its acoustic correlates”, J. Voice, vol. 26, pp. 818.e5-818.e13, 2012.

[18] R.D. Kent, “Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders”, Am. J. Speech Lang. Pathol, vol. 5, pp. 7–23, 1996.

[19] T. Louzada, R. Beraldinelle, G. Berretin-Felix, et al “Oral and vocal fold diadochokinesis in dysphonic women”, J. Appl. Oral Sci, vol. 19, pp. 567– 572, 2011.

[20] L. Lombard, N.P. Solomon, “Laryngeal Diadochokinesis Across the Adult Lifespan”, J. Voice, vol. 34, pp. 651-656, 2020.

[21] M. Hirano, “Clinical Examination of Voice”, in Disorders of Human Communication, Springer: Wien, Germany, 1981, pp. 1–99.

[22] M.B.J. Moerman, J. Martens, M. Van der Borgt, M, “Perceptual evaluation of sub-stitution voices: Development and evaluation of the (I)INFVo rating scale”, Eur. Arch. Otorhinolaryngol, vol. 263, pp. 183–187, 2006.

[23] C.A. Rosen, A.S. Lee, J. Osborne, et al, “Development and validation of the voice handicap index-10”, Laryngoscope, vol. 114, pp. 1549-1556, 2004.

[24] C.L. Ludlow, R.F. Naunton, S.E. Sedory, et al, “Effects of botulinum toxin injections on speech in adductor spasmodic dysphonia”, Neurology, vol. 38, pp. 1220–1225, 1988.

[25] M.P. Cannito, G.E. Woodson, T. Murry, et al, “Perceptual analyses of spasmodic dysphonia before and after treatment”, Arch. Otolaryngol. Head Neck Surg, vol. 130, pp. 1393–1399, 2004.

LONGITUDINAL STUDY OF VOICE PROPERTIES IN MUSIC PEDAGOGY STUDENTS

Jeremias L. Thiele[1], Elke Nagl[1], Malte Kob[1,2]

[1] Antonio Salieri Institute, mdw, Vienna, Austria
[2] Erich Thienhaus Institute, Detmold University of Music, Germany

jeremias.thiele@students.mdw.ac.at
nagl-e@mdw.ac.at
malte.kob@hfm-detmold.de

Abstract: **In this longitudinal study aspects of vocal performance in music teachers (n=42) are investigated at the beginning and at the end of their university studies, as well as at least ten years after graduation. We used voice range profiles (VRP) investigating both the singing and speaking voice to characterize the development. Whereas VRP parameters show a clear increase in vocal abilities during training, the third measurement during working life reveals a stagnation or even regression in the vocal performance represented in the VRP.**
Keywords: **Voice Range Profile, Singing Pedagogy, Voice Development, Teachers**

I. INTRODUCTION

Teaching in schools is a vocally demanding profession, which, according to various studies, increases the risk of voice problems. Roy et al. [1] used a questionnaire to determine a significantly higher rate of voice disorders among teachers compared to non-teachers. Sliwinska-Kowalska et al. [2] can confirm this result with a study that includes both questionnaires and laryngological examinations. Russell et al. [3] report significantly higher rates of voice problems for female teachers compared to male teachers. However, the reported prevalence rates of voice problems in teaching professions vary considerably, which is not surprising given the different definitions and self-assessment methods used.
One clinically used measurement method for assessing vocal performance, but not necessarily health, is the VRP. In the LSME project of the University for music and performing Arts Vienna (mdw), VRPs were used to document the vocal development of teachers in training and in their professional careers. More than in previous studies, the test group received extensive training in singing. Between 2008 and 2025, this longitudinal study documents the vocal development of school teachers in music education. Initial evaluations were carried out by Elke Nagl in 2008 [4] and Carina Kellner in 2019 [5].

The education of the subjects at the mdw usually lasts four years and includes one-to-one singing lessons of 1.5 hours per week to prepare them for a teaching profession. The occupation is characterized by several hours of speaking and singing at high volume in the classroom. However, not all subjects entered a profession as a teacher after completing their studies; some pursued a career in music, while others chose professions outside of teaching.

The pool of test subjects at the mdw is a group that can provide valuable insights for voice research, particularly in relation to the singing voice and under pedagogically controlled conditions. However, the long duration of the project also creates organizational difficulties. Especially inconsistent recording conditions and varying instructions potentially reduce the validity of the results.
The study focused on measuring VRP, but evaluates as well other acoustic aspects of the voice, such as timbre. We focus on VRP analysis here, an extended analysis of the recorded data will follow.

II. METHODS

The measurements were carried out at the beginning (measurement A) and at the end of the education degree program (B). Individuals who only took part in the first measurement were not included in the rest of the study. For some subjects, a further measurement was carried out approximately 10 years after completing their degree (C). As it was not possible to contact all individuals after such a long time, the group size for C is much smaller. In addition, a few VRP from measurements A and B could not be used, reducing the number of subjects to be analyzed intra-individually to 12.

Tab. 1: Overview of subjects

	A	B	C	All
number of subjects	42	42	15	12
number of female subjects	30	30	11	9
number of male subjects	12	12	4	3

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

Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

Since the sample size is relatively small, especially for male subjects after education, this project is considered a pilot study, and the results should be verified and supplemented by further projects with a similar design.
The protocol included several tasks, each of which had to be sung and spoken in a controlled manner. The exercises were repeated as necessary in order to document the best possible vocal control.
The exercises were as follows: A VRP for speaking and singing voices to determine the dynamic range (DR, measured in [dB]) and frequency range (FR, measured in semitones[1] [st]). The VRP is a standardized representation of vocal limits that is also used for clinical applications [6]. It involves singing at the softest and loudest possible voice intensity in a slow, continuous sequence of tones, usually on the vowel /a:/. The two sequences recorded in this process show the control over the voice per pitch and can therefore also be used to display different registers or even the passaggi. The speaking VRP is measured at the volumes soft, normal, loud, and shouted by counting from 20 to 30 at a comfortable pitch.
In addition, a standard spoken text in German ("Der Nordwind und die Sonne"), and a simple singing example, "Summertime", were recorded. In measurement A, no style was specified for "Summertime", while in B and C, the test subjects were asked to sing in pop and classical styles. Further exercises were a glissando across the entire vocal range (usually on the vowel /a:/), and a vowel sequence (/a:/ - /e:/ - /i:/ - /o:/ - /u:/ - /ə/) at constant middle pitch. In the third series of measurements, some more exercises were performed with additional dynamics and vowels. These exercises provide further opportunities to investigate timbre or voice irregularity.
The acoustic conditions could not be maintained due to the long duration of the study. The first two measurements took place in classrooms with medium reverberation time. The third series of measurements was carried out in a soundproof and anechoic booth. The Lingwaves system was used for the VRP measurement as well as for the additional recordings. This system consists of a level-calibrated measurement microphone positioned 30 cm in front of the mouth and converts the audio signal into digital wav format at a sampling rate of 22,050 Hz.
The VRP is evaluated using Lingwaves and software developed at the mdw.
During measurement C, a logopaedic screening was carried out before the measurements started. No voice disorders were detected in this screening.
Questionnaires were used to monitor the development of the participants during the C measurements. These questionnaires assessed the professional development, daily voice use, and the participants' own assessment of their vocal health. The results of these questionnaires were used to divide the participants into different groups, according to gender, voice type, profession, and daily voice use.
All subjects were anonymized using a three-digit ID, which were used for intra-individual investigations.

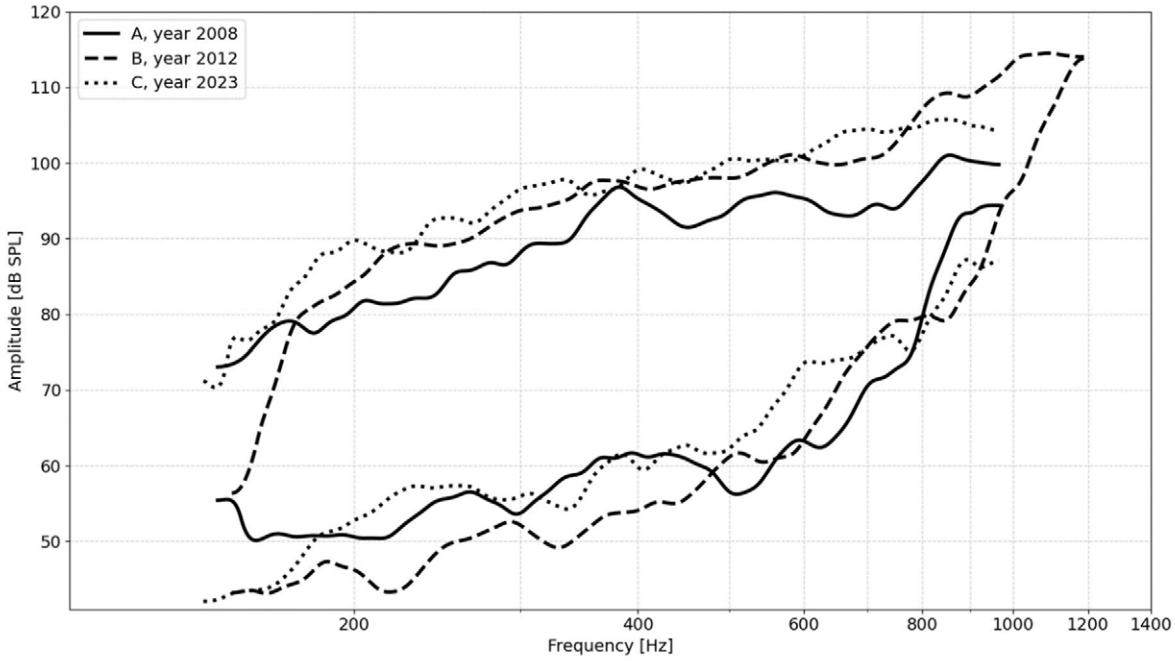


Fig. 1: singing VRPs of subject ID009

[1] Semitone (st) refers to the equal-tempered semitone, i.e. the interval of a frequency ratio of $2^{1/12}$

III. RESULTS

An example of the three singing VRPs for one subject (ID009/female) is shown in Fig. 1, illustrating a marked increase of the VRP coverage between A and B, and a slight decrease towards C. Compared to an average female VRP in the Lingwaves system, the coverage is 82.5% in A, 104% in B, and 102% in C. The gain in DR is particularly strong, and persists partially even after training (DR_A=53 dB, DR_B=75 dB, DR_C=64 dB). FR increased to a lesser extent (FR_A=33 st, FR_B=36 st, FR_C=33 st). After graduating, this subject pursued the career path envisaged during her training and became a teacher. Measurement C shows that, despite the very voice-intensive nature of the subjects' daily work, it was possible to maintain almost the same level of vocal ability as before. Particularly the profile for loud voice use decreases only very little in comparison to B.

The evaluation of all VRPs shown in Tab. 2 reveals significant improvements in B, resulting in 117% coverage compared to an average VRP, particularly due to an increased dynamic range. The differences in the speech VRP ranges and volumes are less clear.

Tab. 2: Results of all subjects VRP measurements

	A		B		C	
	⌀	Std	⌀	Std	⌀	Std
coverage [%]	87	20	117	13	100	20
singing DR [dB]	56	8	64	5	65	5
singing FR [st]	35	4	38	3	38	4
speech DR [dB]	46	6	50	6	40	14
speech FR [st]	13	4	13	4	12	4
$vol_{speech,\ soft}$ [dB]	52	5	48	4	52	6
$vol_{speech,\ normal}$ [dB]	65	5	65	4	64	5
$vol_{speech,\ loud}$ [dB]	78	6	74	4	75	6
$vol_{shouting}$ [dB]	98	5	98	4	92	10

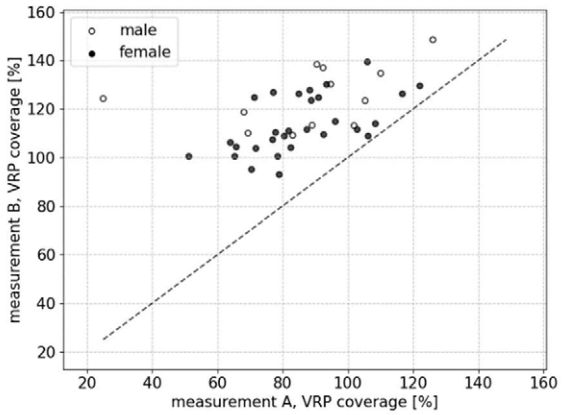


Fig. 2: Intra-individual comparison between VRP coverage in A and B

Fig. 2 shows the development of the singing VRP coverage of all test subjects between A and B. Without exception, all individuals are placed left from the diagonal, i.e., they achieved an increase in VRP coverage. It is also clearly visible that the group studied becomes more harmonious, i.e., that the individuals with the "weaker" results in A show particularly strong improvements. Measurement A indicate the majority of subjects below 100% coverage, while after education only two people were slightly below 100%.

As shown in Fig. 3, all individuals stagnate or even decline in VRP coverage after graduation. The significantly smaller scaled diagram also contains selected characteristics of the subjects, namely their career path and daily voice use according to the questionnaire. It is visible that persons in the teaching profession, marked with squares, show both nearly constant coverage and a significant decline. The time of daily voice use also shows no obvious correlation. All three male subjects are close to the diagonal line, representing comparably less reduction in voice range after education.

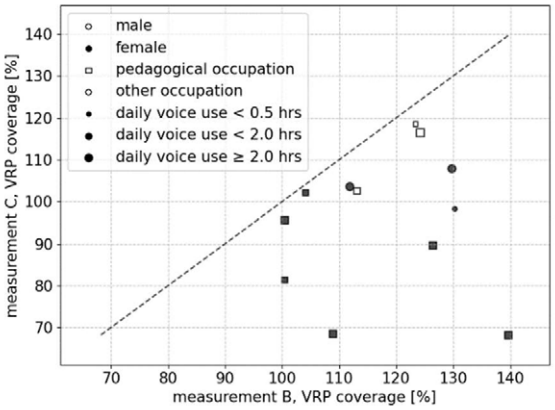


Fig. 3: Intra-individual comparison between VRP coverage in B and C

The most important vocal activity for teaching is speaking aloud. The mean voice level shows a slight decrease between A and B. The development to C shows a relatively constant mean value. In Fig.4, no clear correlation with career path, gender, or daily voice use is apparent.

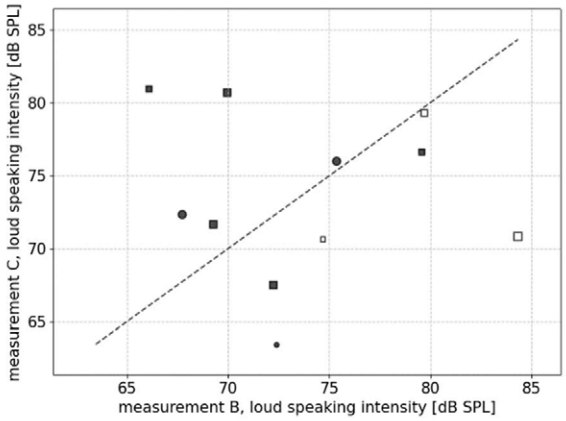


Fig. 4: Intra-individual comparison between loud speaking volume in B and C

IV. Discussion

It is not surprising that the measurements show significant improvements in vocal abilities during the study program, but it does confirm both the effectiveness of the teaching and the significance of the VRP measurements. It is interesting to note that in all recorded cases, vocal ranges decline or stagnate after graduation. While the number of subjects is small, males appear to have a more consistent development in this aspect, and females in some cases show larger declines. Previous studies [3] have shown that women have a higher prevalence of voice disorders in the teaching profession, our findings seem to indicate a similar trend for a greater decline in females.

No clear correlations between VRP coverage and daily voice use or career path can be found. It should be noted that the other career paths do not necessarily involve minimal use of the voice; one subject, for example, worked as a self-employed singer.

No acute voice disorders were detected in the logopaedic screening, which is also reflected in the mean values of the VRPs. However, it must be considered that the subjects do not cover all age groups, but are mostly under 40 years of age, and therefore our study does not yet show any effect over the entire career. Further studies with the trained teachers would be of interest.

While after education generally the mean range of the singing voice decreases and the level of the shouting voice also declines, the level of the loud speaking voice remains relatively constant. This mode of phonation is probably closest to the daily requirements in the classroom and thus shows that the subjects are still able to meet the professional demands.

The findings of this study are based only on acoustic measurements and therefore do not give us a deeper understanding of the techniques and singing strategies that the students might have used in measurements. Changes in singing or speaking technique between measurements A, B and C, and the influence of pedagogical support during education, were not fully documented in this study. However, this remains an interesting question for future research. Other measurement methods, such as EGG or respiratory measurement devices, could provide some insights.

V. Conclusion

The study confirms expectations of a significant change in vocal abilities over the course of the degree program, this is evident in a larger vocal range shown in the VRP. The development of the test subjects in their professional lives, on the other hand, is very heterogeneous; in all cases, the performance level represented in the VRP coverage stagnates or even declines. However, the mean coverage remains average and therefore does not appear to indicate any voice problems despite intensive daily use. Also, the abilities in loud speaking voice are constant on average between B and C.

The results in this paper provide an overview of the findings of the LSME project; more comprehensive evaluations will follow.

Further studies on the details of vocal technique development during the study course and a comparative study of the mdw students with vocal training against untrained singers are possible next areas of research.

ACKNOWLEDGEMENTS

We would like to thank the persons who volunteered to participate as subjects, speech therapist Gerlinde Tichy, and former assistants Carina Kellner, Anna Hurch, and Michaela Mayr for their support.

REFERENCES

[1] N. Roy, R. Merrill, S. Thibeault, R. Parsa, S. Gray, E. Smith. "Prevalence of Voice Disorders in Teachers and the General Population", *Journal of Speech, Language, and Hearing Research*. 47. 281-293. 10.1044/1092-4388(2004/023)

[2] M. Sliwinska-Kowalska, E. Niebudek-Bogusz, M. Fiszer, T. Los-Spychalska, P. Kotylo, B. Sznurowska-Przygocka, M. Modrzewska. "The prevalence and risk factors for occupational voice disorders in teachers", *Folia Phoniatr Logop.* 2006; 58(2):85-101

[3] A. Russell, J. Oates, K. M. Greenwood. "Prevalence of voice problems in teachers", *J Voice*. 1998 Dec;12(4):467-79

[4] E. Nagl, „Zur gesangspädagogischen und akustisch-physiologischen Empirie der Singstimme von Musik- und Gesangspädagoginnen in der universitären Ausbildung und Berufstätigkeit", PhD thesis, 2008, mdw, Vienna

[5] C. Kellner, „Visualisierung stimmlicher Entwicklung mittels Stimmfeldmessungen während einer Gesangsausbildung in musikpädagogischen Studienrichtungen", Bachelor thesis, 2019, mdw, Vienna

[6] B. Schneider-Stickler, W. Bigenzahn, *Stimmdiagnostik*, 2nd ed., Springer, 2013, pp.106-114

THE ACOUSTIC PERFORMANCE OF A TRANSGENDER SINGER

Henry Browne[1], Malte Kob[1,2], Jeremias Louis Thiele[1], Ines Kansy[3], Berit Schneider-Stickler[3]

[1] University of Music and Performing Arts Vienna, Austria
[2] Erich Thienhaus Institute, Detmold University of Music, Germany
[3] Vienna General Hospital (AKH)/Clinical Department of Phoniatrics and Speech Therapy, Vienna, Austria

henry.browne@students.mdw.ac.at, malte.kob@hfm-detmold.de, jeremias.thiele@students.mdw.ac.at, ines.kansy@meduniwien.ac.at, berit.schneider-stickler@meduniwien.ac.at

Abstract: **So far little is known to which extent a transmasculine classical singer can regain his singing abilities after a testosterone-induced vocal transition. This study observes the vocal transformation of a 27-year-old classical mezzosoprano-baritone undergoing testosterone treatment who has undergone professional vocal training for five years to date.**
A wide array of methods, such as Voice Range Profile (VRP) measurements, laryngostroboscopy, electroglottography (EGG), voice acoustics and expert interviews have been applied.
After 24 months of testosterone therapy, the laryngostroboscopy revealed a thickening of the vocal folds. The VRP showed a transition of the voice into the typical male range (baritone). The VRP analysis also reveals a recovery of the dynamic and tonal deficiencies that occurred after initial testosterone application. The voice quality was characterized initially after a pharmacologically induced vocal transition by instability and loss of brilliance which has been improved by pedagogically informed vocal training. The endurance of voice performance was reduced with tendency of improvement. The findings of this study indicate that vocal recovery is possible for transmasculine classical singers. To which extent, is subject to further examinations.
Keywords: Classical singer, Transgender, Voice Range Profile, Laryngostroboscopy

I. INTRODUCTION

Research on the singing voice in transmasculine individuals is limited [2]. The current state of research is based largely on observations and assumptions [1,2]. Research has shown that transmasculine individuals often experience limitations with their singing voice post vocal transition [1]. The extent of the influence of possible functional problems in the use of the voice remains unclear [2].
Exogenous testosterone treatment in transmasculine individuals has shown to induce the following changes: The vocal folds thicken due to muscle hypertrophy [1,4]; the vocal fold length usually stays stable or increases only to a minimal extent [4]. As a result, the voice shifts to a lower range. To date, no scientific evidence has demonstrated growth of the laryngeal cartilages under exogenous testosterone treatment [4].
Vocal registers, and their coordination, are an important feature of the singing voice. The terms M1 and M2 correspond to what is commonly referred to as the modal/chest register (M1) and the falsetto/head register (M2). Their detailed explanation and definition based on electroglottographic (EGG) analysis are provided in [6]. In this study, the distinction between M1 and M2 is based on variations in the vibratory mass of the vocal folds, producing different ratios of open and closed quotients in the vibratory pattern as identified through EGG measurements [6].
The limited research shows trends in the development of the M1 and M2 registers under exogenous testosterone treatment: The M2 register is often lost early into hormone therapy, or gains an airy quality [1,4]. The transition between M1 and M2 registers becomes less stable, resulting in the loss of notes in the transition area [1,4]. After ~6 months, the transition area between M1 and M2 progressively diminishes, while the newly developed M2 register gains stability and strength [4]. However, not all singers regain access to their M2 register [1]. In some transmasculine singers, limitations in range and vocal quality improve over time, whereas in others, these issues persist [1].
To our knowledge, no research has yet examined why the use or coordination of the M1 and M2 registers appears to be sensitive to hormone therapy. This is a question we want to investigate further. In our results to date, we observed changes in the vibratory pattern through laryngostroboscopic examinations. A similar finding was reported by Agha&Hynes [1], who documented alterations in vibratory behavior characterized by an increased muscle tension pattern.
Using the Voice Range Profile (VRP), we have been able to document the voice shift and voice quality change of a transmasculine classical singer for the first time for about 2 years. We have assessed whether the resulting voice range is comparable to that of cisgender male professional operatic singers. The aim was to evaluate whether the voice of this transmasculine classical singer undergoing testosterone therapy can meet the requirements typically expected of professional singers in the Western operatic tradition.

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

II. METHODS

The observation period started in June 2023 and is ongoing. Testosterone treatment began in May 2023.
To investigate this voice transition, a large array of methods from acoustic methods via EGG to medical examinations and surveys have been applied.
Laryngostroboscopic examinations are carried out at a six-month interval at the Medical University of Vienna/Clinical Department of Phoniatrics and Speech Therapy to examine the influence of testosterone on the vocal folds. Testosterone level tracking took place by conducting bloodwork monthly.
VRP recordings are conducted every six months in the Vienna General Hospital (AKH)/Clinical Department of Phoniatrics and Speech Therapy with DIVAS software of the company XION, and on a quarterly basis in the voice research laboratory at the University of Music and Performing Arts Vienna (Mdw)/Antonio Salieri Institute in a recording booth with anechoic conditions using a Røde NT1 microphone and the recording analysis Software LingWaves using a sample rate of 22050 Hz. Voice range profiles (VRP) are graphical representations of a person's maximum and minimum phonation ability [5]. Ascending and descending Glissandi are recorded on vowels /a:/, /i:/ and /u:/ in minimum and maximum phonation intensity.
The Long-term average spectrum (LTAS) is recorded using a Røde NT1 microphone and analysed with praat-based scripts. Simultaneous electroglottographic recordings are conducted using an Electroglottograph of the company Glottal Enterprises, Model EG2-PCX2.
The vocal pedagogy work focuses on the maintenance of controlled exhalation, register negotiation and consistency of the vocal tract shape, especially in the pharyngeal area. The experience of the test subject is tracked through expert interviews and the keeping of a voice diary. Vocal coaching was conducted on average twice a week for 45 minutes.
Our results presented in this manuscript are restricted to the evaluation of VRPs and laryngostroboscopic examinations. Future analysis will be conducted using additional parameters.

III. RESULTS

A. Testosterone levels

At the start of testosterone treatment in May 2023, testosterone levels were at 0.5 ng/ml. Testosterone levels reached their peak by August 2024 with 7.3 ng/ml.
Levels stabilized at around 4.3 ng/ml between February and April 2025.

B. Laryngostroboscopic findings

Comparative pictures of beginning of observation period until May 2025 is shown in Fig. 1.

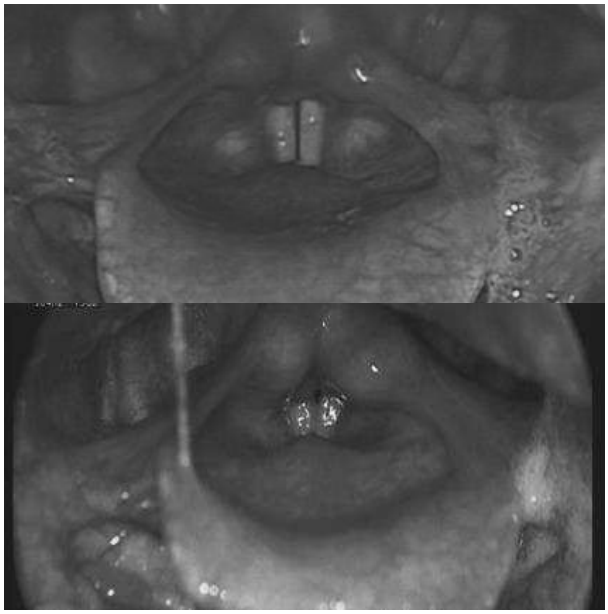


Fig. 1: Laryngoscopy June 2023 (top) and May 2025 (bottom)

A comparison of laryngostroboscopic images taken before and during testosterone treatment shows a significant increase in the mass of the vocal folds.
The increase in muscle mass is leading to a slowing of the vocal fold vibrations, resulting in a lowering of the speaking voice pitch and vocal range. In addition, changes in the ratios of opening, closing and closed quotients are observed, with an increase in the closing and closed quotients after testosterone therapy.

C. Dynamic range and frequency range

The development of the dynamic range and frequency range is shown in Fig. 2.

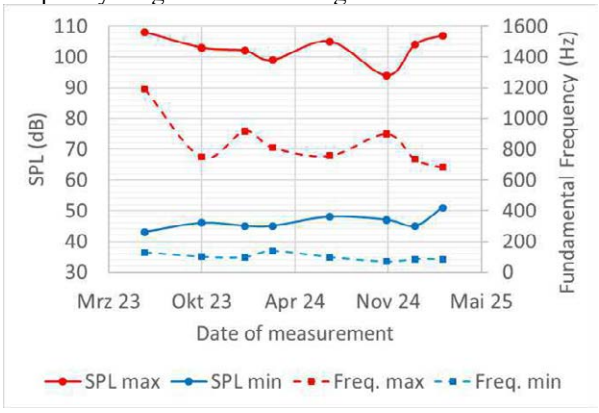


Fig. 2: *Dynamic range and frequency range*

Dynamic range and frequency range were plotted using VRP recordings.
Between June 2023 and November 2024, the maximum intensity fell continuously and reached its lowest point in November 2024. By May 2025, it had recovered

relevantly. From June 2023 to May 2025, the minimum intensity went through fluctuations but did not relevantly change.
The lower limit of the dynamic range of the singing voice reached its lowest point in November 2024 with a range of 47 dB SPL. This was relevantly below the initial dynamic range measurement on June 2023 with 65 dB SPL. By March 2025, it had reached its highest point at 56 dB SPL, before decreasing to 50 dB SPL in May 2025.
Between June 2023 and March 2025, the highest reachable pitch had decreased by a major sixth. Between March and May 2025, the highest reachable pitch expanded again by 4 semitones.
A first relevant drop occurred between June and October 2023. Until March 2025, the highest pitch decreased continuously before expanding again by May 2025. From June 2023 to May 2025, the low range of the singing voice expanded by a fifth/45 Hz.
The smallest vocal range was measured on the 21st of March 2024, consisting of 31 semitones – compared to the baseline measurement in June 2023 with 38 semitones. By May 2025, the range had expanded back to 36 semitones. VRPs recorded at AKH and at Mdw showed very similar results.

D. Detailed analysis of VRP
The detailed analysis of VRPs are shown in Fig. 3.

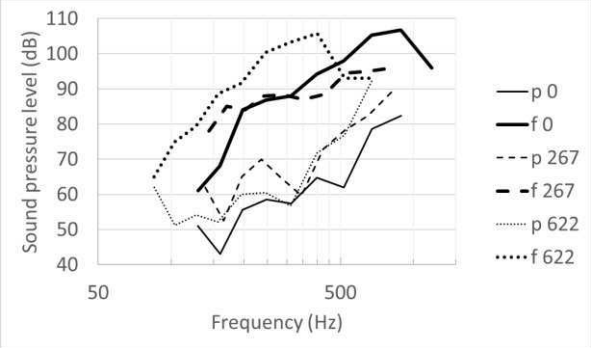


Fig. 3: *Detailed VPR June 2023, March 2024, March 2025. P0=piano curve June 2023, f0=forte curve June 2023. P 267=piano curve March 2024, f267= forte curve March 2024. P 622= piano curve March 2025, f 622= forte curve March 2025*

In the March 2024 VRP, the piano curve revealed a deficit in intensity throughout the frequency range.
The forte curve also displayed a reduction in intensity.
The forte curve displayed collapses in intensity throughout the frequency range. The piano curve also displayed sharp rises in intensity throughout the frequency range.
By January 2025, the piano curve had expanded relevantly in the low frequency range of the voice whereas the forte curve did not yet extend into the lower frequency region. In the middle and higher frequency region, the forte curve exhibited a smoother transition and no longer displayed the collapses in intensity observed in March 2024. The piano curve, however, continued to exhibit marked deficits in intensity throughout the frequency range.
By March 2025, the forte curve extended into the lower frequency range, aligning more closely with the piano curve. Both piano and forte curves displayed less variable transitions in intensity compared to earlier measurements.

E. Long-Term Average Spectrum (LTAS)
The results of the LTAS are shown in Fig. 4.

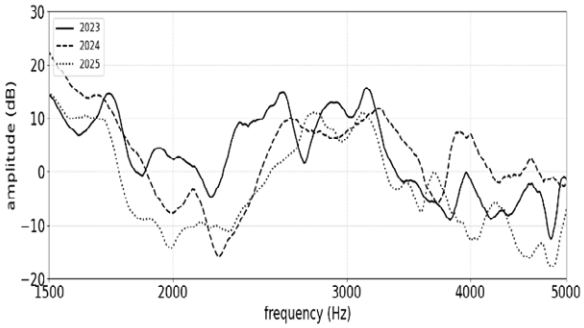


Fig. 4*: LTAS June 2023, March 2024, January 2025*

LTAS analysis was executed for three glissando recordings, one in June 2023, one in March 2024 and one in January 2025.
For all three recordings, there is a clearly discernible singer's formant cluster present at around 3000 Hz, consisting of the merging of the 3rd and 4th formant.
The singer's formant cluster did not significantly shift throughout the observation period but rather showed a consistent presence in the singing voice.

IV. Discussion

A. Testosterone levels
Testosterone levels reached their peak by August 2024, which subsequently coincides with a significant increase in the low range of the singing voice and a simultaneous shrinking of dynamic range between March and November 2024. One can assume that this correlation is due to thickening of the vocal folds by the testosterone treatment, and simultaneous increase of muscle mass of the vocal cords due to its growth, leading to a temporary decline in dynamic range.

B. Laryngostroboscopic findings
To our knowledge, previous studies on transmasculine individuals have only once been able to visually demonstrate the vocal folds thickening under the influence of hormone replacement therapy [1]. Our findings confirmed this observation from Agha & Hynes [1].

C. Dynamic range and frequency range

Between March and November 2024, a significant decline in both the volume and dynamics of the singing voice can be observed (see Fig. 2). The subsequent period (November 2024 to March 2025), on the other hand, shows a noticeable improvement in both parameters.

Whether this positive trend will continue remains to be clarified by further investigations. However, the present results already demonstrate an increase in both the range and dynamics of the singing voice after an initial decline. This is an important finding, as many studies on transgender voices only collect data up to about six months after the start of hormonal transition [3], leading to premature conclusions of permanent deterioration in singing abilities [3]. These initial findings will be confirmed through further analysis.

D. Detailed analysis of VRP

A general trend is observable when comparing the VRP recorded between June 2023 and March 2025: the overall VRP first contracted and subsequently expanded downwards (see Fig. 3). From October 2023 through March 2024, the VRP progressively decreased, reaching its smallest size in March 2024. From July 2024 through March 2025, the VRP demonstrated a steady increase in size. These findings add to the overall observed trend, displaying an initial collapse of vocal range, with a subsequent expansion of the vocal range downwards and a near regaining of the baseline vocal range in a lower range.

E. Long Term Average Spectrum

The persistent presence of the singer's formant (see Fig. 4) throughout the observation period, as documented with the LTAS, represents a promising finding. The finding suggests that this aspect of vocal quality was maintained despite the hormonally induced vocal transition. The LTAS shown here was based on the analysis of glissandi.

Further LTAS analysis of longer music excerpts are planned to confirm this initial finding.

The vocal training accompanying the transition was described by the test subject as having a very positive influence on vocal hygiene. However, the extent to which the observed improvements in the dynamics and range of the singing voice can be attributed to this cannot yet be conclusively assessed.

V. Conclusion

This study offers the opportunity to provide insight into the vocal development of a classical singer who receives professional vocal training and optimal support for their functional vocal hygiene. With the help of VRP and laryngostroboscopy, paired with tight monitoring of hormone levels, we can contribute to understanding the influence of functional vocal training on the outcome of singing voices that undergo testosterone treatment.

The initial findings of this case study demonstrate a promising potential correlation between the upkeeping of functional voice training and improved outcomes of the intensity and frequency range of transgender classical singers undergoing testosterone therapy. More in-depth research is needed to further examine this potential correlation.

An extended observation range has already revealed more insight into the recovery of the voice capabilities. An even longer observation period could increase these insights. Further VRP measurements and laryngostroboscopic examinations are necessary to examine how the initial vocal development trends observed until May 2025 will continue to develop. In addition, the development of register transitions will be investigated with the help of EGG recordings, Glissandi and spectrographic analyses.

While this case study examined the vocal transition of only one transmasculine classical singer, the findings demonstrate that vocal recovery is possible following testosterone-induced voice changes, though its variability across individuals requires further study.

REFERENCES

[1] Agha, A. & Hynes, K. (2022). Exogenous Testosterone and the Transgender Singing Voice: A 30-Month Case Study. Journal of Singing, National Association of Teachers of Singing. 78(4). pp. 441-455. DOI: 10.53830/ykwm3774

[2] Azul, D. & Neuschaefer-Rube, Ch. (2019). Voice Function in Gender-Diverse People Assigned Female at Birth: Results From a Participant-Centered Mixed-Methods Study and Implications for Clinical Practice. J Speech Lang Hear Res. 62(9). Pp. 3320-3338. DOI: 10.1044/2019_JSLHR-S-19-0063

[3] Graham, F. (2022). To T or Not to T: The Transmasculine Singing Voice on Hormone Replacement Therapy. Voice and Speech Review. 16(2). 180–199. DOI: 10.1080/23268263.2022.2038349

[4] Romano, T. (2018). The singing voice during the first two years of testosterone therapy: Working with the trans or gender queer voice. Doctoral dissertation. University of Colorado Boulder

[5] Schneider-Stickler, B. & Bigenzahn W. (2013). Stimmdiagnostik, 2nd ed. Springer. pp.106-114

[6] Roubeau, B., Henrich, N., & Castellengo, M. (2009). Laryngeal vibratory mechanisms: The notion of vocal register revisited. Journal of Voice, 23(4), 425–438. DOI: 10.1016/j.jvoice.2007.10.014

"VOICE RANGE PROFILE" OR "VOICE MAP"? ON TERMS, RATIONALES AND TECHNIQUES

S. Ternström[1] and P. Pabon[2]

[1] Dept of Speech, Music and Hearing, KTH Royal Inst. of Technology, Stockholm, Sweden, stern@kth.se
[2] Voice Quality Systems, Utrecht, The Netherlands (retired)

Abstract: **Let "voice range profile" (a.k.a, "phonetogram") be the term for a graph of the maximum phonatory range of a voice on the f_o×SPL plane, i.e., a closed contour. Let "voice map" be the term for a map of a scalar metric over some relevant range, not necessarily to the extremes, on that same plane, i.e., a 2D scalar field. For imaging several metrics, one voice map can have several "layers", all derived from the same recording. This paradigm for collection and collation of voice data is useful, because it accounts for how the chosen metrics vary systematically with f_o and SPL. Both f_o and SPL are influential and typically nonlinear covariates of other voice metrics. Not accounting for them can obscure the effects of an intervention. Here we summarize some central concepts, rationales and techniques related to voice mapping.**

Keywords: **voice analysis, voice map, voice range profile**

I. WHY MAP VOICES?

Even healthy voices are notoriously variable across their ranges, and very different from person to person. In challenged voices, the diversity increases, due to various possible failures of the normal mechanism, and to the ensuing coping strategies. The vocal mechanism is such that almost all voice metrics vary considerably with f_o and SPL, and usually in a nonlinear fashion. Furthermore, in response to identical elicitations at different times, humans rarely vocalize at the same f_o and SPL [1,2]. In consequence, obtaining robust quantitative evidence for the effects of interventions such as treatment or training has proven elusive. Legacy measurements of single metrics taken from isolated vowel sounds that are sparsely and vaguely elicited (e.g., "soft-comfortable-loud") are subject to what we might call 'elicitation noise.' Small changes in SPL or f_o can in themselves result in consequential changes in the value of the observed metric. Therefore a method is needed that accounts for the two major non-linear metric covariates, f_o and SPL, on an individual basis. Fortunately, individual voices are quite consistent: repeated productions by one person at the same f_o and SPL will yield very similar results, unless the voice function of the person has changed, for whatever reason. The resulting map of the metric can be acquired pre- and post-intervention, with metric values positioned accurately in their f_o×SPL-context. This allows interventions in individuals to be meaningfully assessed, by making a *difference* map. Doing so reveals that the effects of an intervention often are not the same in different parts of the voice range, underscoring the importance of adequate sampling.

Crucially, to make a pre-post comparison, f_o and SPL must be controlled for, or the added variability can obscure the effects of the intervention [3]. This control cannot be demanded of the subject; the task is too difficult for most people, and especially for those with challenged voices. When voice-mapping, however, data are automatically arranged by f_o and SPL, so informant control is unnecessary. If some overlap can be elicited of the regions phonated pre- and post-intervention, the informant's spontaneous behaviour in f_o and SPL when reading or singing can suffice; or overlap may be achieved with additional guidance by a therapist and/or by on-screen visual feedback. If no overlap is achieved, then that outcome in itself probably constitutes the primary effect of the intervention.

II. WHAT IS A VOICE RANGE PROFILE?

A VRP is a graph of the maximum phonatory range of a voice on the f_o×SPL plane, i.e., a closed contour. Both f_o and SPL are then dependent variables that describe the range over which the informant is capable of phonating; while the independent variable is the elicitation from the clinician or researcher (e.g., "as softly as you can"). There is a large body of literature on the VRP and its application in the clinic and the voice studio, reviewed in [4].

III. WHAT IS A VOICE MAP?

Voice mapping is the computerized process of collecting signals from the voice, determining f_o and SPL, and mapping any signal features of interest onto the f_o×SPL plane [4]. When the f_o×SPL plane is

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

sampled with adequate resolution and extent, it is possible to construct reproducibly the full 2D scalar field of any given metric, which is consistent for that person, in her current vocal status. That is the voice map. In this case, f_o and SPL are seen as the *in*dependent variables, and the feature of interest is the dependent variable. In fact, as will be shown, we are usually interested in *multiple* simultaneous features, or metrics, and in how these vary together across the voice range. Examples of acoustic metrics could be the jitter, the spectrum balance (SB), or the cepstrum peak prominence (CPP). If we acquire also the electroglottogram, then EGG metrics could be for instance the EGG contact quotient (Q_{ci}) or the normalized peak EGG derivative (Q_Δ) [6]. For each metric, the result will be a sparse matrix of data with SPL (rounded to integer dB) indexing the rows, f_o (rounded to integer semitones) indexing the columns, and in each element of the matrix some feature of the distribution of the metric (typically, its mean). Each element is called a cell.

For visualization on the screen, the numerical values in the cells are translated to a colour scale. This creates an image in which each cell is a 'pixel'. Our eyesight excels at discerning regions and especially gradients in images. On voice maps, such topological features represent aspects of vocal function.

Note that the voice map contains no temporal information; rather, each cell contains an accumulated result of many intermittent visits to that particular f_o and SPL. This means that voice maps emphasize the overall condition of the vocal instrument and how it is being used, but tell nothing of what is being said or sung.

IV. TO THE EXTREMES, OR NOT?

The VRP gives information on the operating limits of the informant's voice, which often is clinically useful information, and straightforward to interpret. Phonating at the extremes is revealing, but it is also an unfamiliar task that requires a precise protocol and takes time. The voice map, on the other hand, gives information on the voice *quality* within the bounds of whatever region that is relevant, such as habitual speech or singing. If that region is smaller than the VRP, acquisition can be quicker. If the protocol for acquiring a VRP is followed using a voice-mapping system, then the contour of the mapped region ('the coastline of the island') will be equivalent to a VRP.

V. SYSTEM REQUIREMENTS

Several technical conditions must be met if voice mapping is to work well. The dynamic range of the audio chain should be at least 100 dB, which mandates recording with a low-noise microphone and a free-standing pro-quality 24-bit audio interface. Multiple display screens are helpful, especially if visual feedback is part of the procedure. A high-end personal computer is needed—a budget laptop will be too slow. A good loudspeaker and/or headphones are also needed.

Although voice maps are typically made from the acoustic signal only [7], we have found that including also the EGG signal has great explanatory power. In particular, crossing the vocal fold contacting threshold has interpretable consequences also for the acoustic metrics. An EGG device is therefore part of the recommended setup. For making voice maps, the EGG device should have an analog output that can be used in parallel with a studio quality microphone. We know of no commercial EGG device in which the built-in microphone channel has a sufficient dynamic range for making full-range voice maps.

VI. ANALYSIS CONSIDERATIONS

Looking at the different layers of a voice map, it becomes apparent that the co-variation between metrics is often high, but not complete. For instance, the acoustic crest factor tends to plays a major role in the north-west part of the phonated area, but not elsewhere. One may then ask which metrics will contribute most usefully to general voice assessment, and whether some metrics are more relevant than others, to specific voice problems. This is ongoing work. It is clear that there exists no single 'holy grail' metric that alone will suffice to characterize vocal status over the full range.

Much debate and effort has gone into trying to reach consensus on exactly how to define and compute particular voice metrics. Even seemingly straightforward metrics, such as period jitter [8] or the EGG contact quotient have given rise to very many definitions [9]. For complex metrics such as the CPP, it seems to us unlikely that full consensus on the definition and implementation can ever be reached. For instance, a sophisticated computation of the CPP can be arbitrarily complex in an off-line system, where response time is not critical; while for a real-time system, one may have to resort to approximations. Also, the instrumentation at hand and operator skills invariably introduce noise and idiosyncrasies of their own.

Now, we propose that much, if not most, of the information that we seek about the voice lies in the *topology* of the metric maps, rather than in the absolute metric values. So, if instead we move our focus of attention, from absolute values to how metrics change over the voice range, then voice *function* is emphasized, and the exact definition, calibration and computation of a metric become less crucial. In other words, the important thing is not whether the CPP (for instance) computes to 18.34 dB or 22.91 dB, but rather how it has *changed* across an intervention.

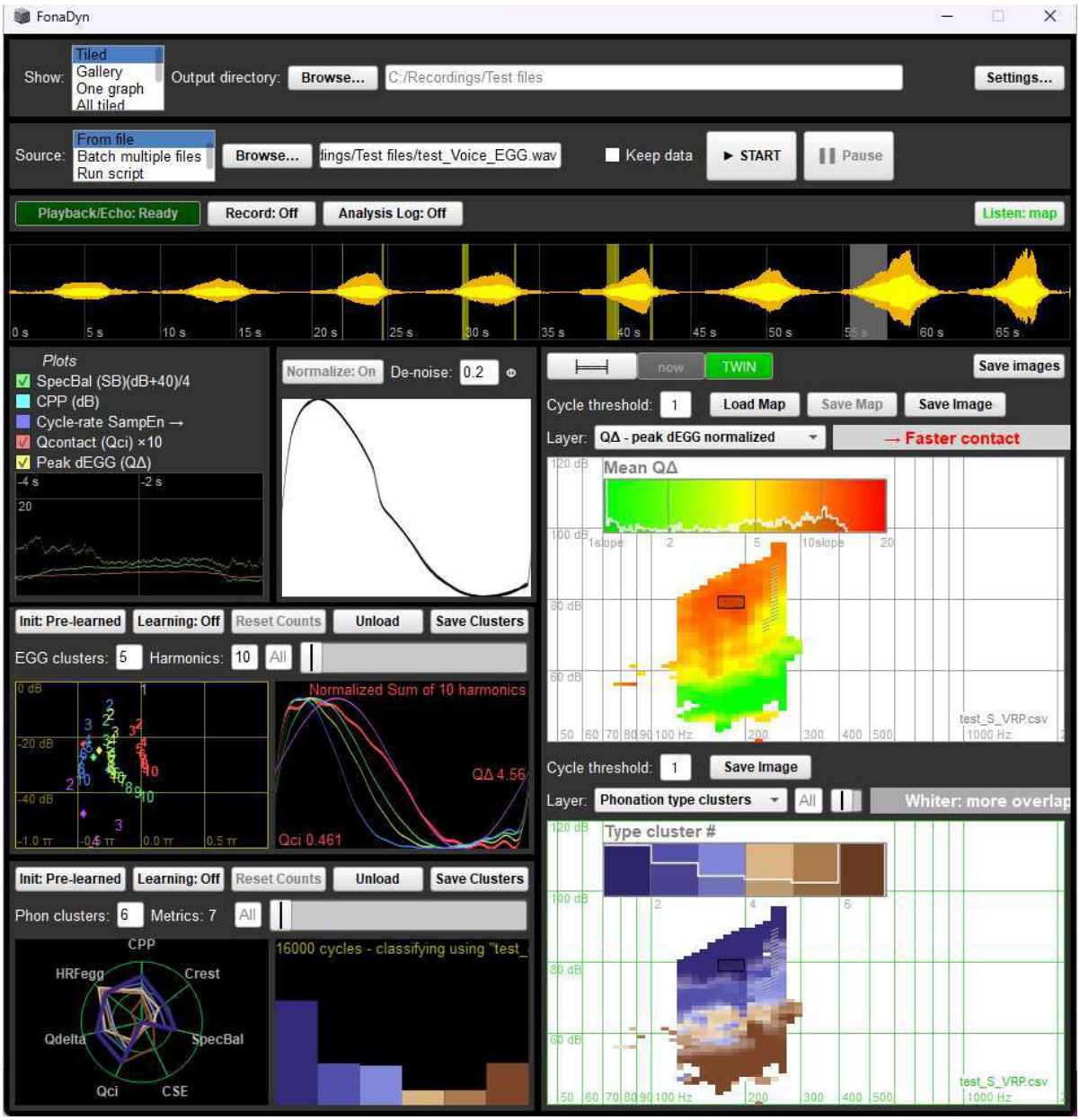


Figure 1. FonaDyn is a public-domain system for real-time voice mapping that analyzes the voice and EGG signals in parallel. It has a large repertoire of voice metrics, computes pre-post difference maps, and clusters phonation types as well as EGG wave-shapes, automatically or guided by the user. All input and result files are in standard formats, to facilitate further processing. FonaDyn is in continued development by author S.T. For more information, and a free open-source download for Windows or Mac, visit https://www.kth.se/profile/stern.

Since every cell of a voice map typically contains an average of tens or hundreds of observations, that difference can be statistically tested for every cell [5]. Where a few cells are vacant but surrounded, the map can be interpolated and smoothed, also saving acquisition time. We still require each metric to have a relationship to one or more physical entities that is systematic and unchanging—but not necessarily linear or monotonic. This leads us straight into the topic of clustering.

VII. CLUSTERING OF METRICS

For comparing VRPs across individuals, several studies have explored ways of normalizing VRP contours, by translation and scaling, e.g., to percent of full range (for an overview, see [10]). Unfortunately, such transformations have the effect of warping the axes of the map; the connection to the physics is lost. With voice maps, we can instead characterize phonation

types with combinations of metrics, obtained by statistical clustering. For example, healthy breathy soft phonation typically has a low SB, a low CPP, a Q_{ci} of 0.5 and a Q_{Δ} just over 1; while healthy pressed loud phonation has a high SB, a high CPP, a $Q_{ci} > 0.5$ and a $Q_{\Delta} > 5$. Such combinations can be found automatically, and then used to classify previously unseen signals. The resulting categories are visualized with different colours on the voice map. For mapping, it is important *not* to include f_o or SPL as clustering features, since the map itself already embodies them.

By using a common clustering scheme based on normal voices to map pathological voices, it becomes easy to interpret the vocal status of the patient, and recordings of different patients can be compared. With the FonaDyn system (Figure 1), the clinician can also interactively define custom categories that are particularly relevant to a specific patient's voice disorder [11]. This further allows the cluster maps to represent quite closely the *perceived* character of a voice.

VIII. CAN MACHINES LEARN FROM MAPS?

The current upwelling of research based on AI and machine learning, while full of promising possibilities, also carries the risk that basic knowledge of voice production will be downplayed or even ignored. It may be noted that the preprocessing of signals involved in making maps is well posed to give machine-learning models a head start, since the data is already reduced to a more compact form, and structured in a way that emphasizes the mechanisms of voice production. For instance, it seems likely that algorithms such as variational autoencoders, when applied to many maps, could uncover latent representations that correlate with the internal biomechanics, thus addressing the general inversion problem. Work along these lines has only just started. Of course, the voice map matrix representation is directly accessible to the vast toolboxes of legacy matrix algebra and image processing.

IX. CONCLUSION

We hope that others will be tempted to explore the rich world of voice mapping for their voice research. For a detailed position paper, please see [12]. While FonaDyn goes a long way, it is primarily a research tool. Hopefully, enough market pull will accumulate to resource the development and deployment of practical implementations for the clinic and the voice studio.

If you find the prospect of voice mapping interesting, you might want to join the group forum *Voice Mapping Central* [13], for more literature references, application notes and volunteered technical support, as well as other systems, and voice mappers with whom to chat.

REFERENCES

[1] W. S. Brown, T. Murry and D. Hughes. Comfortable effort level: An experimental variable. *Journal of the Acoustical Society of America*, *60*(3), pp. 696–699, 1976. https://doi.org/10.1121/1.381141

[2] W. S. Brown, R.J. Morris and T. Murry. Comfortable effort level revisited. *J Voice*, 10 (3), pp. 299–305, 1996. https://doi.org/10.1016/s0892-1997(96)80011-7

[3] N. A. Iob, L. He, S. Ternström, H. Cai, and M. Brockmann-Bauser. Effects of Speech Characteristics on Electroglottographic and Instrumental Acoustic Voice Analysis Metrics in Women With Structural Dysphonia Before and After Treatment. *J of Speech, Language, and Hearing Res.* 67(6), pp. 1660–1681, 2024. https://doi.org/10.1044/2024_JSLHR-23-00253

[4] A. Lamarche. *Putting the Singing Voice on the Map - Towards Improving the Quantitative Evaluation of Voice Status in Professional Female Singers* (Issue 2009:03). Ph.D. thesis, KTH, Stockholm.

[5] Ternström, St., Pabon, P. From Voice Signals to Voice Maps. (2025, in press) *International Journal of Voice Science*, Vol. 1, DOI: 10.2478/ijvs-2025-0002

[6] S. Ternström. Normalized time-domain parameters for electroglottographic waveforms. *Journal of the Acoustical Society of America*, 146(1), pp. EL65–EL70, 2019. https://doi.org/10.1121/1.5117174

[7] Pabon, P. and S. Ternström. Feature Maps of the Acoustic Spectrum of the Voice. *J Voice*, 34(1), 161.e1-26, https://doi.org/10.1016/j.jvoice.2018.08.014

[8] R. M. Roark. Frequency and Voice: Perspectives in the Time Domain. *J Voice*, 20(3), pp. 325–354, 2006. https://doi.org/10.1016/j.jvoice.2005.12.009

[9] C. Herbst and S. Ternström. A comparison of different methods to measure the EGG contact quotient. *Logoped. Phoniatr, Vocol.*, 31(3), 126–138, 2006. https://doi.org/10.1080/14015430500376580

[10] P. Pabon, S. Ternström, and A. Lamarche. Fourier Descriptor Analysis and Unification of Voice Range Profile Contours: Method and Applications. *J Speech, Language and Hearing Res.*, 54(3), 755–776, 2011. https://doi.org/10.1044/1092-4388(2010/08-0222)

[11] S. Capobianco, G. Björck, F. Forli, L. Bruschini, A. Nacci and S. Ternström. Voice mapping in clinical practice: tracking objective changes after injection laryngoplasty. In (C. Manfredi, ed.) *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop*, December 16-17, 2025, Firenze University Press.

[12] S. Ternström and P. Pabon. Voice Maps as a Tool for Understanding and Dealing with Variability in the Voice. *Applied Sciences* (Switzerland), 12(22), 11353, 2022. https://doi.org/10.3390/app122211353

[13] *Voice Mapping Central* user interest group, https://voicemapping.groups.io.

SESSION IV
VOICE EMOTIONS AND PERSONALITY

HARMONY IN SPEECH: MUSICAL STRUCTURE AS A MARKER OF PERSONALITY TRAITS

D. Valeeva,

Chair of Communication Science, Institute of Language & Communication, Technische Universität Berlin, Germany
valeeva@campus.tu-berlin.de

Abstract: **This paper argues that the perception of human personality through voice is deeply connected to the principles of music perception. By analyzing speech melody through the lens of harmony, melodic contour, and acoustic expression, we can uncover the specific vocal cues that signal personality traits such as neuroticism and extraversion.**

Keywords: **dissonance in speech, vocal markers of personality, acoustic features of extraversion & neuroticism**

I. INTRODUCTION

Music and language are two of the most intricate and defining forms of human communication. While their surface features may seem distinct, both rely on shared biological and cognitive infrastructures. Studies using neuroimaging and electrophysiological methods show that speech and music share common processing paths at early auditory stages [1]. Several key components where music and language research converge include pitch perception, syntax and harmonic structure, emotional expression, timbre and speaker identification, auditory attention, and rhythm [2], [3].

Although the psychoacoustic study of vocal expression has deep historical roots, with ancient Greek rhetoricians linking vocal traits to personality, empirical research into voice-personality correlations began in the 1930s, notably in England and Germany. In pre-WWII Germany, vocal evaluation influenced officer selection - warm, melodic voices were linked to empathy, while harsh, monotone voices suggested willpower [4]. Moses [5] found neurotic traits reflected in insecure intonation and unstable pitch. Fährmann [6] showed that low self-confidence manifested as quiet, high-pitched speech with erratic accentuation, while dominant personalities used faster tempo and clearer articulation.

Scherer [7] introduced a systematic, cross-cultural framework for identifying vocal personality markers. German listeners accurately judged American speakers' extraversion based on voice quality. Pitch and vocal effort emerged as key indicators: louder, nasal voices were perceived as extroverted and emotionally stable; warm, resonant voices signaled conscientiousness and stability [8].

Complementing these findings, Müller [9] examined F0 and formants. Lower F0 correlated with perceptions of competence and emotional stability. Formant manipulation enhanced perceived altruism in female voices, especially those with lower pitch. Later studies by McAleer et al. [10], Sendlmeier [11], Belin et al. [12], and Pearsell et al. [13] further confirmed the link between acoustic features and personality traits.

Experimental research into the musical structure of speech is still emerging. Raven [14] showed that emotional expression in spoken language often mirrors musical interval patterns: dissonant intervals - such as minor seconds and augmented fourths - were predominantly used to convey negative emotions like fear, anger, and sadness. In a related line of inquiry, Ploug & Niebuhr [15] found that pitch interval usage also correlates with perceived personality traits. Less charismatic speakers favored narrow, dissonant intervals (minor/major seconds, minor thirds), while more charismatic individuals used broader, consonant intervals (perfect fourths, fifths, major sixths), suggesting that melodic contour in speech may reflect underlying personality dimensions.

II. METHODS

The experiment aimed to explore the relationship between vocal expression and personality perception, specifically identifying which traits are reflected in voice and speech patterns. To this end, correlations were examined between self-reported personality dimensions and traits perceived by listeners. The assessment involved 11 native German speakers and 28 listeners, using a perception test based on the NEO-Five Factor Inventory model [16].

Three types of speech samples were analyzed: a sustained vowel [a] (6 seconds), a read passage ("The North Wind and the Sun" by Aesop), and spontaneous speech describing weekly routines (15–17 seconds each). Listener ratings were statistically evaluated with the help of SPSS software [17], and mean values for

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

perceived traits were compared with self-assessments to determine correlations.

In addition to auditory evaluation, acoustic analysis was conducted using the *PRAAT* software [18] to identify vocal markers linked to personality traits. Speech signals were also converted into musical notation using *Neuratron AudioScore Professional* [19], allowing for the identification of pitch intervals and the classification of harmonic versus disharmonic transitions.

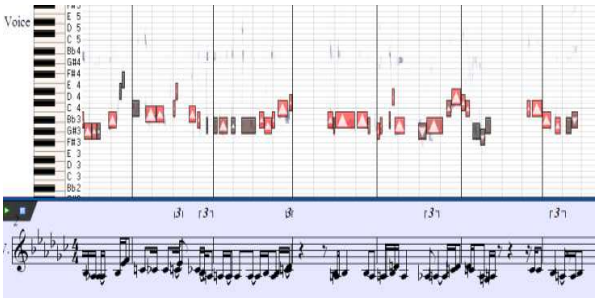


Figure 1 *- A spoken utterance converted into musical notation using Neuratron AudioScore Professional.*

III. RESULTS

Analysis of the hearing test data revealed significant correlations between self-assessed personality traits and those perceived auditorily by unfamiliar listeners. Neuroticism showed the strongest alignment across all three speech samples: sustained vowel, read text, and spontaneous speech. Extraversion demonstrated a weaker but still significant correlation, particularly in the vowel sample. No meaningful correlations were found for openness, conscientiousness, or agreeableness.

Two speakers with contrasting personality profiles were selected from the original group of 11 for detailed acoustic analysis: one perceived as highly neurotic and introverted, the other one perceived as emotionally stable and extroverted. The neurotic/introverted speaker exhibited the highest average F0 with a flat pitch contour. The stable/extroverted speaker showed greater pitch variation.

Differences in microprosodic markers were also observed: the neurotic/introverted speaker showed nearly double the values of *jitter* (subtle irregularities in the timing of vocal fold vibrations affecting pitch stability) and *shimmer* (variations in amplitude between successive cycles) compared to her counterpart, which resulted in a perceptible vocal tremor.

Concerning the Harmonics-to-Noise Ratio (HNR) - which compares the energy of the harmonic (periodic) components, generated by regular vocal fold vibrations, to the energy of the aperiodic (noise-like) components caused by turbulent airflow or irregular vocal fold behavior - the neurotic/introverted speaker exhibited the lowest HNR value.

While overall intensity levels were similar across speakers, the energy distribution differed: stable/extroverted individuals concentrated more vocal energy in higher frequency ranges.

The emphasis strategy also varied. The stable/extroverted speaker primarily employed pitch accents, whereas the introverted speaker emphasized through vowel lengthening, resulting in a slower perceived speech tempo.

In terms of vowel articulation, the stable/extroverted speaker produced decentralized vowels with clearer articulation. In contrast, the introverted speaker showed central vowel placement and reduced mouth opening, as indicated by lower F1 values.

Consonant articulation revealed further contrasts. The introverted speaker articulated consonants more precisely, with minimal assimilation and frequent devoicing (e.g., [d] → [t], [b] → [p]). The stable speaker, on the other hand, tended to voice voiceless sounds (e.g., [t] → [d]) and produced more forceful plosives with longer voice-onset times.

Speech melody analysis indicated that the neurotic/introverted speaker's speech featured dissonant intervals - particularly the minor second - and a descending melodic contour. In contrast, the stable/extroverted speaker's speech included consonant intervals, such as minor thirds and fourths, and exhibited a bell-shaped contour with greater pitch variability.

IV. DISCUSSION

The observed correlations between self-assessed and perceived personality traits suggest that vocal cues play a measurable role in personality attribution. Neuroticism emerged as the most acoustically salient trait, particularly in read and spontaneous speech, while extraversion was more detectable in vowel phonation.

The detailed comparison of two speakers highlighted how specific acoustic markers - such as elevated F0, jitter, shimmer, and reduced HNR - may signal emotional lability and introversion. These findings align with prior research linking vocal instability to emotional states like fear, boredom, and fatigue.

Articulatory differences further supported these impressions. The introverted speaker's precise consonant articulation and vowel centralization contrasted with the extroverted speaker's dynamic pitch use and decentralized vowel production, suggesting that both segmental and suprasegmental features contribute to perceived personality.

The integration of music psychology principles - particularly sensitivity to consonant/dissonant intervals

and melodic contour - offers a compelling framework for understanding how listeners intuitively assess personality traits through speech. The predominance of dissonant intervals and descending contours in emotionally unstable speech, versus consonant intervals and bell-shaped contours in stable speech, underscores the musical structure of spoken personality cues. One of the analyzed speakers was perceived as both emotionally unstable and introverted, suggesting that vocal and prosodic features may simultaneously influence the assessment of neuroticism and extraversion. To disentangle these dimensions, future studies should include individuals who score high in only one trait - such as stable but introverted profiles - to isolate relevant acoustic cues. Expanding the sample size of speakers and listeners would also enhance the representativeness of the findings.

Another consideration concerns the influence of linguistic content on spontaneous speech perception. In this study, such influence appears minimal: despite similar statements about weekend activities (e.g., "meeting friends," "going out"), participants were rated differently in terms of neuroticism and extraversion. This underscores the dominant role of vocal impression over semantic content in personality perception.

Interestingly, the speaker in question reported average scores for neuroticism and extraversion on the NEO-FFI self-assessment yet was perceived as unstable and introverted by listeners. She later admitted being under emotional distress following a breakup, which likely influenced her vocal delivery during the recording of her stimuli. The acoustic analysis focused on female voices due to the pronounced divergence in personality ratings; however, further studies involving male voices are needed to generalize findings across genders.

To tone down claims of robustness, it is important to emphasize that these findings remain preliminary. The sample size was limited, and while patterns were consistent across multiple speech types, further empirical validation is needed to generalize these results across broader populations and speech contexts.

AI-based manipulations of segmental and suprasegmental parameters - such as pitch contour adjustments targeting consonance and dissonance - could serve as a promising direction. Such controlled modifications would allow for targeted perception tests to better understand how vocal features shape personality judgments.

Beyond experimental refinement, the findings hold promise for practical applications in fields such as clinical diagnostics, human-computer interaction, and intercultural communication. In therapeutic contexts, acoustic profiling could assist in identifying emotional distress or personality-related vocal patterns that may not be explicitly reported. In voice-based AI systems, integrating personality-sensitive acoustic parameters could enhance the naturalness and adaptability of synthetic voices, tailoring them to specific user preferences or communicative goals. Moreover, in cross-cultural settings, understanding how vocal cues are interpreted across linguistic and cultural boundaries may inform training programs for professionals in diplomacy, education, or customer service. As research advances, the development of standardized acoustic-personality mapping protocols could support both empirical validation and ethical deployment of voice-based personality inference tools.

V. Conclusion

This study confirms that vocal characteristics play a significant role in how personality traits - particularly neuroticism and extraversion - are perceived by listeners. Acoustic parameters such as pitch variability, jitter, shimmer, formant distribution, and melodic contour were found to correlate with these traits, often overriding the influence of linguistic content. The presence of dissonant intervals and reduced prosodic variation aligned with perceptions of emotional instability and introversion, while consonant intervals and dynamic pitch contours supported impressions of stability and extraversion.

These findings underscore the relevance of integrating music psychology and speech acoustics in personality research. They also point to promising avenues for future studies using AI-based manipulations to further isolate and test vocal markers of personality. Controlled adjustments of segmental and suprasegmental features - such as pitch contour, spectral energy distribution, and interval structure - could enable targeted perception experiments that disentangle overlapping trait impressions and validate acoustic-personality mappings across diverse populations.

In addition to advancing theoretical understanding, such research holds potential for applied domains including clinical diagnostics, affective computing, and cross-cultural communication. However, to temper claims of robustness, it is important to emphasize that the current findings remain preliminary. The limited sample size and gender-specific focus necessitate further empirical validation to ensure generalizability across broader demographic and linguistic contexts.

References

[1] T. Shan, M.S. Cappelloni, R.K. Maddox, "Subcortical responses to music and speech are alike while cortical responses diverge," *Scientific Reports, Nature Portfolio,* ed.14:789, 2024.

[2] M. Ogg, R.L. Slevc, "Neural Mechanisms of Music and Language", in *Oxford Handbook of*

Neurolinguistics, G. Zubicaray, N. O. Schiller. Oxford, 2019, pp.906-952.
[3] B. Maess, S. Koelsch, T.C. Gunter, and A.D. Friederici, "Musical syntax is processed in Broca's area: an MEG study," *Nature Neuroscience*, vol. 4, no. 5, pp. 540–545, 2001.
[4] J. Kreiman and D. Sidtis, *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception*, UK: Wiley-Blackwell, 2011, p. 342.
[5] P. Moses, *Die Stimme der Neurose*, Stuttgart: Thieme Verlag, 1956, p. 107.
[6] R. Fährmann, *Die Deutung des Sprechausdrucks. Studien zur Einführung in die Praxis der charakterologischen Stimm- und Sprechweise*, Bonn: Bouvier Verlag, 1967, p. 74.
[7] K.R. Scherer, "Social markers in speech," Cambridge: Cambridge University Press, 1979, p. 157.
[8] K.R. Scherer, *Vokale Kommunikation: Nonverbale Aspekte des Sprachverhaltens*, Weinheim: Beltz Verlag, 1982, p. 198.
[9] R. Müller, *Stimme und Persönlichkeit*, Hildesheim/Berlin: Franzbecker Verlag, 2009, pp. 106–119.
[10] P. McAleer, A. Todorov, and P. Belin, "How do you say 'Hello'? Personality impressions from brief novel voices," *PLOS ONE*, vol. 9, no. 3, e90779, 2014.
[11] W.F. Sendlmeier, *Sprechwirkungsforschung: Grundlagen und Anwendungen mündlicher Kommunikation*, vol. 10, Berlin: Logos Verlag, 2019, pp. 82, 277.
[12] P. Belin, B. Boehme, and P. McAleer, "The sound of trustworthiness: Acoustic-based modulation of perceived voice personality," *PLOS ONE*, vol. 12, no. 10, e0185651, 2017.
[13] S. Pearsell and D. Pape, "The effects of different voice qualities on the perceived personality of a speaker," *Frontiers in Communication*, vol. 7, article 909427, 2023.
[14] H. Raven, A. Bartels, and W.F. Sendlmeier, "Gemeinsame Kommunikationsstrategien von Sprache und Musik – Musikalische Intervalle im Grundfrequenzverlauf emotionaler Äußerungen," in *Stimmlicher Ausdruck in der Alltagskommunikation*, vol. 4, W.F. Sendlmeier and A. Bartels, Eds. Berlin: Logos Verlag, 2005, pp. 109–132.
[15] M. Ploug and O. Niebuhr, "There is music in speech melody! – How pitch intervals shape speaker charisma," in *Proceedings of the 1st International Seminar on the Foundations of Speech: Pausing, Breathing and Voice*, Sonderborg, Denmark, 2019, pp. 76–68.
[16] P. Borkenau and F. Ostendorf, *NEO-Fünf-Faktoren-Inventar (NEO-FFI). Handanweisung*, Göttingen: Hogrefe, 1993.
[17] SPSS Statistics for Windows. Version 27.0, IBM Corp., 2020.
[18] P. Boersma and D. Weenink, *Praat: doing phonetics by computer*, Version 6.4.01, 1992–2023. Available: http://www.praat.org
[19] Neuratron AudioScore Professional. Version 8.9, https://download.cnet.com/audioscore-ultimate/, Neuratron Ltd., 2023.

A PRELIMINARY ANALYSIS ON LONGITUDINAL EFFECTS OF EXAM STRESS AND PERSONALITY TRAITS OVER ACOUSTIC PROPERTIES

F. Calà[1*], I. Colpizzi[2,3*], C. Sica[2], C. Caudek[4], A. Lanatà[1], L. Frassineti[1]

[1]Department of Information Engineering, Università degli Studi di Firenze, Firenze, Italy
[2]Department of Health Sciences, Università degli Studi di Firenze, Firenze, Italy
[3]Department of Life Sciences, Università degli Studi di Trieste, Trieste, Italy
[4]Department Neurofarba, Università degli Studi di Firenze, Firenze, Italy
{federico.cala, ilaria.colpizzi, claudio.sica, corrado.caudek, antonio.lanata, lorenzo.frassineti}@unifi.it
*Contributed Equally

Abstract: In recent years, the combination of physiological and psychological domains has gained increasing interest for capturing emotional and stress-related changes, especially through quantitative analysis of voice and speech. This study explores the relationship between personality traits and stress-induced vocal alterations within an Ecological Momentary Assessment (EMA) framework. Thirty-six male undergraduate students completed baseline assessments of personality traits (PID-5) and provided voice recordings before and after an academic exam. Acoustic analyses focused on sustained vowels and a standardized sentence, extracting acoustic features and mel-frequency cepstral coefficients. The results revealed significant vocal changes, including increased fundamental frequency, alterations in articulation and prosody, along with distinct acoustic patterns associated with higher scores in specific PID-5 dimensions. These preliminary findings suggest that integrating psychological and physiological data can enhance the understanding and monitoring of emotional states and stress responses, paving the way for future development of acoustic quantitative analysis within EMA paradigms as a tool to support identification of potential voice-based biomarkers for stress and psychopathology vulnerability.

Keywords: **EMA, BioVoice, MFCC, psychological traits, PID-5**

I. INTRODUCTION

Traditional evaluations of psychological states have predominantly relied on self-report questionnaires, clinical interviews, and behavioral observations. These methods provide valuable insights; however, they may suffer from some drawbacks such as low dimensionality, reporting biases and recall limitations [1]. On the other hand, physiological measurements offer objective and robust markers that, if combined with psychometric measures, can better describe emotional responses. Beyond heart rate variability and electrodermal activity, which require contact between the skin and the recording wearable device, voice and speech acquisitions are cost-effective, contactless and can be performed with ubiquitous instruments such as smartphones [2]. Speech can easily deteriorate during situations that demand challenging human performance. Thus, acoustic analysis has experienced a constantly growing interest as a supportive assessment technique for emotion and mood disruption and disorders, among which stress [3]. Specifically, it can be hypothesized that individuals characterized by dysfunctional personality traits, such as high levels of negative affectivity or disinhibition as assessed by the Personality Inventory for DSM-5 (PID-5) [4], may display a personality configuration more prone to psychopathological outcomes. However, the relationship between stable dysfunctional personality traits and stress-induced vocal alterations remains poorly understood, especially regarding their potential role as markers of vulnerability to psychopathology. Furthermore, the possibility of investigating, through quantitative analysis, how the physiological and psychological domains vary together along with their associated temporal dynamics represents a potential area for improvement in voice and speech analysis.
To address this gap, we conducted a study in which dysfunctional personality traits were measured at baseline, while negative emotional states and vocal markers were assessed before and after an ecologically valid stressor, i.e., an academic examination.

II. METHODS

Participants were undergraduate psychology students at the University of Florence, recruited through university advertisements. Participation was voluntary, and inclusion criteria required participants to be at least 18 years old, proficient in Italian, experienced with smartphones, and free from current or past psychiatric disorders or substance addictions.

As a baseline psychological evaluation (hereinafter referred to as the baseline assessment), participants completed the PID-5 questionnaire to assess dysfunctional personality traits [5]. The PID-5 is a self-report instrument consisting of 220 items organized

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

into five domains: Antagonism (A), Detachment (De), Negative Affectivity (NA), Disinhibition (Di), and Psychoticism (P). In addition, participants were trained to use the m-Path mobile application [6], a smartphone-based tool for Ecological Momentary Assessment (EMA) data collection. EMA is a methodology that captures psychological phenomena as they occur in daily life, thereby reducing recall bias [7]. Through m-Path, participants received smartphone notifications scheduled once per day on two non-consecutive days per week over a three-month period. Prompts were delivered between 6:00 and 8:00 PM and included affect ratings and brief self-report items. In addition, two exam-related prompts were administered immediately before and after an academic examination. On these occasions, participants were also asked to provide short voice recordings via their smartphones. The recording protocol included three repetitions of the sustained Italian cardinal vowels /a/, /i/ and /u/ for at least 3s, a coarticulation task consisting in listing the number from 1 to 10 and standardized constantly voiced sentence. Participants were instructed to perform such recordings with conversational pitch and loudness, in quiet rooms, keeping the smartphone's microphone at a fixed distance of 15cm from their mouth and angled at 45° to reduce lateral distortions [8]. Individuals with <50% compliance to EMA prompts were excluded. The final sample comprised 556 participants. As a preliminary analysis, this work focused on the sole male subsample, consisting of 36 subjects (mean age = 22.6 ± 2.8 years old) that completed the assessments for the baseline, pre- and post-exam conditions, and considered the sole PID-5. Indeed, identifying significant acoustic effects here will provide a stringent test of the hypothesized associations. If such effects are detectable in this group, it will be expected to find the same also in the female sample or in the general population, given their reported higher vulnerability to stress-related disorders, including those concerning voice [3].

Audio recordings were automatically partitioned in 20 segments, containing each repetition of the three cardinal vowels, each number and the standardized sentence. Only the firsts and the latter segments were considered for further analysis in this preliminary study. Sustained vowels were processed through the open-source BioVoice software [9], which extracted 37 parameters from both frequency and time domains (e.g., fundamental frequency F0, jitter, voiced unit duration). Indeed, an increase of F0 has been recognized as one of the most prominent effects of an altered psychological status [3].

The acoustic properties of the standardized sentence were quantified by the mel-frequency cepstral coefficients (MFCCs). These parameters are known to approximate the human auditory system's response [10]. Recently, they have been implemented to capture emotional activation and alterations in several experimental paradigms [10,11,12]. The MFCC extraction pipeline was developed in MATLAB 2023a (The Mathworks, Naticks, MS, USA): it computed 13 measures through a per-frame approach, specifically with 25ms long windows and 15ms overlap [10]. The resulting array was then synthesized considering the following statistical metrics: mean, standard deviation (std), median, interquartile range (IQR), skewness, kurtosis, 25th and 75th percentiles.

A first hypothesis aimed at detecting possible alterations between the baseline, pre- and post-exam conditions. Thus, each acoustic parameter underwent a repeated measure ANOVA, with the α level of significance set at 0.05. In case of significant differences, multiple comparisons were performed across groups with Tukey-Kramer correction. Moreover, a second hypothesis was tested to uncover possible interactions linking psychological traits and acoustic features. Accounting for the median value of each PID-5 subscale, data from pre- and post-exam conditions were divided into two groups characterized by lower and higher psychometric scores to perform comparisons between them ($\alpha = 0.05$).

III. RESULTS

Mean and standard deviation (in parentheses) for each of the PID-5 dimension were: A = 0.73 (0.25), De = 1.08 (0.37), Di = 1.02 (0.32), NA = 1.08 (0.39), P = 0.97 (0.39).

The repeated measures ANOVA on acoustic features highlighted four statistically significant differences for the sustained vowels between the three conditions, specifically F0 median /i/, signal duration /i/, F0 mean /u/ and F1 minimum /u/, as Figure 1 displays.

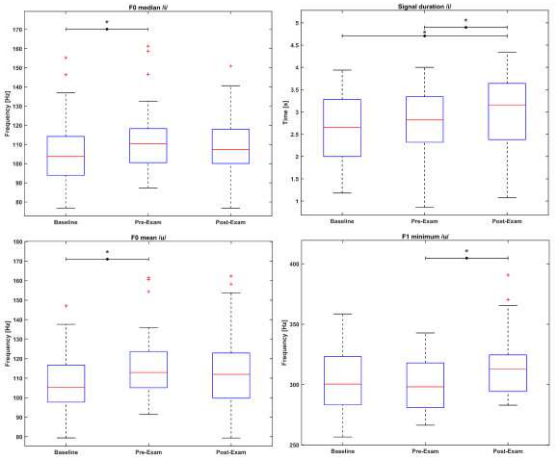


Figure 1: Boxplots for the sustained vowels' significant differences. A ● represents a p-value < 0.05, whereas a (*) to an effect size d < 0.5

As far as the standardized sentence is concerned, statistical analysis found 11 significant differences across MFCC parameters. Figure 2 shows their distribution.

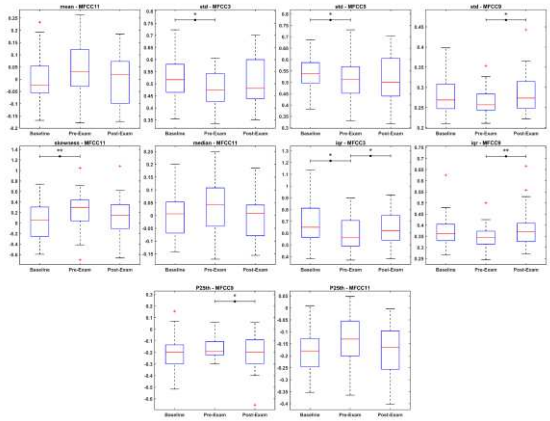


Figure 2: Boxplots for the sustained vowels' significant differences. A ● represents a p-value < 0.05. A (*) to an effect size d < 0.5, whereas (**) to d < 0.8.

The analysis between the populations with low and high values of the single PID-5 dimensions highlighted several significant differences. Here, only acoustic parameters that presented an alteration in both pre- and post-exam conditions are reported. Subjects with low De values presented a significantly higher value for T0 (F0 max) /a/, i.e., the time instance at which the F0 maximum occurs, and a lower F0 std /u/ before the exam (p_{pre} = 0.047 and p_{pre} = 0.029, respectively). After it, the same parameters preserved the same effect direction (p_{post} = 0.027 and p_{post} = 0.045, respectively). For the A dimension, MFCC3 mean was significantly higher in both pre- and post-exam conditions in subjects with higher Antagonism scores (p_{pre} = 0.030 and p_{post} = 0.027). Finally, MFCC4 median was found to be significantly higher in participants with higher values of the P dimension (p_{pre} = 0.026 and p_{post} = 0.0089).

IV. Discussion

In this study, a preliminary analysis was conducted to investigate whether, and how, an individual's acoustic properties vary not only in response to stressful events but also according to intrinsic psychological traits.

Indeed, as also reported in [1, 13], the intrinsic information of the human voice demonstrated that under specific cognitive or psychological conditions, e.g., university exams, voice and speech parameters change due to an overall increased sympathetic activity. Unlike spontaneous speech, sustained vowels and standardized sentences provide a proper repeatability framework to study stressors' influence on vocal emissions. Moreover, using additional utterances such as /i/ and /u/ allows for a deeper understanding.

In fact, this choice proved to be successful in detecting vocal alterations between the three conditions (as shown in Figure 1). The mean and median value of F0 for /u/ and /i/, respectively, were significantly higher in the stressful condition, i.e., the day before the exam, than in the baseline. This aligns with several other literature works that indicate the F0 as one of the most relevant acoustic markers of an altered psychological status [1,3,12]. It has been hypothesized that stress increases the general degree of muscular tone, also at the larynx level, causing the vocal folds to stretch and vibrate more quickly [3]. Such a hypertonia provokes an alteration in the articulation of /u/, specifically causing the vocal tract to be more closed, as a significantly lower minimum of the F1 underlines during the pre-exam period.

MFCCs provide a useful insight about the shape of the short-term spectral envelope of a vocal emission, condensing details of both low and high frequency components. Several studies successfully implemented these coefficients in emotion, and specifically stress as well, recognition from audio recordings [10,14,15]. Smaller significant standard deviations of MFCC3 and MFCC5 in the pre-exam condition than the baseline one reflect a loss of physiological variability that could be associated with constrained articulators (especially tongue and jaw), thus altered resonances. High frequency components were also significantly less variable in the stressful condition, as the boxplot for MFCC11 shows in Figure 2. Such an alteration may be related to aspiration noise between words rather than consonantal sounds due to the phonetic properties of the used sentence. Interestingly, the IQR detected for MFCC3 a difference also between pre- and post-exam conditions. This suggests a high presence of extreme observations in the studied sample, to which the IQR is less sensitive to, that could be caused by how students cope with such a stressor based on their specific stable dysfunctional personality traits. Nonetheless, the male subsample values from PID-5 questionnaires are consistent with normative data and previously reported findings in non-clinical populations of typical levels of dysfunctional personality traits [16]. By dividing the population with the PID-5 dimensions, it was discovered that some acoustic features may serve as robust biomarkers to differentiate participants with low and high scores of Detachment, Antagonism and Psychoticism in both pre- and post-exam conditions. Indeed, recent neurophysiological evidence showed that personality traits systematically modulate biological stress reactivity across multiple systems, including the hypothalamic–pituitary–adrenal axis, the autonomic nervous system, and immune responses. A

high level of Detachment marks blunted emotional reactivity: unsurprisingly, subjects belonging to this category presented an early maximum F0, reflecting a hyperfunctional phonation characterised by abrupt and unstable vocal folds' adduction. On the other hand, Psychoticism can be related to perceptual dysregulation and Antagonism to grandiosity: both facets may explain the altered motor-articulatory reflected in the higher MFCC4 and MFCC3 mean parameters [17].

V. Conclusion

Stressful events, e.g., university exams, determine changes in voice and speech properties, even in a non-clinical population, which become more evident by considering the students' stable personality traits. This outcome supports the hypotheses that voice features can represent a non-invasive technique to explore trait-dependent stress physiology, bridging dimensional models of personality and objective biomarkers of emotional reactivity. Moreover, this study proposes for the first time a long-term framework for monitoring stress-related responses. While encouraging, these results should be considered preliminary. The associations found between the physiological and psychological domains were derived from a limited cohort consisting of only male participants and did not account for the dynamic variations at the time points collected in the original study (i.e., EMA at baseline, pre-, and post-conditions). Such aspects will be integrated into future studies to better understand the complex relationship between emotional responses and the physiological characteristics, particularly regarding longitudinal studies of voice and speech. Research will also investigate the impact of demographic factors (e.g., age, gender, exam mark) and include more psychophysiological data from longer monitoring periods.

References

[1] Wang, Q., …, & Liu, X. (2025). How Anxiety State Influences Speech Parameters: A Network Analysis Study from a Real Stressed Scenario. Brain Sci, 15(3), 262.

[2] Manfredi, C., …, & DeJonckere, P. H. (2017). Smartphones offer new opportunities in clinical voice research. J Voice, *31*(1), 111-e1.

[3] Giddens, C. L., …, & Winter, A. S. (2013). Vocal indices of stress: a review. J Voice, 27(3), 390-e21.

[4] Krueger, R. …, & Skodol, A. E. (2012). Personality inventory for DSM-5. Psychiatry Res

[5] Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. Behav Res Ther, 33(3), 335–343.

[6] Mestdagh, M., ..., & Dejonckheere, E. (2023). M-path: An easy-to-use and highly tailorable platform for ecological momentary assessment and intervention in behavioral research and clinical practice. Front Digit Health, 5, 1182175.

[7] Trull, T. J., & Ebner-Priemer, U. W. (2020). Ambulatory assessment in psychopathology research: A review of recommended reporting guidelines and current practices. J Abnorm Psychol, 129(1), 56.

[8] Calà, F., Frassineti, L., ... & Zampino, G. (2023). Artificial intelligence procedure for the screening of genetic syndromes based on voice characteristics. Bioengineering, 10(12), 1375.

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

[10] Kim, S., Kwon, N., …& Bartlett, M. (2020, July). "How are you?" Estimation of anxiety, sleep quality, and mood using computational voice analysis. IEEE EMBC 2020

[11] Olah, J., … & Cummins, N. (2023). Online speech assessment of the psychotic spectrum: exploring the relationship between overlapping acoustic markers of schizotypy, depression and anxiety. Schizophr Res, 259, 11-19.

[12] König, A., …& Robert, P. (2021). Measuring stress in health professionals over the phone using automatic speech analysis during the COVID-19 pandemic: observational pilot study. J Med Internet Res, 23(4), e24191.

[13] Pisanski, K., Nowak, J., & Sorokowski, P. (2016). Individual differences in cortisol stress response predict increases in voice pitch during exam stress. Physiol Behav, 163, 234-238.

[14] Nordin, M. S., …& Azmin, N. F. M. (2022, November). Stress Detection based on TEO and MFCC speech features using Convolutional Neural Networks (CNN). IEEE ICOCO 2022.

[15] Narzary, D., & Sharma, U. Analysis of Mental Stress with Machine Learning Methods. Springer ICICC 2024.

[16] Miller, J. D., … & Lynam, D. R. (2022). Normative data for PID-5 domains, facets, and personality disorder composites from a representative sample and comparison to community and clinical samples. PD:TRT, 13(5), 536.

[17] Soliemanifar, O., Soleymanifar, A., & Afrisham, R. (2018). Relationship between personality and biological reactivity to stress: a review. Psychiatry Investig, 15(12), 1100.

PROSOVR: EMOTIONAL PROSODY ACOUSTIC ASSESSMENT IN HEALTHY AND SCHIZOPHRENIA PATIENTS THROUGH VIRTUAL REALITY ASSISTED DATA COLLECTION

A. Araujo[1], S. Sousa [1], A. C. Gaspar [1], C. Silveira[2], J. Silva [1], C. Queirós[1], M. Leite[1], S. Ferreira [1], J. Martins [3], F. Torres [3], A. Campos [3]

[1] CIR/E2S, Polytechnic of Porto, Porto, Portugal
[2] Psychiatry Department, ULSSJ, Porto, Portugal
[3] ENCONTRAR+SE—Association for the Promotion of Mental Health, Porto, Portugal

Abstract: **This study aimed to evaluate the sensitivity of ProsoVR, a virtual reality (VR) tool, in collecting speech samples for the assessment of emotional prosody in healthy adults and patients with schizophrenia. The research compared fundamental frequency (f_0) measures between 24 individuals with schizophrenia and 24 healthy individuals. The results indicated deficits in expressive emotional prosody in the schizophrenia group. In female participants with schizophrenia, mean f_0 was consistently lower, while in male participants, f_0 variability (standard deviation) was reduced in half of the emotions, suggesting more monotonous speech. The tonal range proved to be particularly sensitive in detecting differences between the groups, with lower values in the clinical group. The study concluded that ProsoVR is a promising tool for clinical practice in mental health, providing objective acoustic measurements that can identify changes in emotional expression.**

Keywords: **Emotional prosody, virtual reality, acoustic analysis**

I. INTRODUCTION

Prosody, commonly known as the melody of speech, encompasses properties like fundamental frequency, duration, and intensity, which are used to encode linguistic and paralinguistic phenomena in speech [1]. These attributes allow the speaker to make use of a high variety of subtle melodic intonations, transmitting to the listener explicit or subliminal messages [2].

Prosody is typically delineated into two principal domains: linguistic and emotional prosody. Linguistic prosody encompasses suprasegmental features that fulfill grammatical functions, including word stress, sentence focus, boundary demarcation, and intonation patterns. Conversely, emotional (or affective) prosody is regarded as a paralinguistic attribute and pertains to the conveyance of affective states through spoken language [3].

Prosody plays a fundamental role in the paralinguistic aspects of communication, making it essential for effective human interaction [4]. Consequently, individuals with various communication disorders often exhibit impairments in prosody, among other difficulties [5]. Communication disorders and clinical conditions with described prosody limitations are: dysarthria, stuttering, neurological lesions, deafness, autism, dementia, and schizophrenia [6].

According to the Diagnostic and Statistical Manual of Mental Disorders (DSM-5-TR), schizophrenia is a serious mental illness characterised by positive symptoms (like hallucinations and delusions), negative symptoms (such as apathy and emotional blunting), and cognitive deficits [7]. Prosodic deficits are considered a negative symptom of schizophrenia and can impact a person's quality of life and social interactions [8].

Emotional prosody is a key component of communication and emotion sharing and has been studied for decades [9]. When an individual has prosodic deficits, their speech may sound monotonous and emotionless, which can make it difficult for them to express or identify their own or others' emotional intentions. This can ultimately lead to a message being incorrectly conveyed or misunderstood [10].

Existing international protocols for assessing prosodic skills in adults are limited. Most rely on the perceptual judgments of clinical experts, which can introduce bias [11]. Alternatively, acoustic analysis can be used to describe prosody based on recorded speech samples, including parameters as fundamental frequency (f_0), as a descriptor of pitch, intensity and duration (including rate and pauses) [12, 13]. However, although software options exist (i.e. Prosogram) and have been tested in speech samples of schizophrenia patients [14], their use is still not widely seen in clinical settings.

Many protocols have limited ecological validity because they collect speech samples in non-natural settings, such as when people read or repeat sentences. Virtual reality (VR) has proven to be a useful

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Lorenzo Frassineti, Antonio Lanatà, Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 14th International Workshop, December, 16-17, 2025*, © 2025 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0821-5, DOI 10.36253/979-12-215-0821-5

technology in various neuroscience fields [15] and is being developed and tested for communication disorders and mental health.

ProsoVR is a VR instrument, developed by first author, to assist clinicians and researchers in collecting speech samples for emotional prosody assessment. ProsoVR includes a neutral environment and three emotion-inducing scenarios (joy, sadness, and anger), followed by guided sentence reading. The experience incorporates a dialogue including three key sentences repeated in all the scenarios. ProsoVR is the ground of a research project aimed at creating a large-scale database of speech corpora to study emotional prosody and expressivity: ProsoVR database.

This study aims to evaluate the sensitivity of ProsoVR, as a VR tool designed to collect ecologically valid speech samples for emotional prosody assessment, by comparing f_o measures between healthy individuals and patients diagnosed with schizophrenia.

II. METHODS

This quantitative observational cross-sectional study evaluated expressive emotional prosody in 48 adults. The group was composed of 24 individuals with stabilised schizophrenia (schizophrenia group - SG) and 24 healthy controls (control group - CG). Participants were at least 18 years old, fluent in Portuguese, and could read. Exclusions for both groups included self-reported significant hearing or visual impairments, diagnosed speech-language disorders, or being a voice professional. The control group also excluded individuals who self-reported mental disorders.

Each participant completed a sociodemographic questionnaire and a 30-minute session in a controlled environment. Speech was recorded using a microphone built into Meta Quest 2 glasses and analysed using Praat software. Participants read a European Portuguese phonetically balanced text, used as neutral speech sample, and then took part in the ProsoVR experiment, which included scenarios for joy, sadness, and anger, presented in a random order. The acoustic analysis focused on f_o-related parameters: mean f_o and standard deviation (f_o SD) in Hz, and tonal range in semitones (ST). Data was processed in SPSS, with statistical comparisons between groups using Student's t-test or Mann-Whitney tests, with a significance level (α) of 0.050.

III. RESULTS

Table 1. Acoustic characterization of emotional prosody in different VR contexts

Emotion / VR context	Acoustic parameter	Diagnosed with Schizophrenia A ± SD		Control Group A ± SD		p-value	
		Female	Male	Female	Male	Female	Male
Neutral	Mean f_o (Hz)	201.75 ± 18.66	122.71 ± 19.67	218.45 ± 16.65	123.45 ± 19.82	**.016***	.776 [i]
	f_o SD (Hz)	24.29 ± 9.01	11.32 ± 6,63	26.33 ± 8.09	18.31 ± 6.81	.241 [i]	**.006**** [i]
	Tonal Range (ST)	13.92 ± 4.53		16.67 ± 4.14		**.004****	
Joy	Mean f_o (Hz)	234.29 ± 39.02	147.01 ± 39.70	265.19 ± 18.19	156.65 ± 26.76	**.020***	.538
	f_o SD (Hz)	40.48 ± 19.37	20.73 ± 13.82	47.70 ± 7.24	34.61 ± 21.53	.226	**.033*** [i]
	Tonal Range (ST)	15.38 ± 5.13		21.17 ± 6.20		**< .001***** [i]	
Sadness	Mean f_o (Hz)	200.39 ± 20.83	123.11 ± 21,42	219.09 ± 14.26	119.33 ± 10.41	**.008****	.722 [i]
	f_o SD (Hz)	21.85 ± 8.40	13,81 ± 8,43	26.50 ± 10.89	16.54 ± 5.82	.241	.177 [i]
	Tonal Range (ST)	12.75 ± 5.55		13.79 ± 5.53		.563 [i]	
Anger	Mean f_o (Hz)	223.62 ± 34,03	144.62 ± 39.98	248.35 ± 16.70	141.40 ± 13.33	**.015***	.320 [i]
	f_o SD (Hz)	35.42 ± 15.24	20.00 ± 15.01	44.45 ± 7.21	24.84 ± 8.98	.074	.065 [i]
	Tonal Range (ST)	14.54 ± 4.58		19.63 ± 5.72		**< .001***** [i]	

*Caption: *p ≤ 0.050, **p ≤ 0.010, ***p ≤ 0.001; p-value obtained using the Student's t-test for two independent samples, except for values marked with [i], where the p-value was obtained using the Mann-Whitney test; SD – standard deviation; ST: Semitones.*

Participants in the SG were balanced in gender (50% female) and aged in average 41 years (SD=11,89) ranging from 23 to 63 years-old. CG had more female participants (62,5%), and was younger, with mean age of 20,75 years (SD=2,23), ranging from 18 to 25 years-old. Participants with schizophrenia were all under stable pharmacological treatment prescribed by their psychiatrists at the time of assessment. Medication mainly included atypical antipsychotics.

Acoustic analysis of f_o-related parameters is shown in Table 1. Considering that pitch in male and female voices is a physiological distinctive mark, mean f_o and f_o SD are presented by sex. In contrast, tonal range is exposed with combined sex data, as in this parameter, Hertz (Hz) measures are converted into semitones (ST), whose interval is comparable between sexes.

Concerning the mean f_o values varied from 200 Hz (sadness) to 234 Hz (joy) in SG females, and from 218 Hz (neutral) to 265 Hz (joy) in CG females. Statistical differences between groups (SG vs CG) were found for all emotions in females, mostly explained by lower values in the schizophrenia condition. As for males, they varied from 122 Hz (neutral) to 147 Hz (joy), and from 119 Hz (sadness) to 156 Hz (joy) in CG. No statistical differences were found between groups in males. Figure 1 presents these results in visuals.

Figure 1. Mean f_o variation in SG and CG

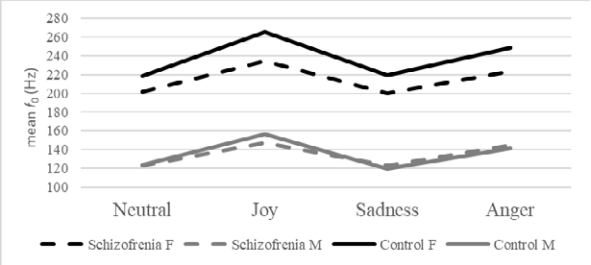
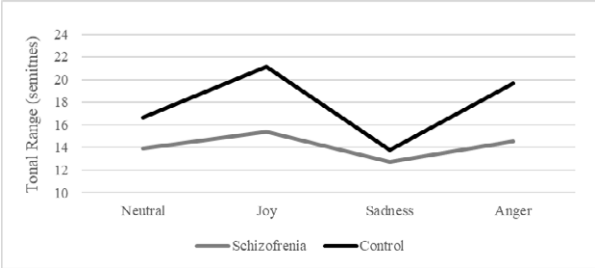


Figure 2. Tonal range variation in SG and CG



Concerning f_o SD in female groups, values varied from 21.85 Hz (sadness) to 40.48 Hz (joy) in SG, and from 26.33 Hz (neutral) to 47.70 Hz (joy) in CG. No statistical differences were found in females. As for males, values varied from 23.81 Hz (sadness) to 20.73 Hz (joy) in SC, and from 16.54 Hz to 34.61 Hz (joy) in CG. Differences were found in males between SG and CG, but only in neutral and joy samples.

Finally, tonal range varied from 12.75 ST (sadness) to 15.38 ST (joy) in SG, and from 13.79 ST (sadness) to 21.17 ST (joy) in CG. Statistical differences were found between groups in all emotions, except for sadness. These results are presented in Figure 2.

IV. Discussion

This study, the first of its kind, used a VR-assisted dialogue to analyse speech prosody in a functional, realistic context. The findings indicate that individuals with schizophrenia have deficits in expressive emotional prosody, as shown by acoustic analysis.

Acoustic differences were observed between genders and emotions:

- Female SG: In female participants with schizophrenia, the mean f_o was lower across all contexts compared to the CG, suggesting a lower vocal tone. Their tonal variation (f_o SD) was similar to the CG.
- Male SG: In male participants with schizophrenia, the average f_o was similar to the CG, but their tonal variation was consistently lower, which points to more monotonous speech, especially in neutral and joyful contexts.
- Tonal Range: Tonal range was a particularly sensitive measure, with lower values in the SG, particularly during the joy scenario. The sadness scenario showed no significant differences, which is consistent with the naturally monotonous expression of that emotion.

The results support the existing literature that links schizophrenia to specific deficits in expressive emotional prosody [14, 16], particularly for emotions that demand more expression. The study also notes that differences between sexes have not yet been fully explored and present opportunities for future research.

The observed prosodic deficits in the SG may be a result of the compromises schizophrenia makes on social cognition and higher cognitive processes, including the recognition and expression of emotions, processing speed, and reading. These changes directly affect social life, leading to social isolation and loneliness [17, 18].

The study validates ProsoVR as a valuable tool for assessing expressive emotional prosody and highlights its potential as a complementary tool in mental health practice. The use of VR is also seen as beneficial for patients, as it offers a motivating and interactive way to deliver therapy, reducing shame and frustration.

Certain demographic and clinical characteristics, such as age and medication status, may have influenced the results obtained in this study. Natural

age-related differences in vocal expression and emotional perception could partially explain the variability observed among participants. Furthermore, all participants with schizophrenia were under stable pharmacological treatment prescribed by their psychiatrists, predominantly with atypical antipsychotics. Such medication may influence emotional prosody through its effects on motor expression and affective processing, potentially contributing to the reduced expressiveness observed in the clinical group. Also, other possible selection bias could be pointed out, especially concerning non-controlled individual behaviours as tobacco consumption, which could contribute to the lower mean f_o found in SG females. Future studies should seek to minimise these potential confounding factors by including unmedicated or drug-naïve participants, ensuring age-matched samples, and adopting longitudinal or cross-sectional designs to explore how prosodic performance evolves over the course of the disorder.

V. Conclusion

Individuals with schizophrenia show deficits in expressive emotional prosody, particularly with emotions that require a high degree of expression, as joy and anger. This study found that the ProsoVR tool can provide relevant objective speech samples for acoustic measurements to confirm these changes in emotional expression in a clinical population. Basic f_o-related measures were able to provide clinically relevant information, and the differences between the schizophrenia and healthy groups were most evident in females using mean f_o and tonal range measures.

References

[1] A. Arvaniti, *The Phonetics of Prosody.* Oxford Research Encyclopedia of Linguistics, 2020.

[2] G. H. Monrad-Krohn, "The third element of speech: prosody in the neuro-psychiatric clinic", *J Ment Sci*, vol.103(431), pp. 326-331, 1957.

[3] V. Raithel, M. Hielscher-Fastabend, "Emotional and Linguistic Perception of Prosody", *Folia Phoniatr Logop*, vol. 56(1), pp. 7-13, 2004.

[4] D. Barth-Weingarten, E. Reber, M. Selting. *Prosody in Interaction*, John Benjamins Publishing Company, 2010.

[5] V. Coulombe, M. Joyal, V. Martel-Sauvageau, L. Monetta, "Affective prosody disorders in adults with neurological conditions: A scoping review", *Int J Lang Commun Disord*, vol. 58(6), pp. 1939-1954, 2023.

[6] V. Saccone, S. Trillocco, M. Moneglia, "Markers of schizophrenia at the prosody/pragmatics interface. Evidence from corpora of spontaneous speech interactions", *Front. Psychol.*, vol. *14*, pp. 1233176. 2023.

[7] American Psychiatric Association. (2022). Diagnostic and statistical manual of mental disorders (5th ed., text rev.; DSM-5-TR). American Psychiatric Publishing.

[8] V. Lucarini, M. Grice, F. Cangemi, J.T. Zimmermann, C. Marchesi, K. Vogeley, M. Tonna, "Speech Prosody as a Bridge Between Psychopathology and Linguistics: The Case of the Schizophrenia Spectrum", *Front. Psychiatry*, vol. *11*, pp. 531863, 2020.

[9] P. Larrouy-Maestri, D. Poeppel, M.D. Pell, "The sound of emotional prosody: Nearly 3 decades of research and future directions", *Perspectives on Psychological Science*, vol. 20(4), pp. 623-638, 2025.

[10] Y. Lin, H. Ding, Y. Zhang, "Emotional Prosody Processing in Schizophrenic Patients: A Selective Review and Meta-Analysis". *J. Clin. Med.*, vol. *7*(10), pp. 363, 2018.

[11] F. Fekar-Gharamaleki, N. Dardani, S.M. Khoddami, S. Jalayi, "The speech prosody tests: A narrative review", *J Res Rehabil Sci*, vol. 15(1), pp. 58-64, 2019.

[12] A. Arvaniti, "The Phonetics of Prosody", *Oxford Research Encyclopedia of Linguistics*, 2020.

[13] K. Hammerschmidt, U. Jürgens, "Acoustical correlates of affective prosody", *J. Voice*, vol. 21(5), pp. 531-540, 2007.

[14] F. Martínez-Sánchez, J. A. Muela-Martínez, P. Cortés-Soto, J. J. G. Meilán, J. A. V. Ferrándiz, A.E. Caparrós, I. M. P. Valverde, "Can the acoustic analysis of expressive prosody discriminate schizophrenia?", *Span. J. Psychol.*, vol. 18, E86, 2015.

[15] T.D. Parsons, "Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences", *Front. Hum. Neurosci.*, vol. 9, p. 660, 2015.

[16] K.M. Putnam, A.M. Kring, "Accuracy and intensity of posed emotional expressions in unmedicated schizophrenia patients: Vocal and facial channels", *Psychiatry Res.*, vol. 151(1-2), pp. 67-76, 2007.

[17] Y. Gebreegziabhere, K. Habatmu, A. Mihretu, M. Cella, A. Alem, "Cognitive impairment in people with schizophrenia: an umbrella review", *Eur Arch Psychiatry Clin Neurosci*, vol. 272(7), pp. 1139-1155, 2022.

[18] A.K.J. Fett, W. Viechtbauer, M. Dominguez de G, D.L. Penn, J. van Os, L. Krabbendam, "The relationship between neurocognition and social cognition with functional outcomes in schizophrenia: A meta-analysis", *Neurosci Biobehav Rev*, vol. 35(3), pp. 573-588, 2011.

INDEX OF AUTHORS

MAVEBA
2025
Firenze, Italy

ISSN 2704-601X (print)
ISSN 2704-5846 (online)
ISBN 979-12-215-0820-8 (Print)
ISBN 979-12-215-0821-5 (PDF)
ISBN 979-12-215-0822-2 (XML)
DOI 10.36253/979-12-215-0821-5
www.fupress.com