eVoiceNet

IVBC

International Vocal Biomarkers Conference

29-30 June 2026 - Luxembourg

ABSTRACT BOOK

cost
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

This publication is based upon work from COST Action eVoiceNet, CA24128, supported by COST (European Cooperation in Science and Technology).

COST (European Cooperation in Science and Technology) is a funding agency for research and innovation networks. Our Actions help connect research initiatives across Europe and enable scientists to grow their ideas by sharing them with their peers. This boosts their research, career and innovation.

www.cost.eu

DOI: 10.5281/zenodo.21641261

Funded by
the European Union

IVBC2026 — Book of Abstracts

International Vocal Biomarkers Conference
29–30 June 2026, Luxembourg

Edited by
Sybille Barvaux (Luxembourg Institute of Health, Luxembourg)
Ján Mucha (Scicake, Czech Republic)
Marcos Faundéz-Zanuy (Tecnocampus Mataró, Universitat Pompeu Fabra, Spain)
Dijana Petrovska-Delacrétaz (SAMOVAR, Télécom SudParis, France)
Eralp Doğu (Muğla Sıtkı Koçman University, Türkiye)

Published by
eVoiceNet — COST Action CA24128
European Cooperation in Science and Technology (COST)

Publication date
August 2026

DOI: 10.5281/zenodo.21641261
Published on Zenodo: https://doi.org/10.5281/zenodo.21641261

Suggested citation
eVoiceNet COST Action CA24128 (2026). *IVBC2026 — Book of Abstracts, International Vocal Biomarkers Conference, 29–30 June 2026, Luxembourg.*
Zenodo https://doi.org/10.5281/zenodo.21641261

This publication is based upon work from COST Action eVoiceNet, CA24128, supported by COST (European Cooperation in Science and Technology).

About the International Vocal Biomarkers Conference

The International Vocal Biomarkers Conference (IVBC) brings together researchers, clinicians, industry representatives, data scientists, engineers, and policy stakeholders working at the intersection of voice, speech, audio, artificial intelligence, and health. The conference focuses on the development, validation, and implementation of vocal biomarkers for clinical research, healthcare applications, disease monitoring, and digital health innovation.

By fostering dialogue across disciplines, IVBC aims to strengthen collaborations between academia, healthcare, industry, and technology stakeholders, accelerating the translation of vocal biomarker research into meaningful clinical and societal impact.

The first edition of IVBC, held in Luxembourg in 2026, gathered 130 participants from 35 countries and provided a unique international forum to share advances, discuss challenges, and explore future directions in the rapidly evolving field of vocal biomarkers.

eVoiceNet Chair Foreword

Guy Fagherazzi
Chair eVoiceNet & Conference Chair
Luxembourg Institute of Health (LIH)

Dear Attendees, Colleagues, and Friends,

When we launched eVoiceNet in October 2025, we wanted to turn a siloed set of promising ideas into a real, structured scientific field with impact. The first **International Vocal Biomarkers Conference**, held in Luxembourg on 29 and 30 June 2026, showed that we are well on our way.

This was our first Annual Meeting, and it went better than we hoped. We welcomed 130 participants from 35 countries, and the Scientific Committee reviewed 73 high-quality abstracts covering the whole breadth of our field of vocal biomarkers: AI and signal processing, clinical validation and standardisation, ethics and data governance, voice-based digital phenotyping, and the translational work needed to bring vocal biomarkers to patients. There is a proverb I like: if you want to go fast, go alone; if you want to go far, go together.

We have been fast. Now we have the collective strength to go far. eVoiceNet was built as a pre-competitive, open space grounded in transparency, inclusion, and scientific excellence, and our shared aim is simple: move beyond the hype and ground vocal biomarkers in solid clinical evidence.

Thank you to all of you who made this happen: the presenters and pitchers who put their work forward, the Working Group leaders and Core Group who keep this Action running, and the funders who made this event free and open to everyone. A special thanks to the team at the Luxembourg Institute of Health and our local organising committee, who turned an idea into two great days plus a Training School.

This is only the beginning. To give this community a lasting home, we are actively preparing the launch of the **International Society for Vocal Biomarkers**. More on that soon, so stay tuned.

This abstract book is the record of our first steps, and I am already looking forward to the next edition.

With gratitude,

Guy Fagherazzi, Ph.D.
Chair, eVoiceNet (COST Action CA24128)
Conference Chair, IVBC 2026

Contents

Science Communication Coordinator Foreword

Sybille Barvaux
Science Communication Coordinator & WG5 Leader
Luxembourg Institute of Health (LIH)

Dear Colleagues, Partners, and Participants,

Scientific progress does not happen in isolation. It grows through the exchange of ideas, the sharing of perspectives, and the ability to make knowledge accessible to a wider community. As Science Communication Coordinator and Dissemination Working Group Leader of eVoiceNet, I am delighted to see this first abstract book bring together the diversity, ambition, and collaborative spirit of our network.

The field of vocal biomarkers is rapidly evolving, bringing together researchers from multiple disciplines, including artificial intelligence, digital health, clinical sciences, signal processing, ethics, and data governance. Communicating advances across these different communities is essential to foster collaboration, ensure transparency, and accelerate the translation of scientific discoveries into meaningful impact.

The first International Vocal Biomarkers Conference represents an important milestone for eVoiceNet. Beyond the scientific presentations, it created a space where researchers, clinicians, early-career investigators, and stakeholders could connect, exchange ideas, and shape the future of this emerging field together.

This abstract book captures the breadth of contributions presented during this first edition and provides a snapshot of the collective efforts driving vocal biomarker research forward. It also reflects the commitment of our community to openness, collaboration, and scientific excellence.

I would like to thank all authors, reviewers, Working Group members, organisers, and participants who contributed to making this event a success. Through our shared efforts, we are not only advancing research but also building a connected and visible scientific community.

I look forward to continuing this journey together and to sharing the next milestones of eVoiceNet with an ever-growing community.

With best wishes,

Sybille Barvaux, Ph.D.
Science Communication Coordinator &
Dissemination Working Group Leader eVoiceNet (COST Action CA24128)

eVoiceNet

INTERNATIONAL VOCAL BIOMARKERS CONFERENCE 2026

Programme - Day 1

TIME	SESSION	SPEAKER(S)	TITLE OF PRESENTATION
13h30-14h	Opening session	Guy Fagherazzi	Vocal Biomarkers: Grounding Innovation in Evidence
14h-14h45	Keynote session	Nicholas Cummins	The Vocal Biomarker Gap: Promise, Evidence, and What's Missing
14h45-15h45	Oral presentations	Selected abstracts	Robust AI Methods, Generalisation & Signal Processing
15h45-16h15	Break		Coffee Break and Networking
16h15-16h45	Pitch session	Selected abstracts	Methods, Infrastructure, Ethics & Privacy
16h45-17h30	WG updates session	WG Leaders	eVoiceNet WG Updates: Progress, Challenges & Futures Directions
17h30-18h30	Oral Presentations	Selected abstracts	Neurocognitive & Mental Health Vocal Biomarkers
18h30	Drinks		Networking

cost
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

eVoiceNet

Programme - Day 2

TIME	SESSION	SPEAKER(S)	TITLE OF PRESENTATION
8h30-9h30	MC Meeting	Core Group & MC Members	Restricted Session
9h-9h30	Welcome Coffee		Welcome Coffee & Networking
9h30-10h	Keynote Session	Paula Perez	Multimodal AI for Speech-Based Health Assessment
10h-11h	Oral Presentations	Selected Abstracts	Clinical Vocal Biomarkers for Physical Health & Diseases Monitoring
11h-11h30	Break		Coffee Break & Networking
11h-30-12h30	Pitch Session	Selected Abstract	Neurocognitive, Mental Health & Vocal Biomarkers
12h30-13h30	Lunch Break		Lunch Break & Networking
13h30-14h30	Oral Presentations	Selected abstracts	Data Infrastructure, Validation & Voice Assessment
14h30-15h	Keynote Session	Jiri Mekyska	Digital Vocal Biomarkers in a Pen-and-Paper World
15h-16h	Pitch Session	Selected Abstracts	Clinical Health Applications & AI for Vocal Biomarkers
16h-16h30	Break		Coffee Break & Networking
16h30-17h	Awards & Closing	Guy Fagherazzi	Awards Ceremony & Closing Remarks

BREAKOUT ROOM

The Vocal Biomarkers Path to the European Market: Navigating Challenges, Obstacles & Opportunities

Reviewers of IVBC2026 Abstracts

The Scientific Committee gratefully acknowledges the reviewers listed below, whose careful evaluation of the submitted abstracts ensured the scientific quality of the IVBC2026 programme.

Guy Fagherazzi — Luxembourg
Luxembourg Institute of Health

Philipp Aichinger — Austria
Medical University of Vienna

Mathew Magimai Doss — Switzerland
Idiap Research Institute

Nicholas Cummins — United Kingdom
King's College London

Paula Andrea Perez-Toro — Germany
Friedrich-Alexander-Universität Erlangen-Nürnberg

Sneha Das — Denmark
Technical University of Denmark

Jiří Mekyska — Czech Republic
Brno University of Technology

Sybille Barvaux — Luxembourg
Luxembourg Institute of Health

KEYNOTE SPEAKERS

The Vocal Biomarker Gap: Promise, Evidence, and What's Missing

Nicholas Cummins
WG1 Leader
King's College London, United Kingdom

This keynote addressed "*the vocal biomarker gap*", the disconnect between promising research results and the near-total absence of vocal biomarkers in routine clinical research or practice. It broke the gap down into three main areas of action, namely the need for measurement standardisation, improved reproducibility, and stronger machine learning practices.

Speech technology for health is attracting significant interest and investment, but with this, marketing and grey literature are outpacing rigorous evidence. There is a need for the field to slow down on unsubstantiated hype and invest effort in developing core infrastructure such as protocols, terminology, and validation standards, which are urgently needed for vocal biomarkers to become trustworthy clinical tools.

True cross-disciplinary collaboration is essential, embedding end-users as co-designers from the start, rather than downstream validators.

CORE CONFERENCE TAKEAWAY

Vocal biomarkers hold real promise, but they are not a solved problem. Progress requires harmonised collection and reporting protocols, tasks and features that are rigorously tied to the clinical constructs they claim to measure, and machine learning practices robust enough to distinguish genuine clinical signal from confounds and overfitting.

Multimodal AI for Speech-Based Health Assessment

Paula Andrea Perez-Toro
WG2 Leader
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

Speech is often treated as a single acoustic signal. Yet speech production is inherently multimodal: it begins with cognition and linguistic planning, continues through motor control, respiration, and articulation, and results in acoustic and linguistic information that is interpreted through human interaction. Changes anywhere along this chain may therefore provide complementary information about neurological, psychiatric, respiratory, or structural disorders.

During the keynote, we discussed how multimodal artificial intelligence can move beyond simply combining independent data streams. Using examples from speech-based health assessment, I show how one modality can guide, reconstruct, segment, explain, or even generate another. Speech can be used to synthesize vocal-tract MRI sequences, acoustic information can improve anatomical segmentation, and imaging can provide a physiological interpretation of patterns observed in the voice. Conversely, linguistic and interactional information can complement acoustic biomarkers when assessing disorders such as mild cognitive impairment and Alzheimer's disease.

These examples illustrate a broader shift in multimodal AI: from combining modalities toward models that explicitly learn the relationships between them. Cross-modal generative models, speech-guided vision systems, and clinically informed multimodal representations offer new opportunities to work with incomplete or difficult-to-acquire data and to produce more interpretable assessments. At the same time, multimodality does not automatically guarantee better performance. The contribution of each modality depends on the clinical question, the task, the population, and the availability and quality of the data.

The next generation of speech-based health technologies will therefore depend not simply on building increasingly large unimodal models, but on understanding speech as the product of interconnected cognitive, physiological, anatomical, linguistic, and interactional processes. By explicitly modelling these relationships, multimodal AI can enable the different dimensions of speech to complement, generate, and explain one another.

CORE CONFERENCE TAKEAWAY

Speech is not merely an acoustic signal, but the observable outcome of interconnected cognitive, physiological, anatomical, linguistic, and interactional processes. Multimodal AI can exploit these connections not only by combining modalities, but also by allowing them to guide, generate, and explain one another, opening new directions for more precise and interpretable health assessment.

Digital Vocal Biomarkers in a Pen-and-Paper World

Jiří Mekyska
WG4 Leader
Scicake & Brno University of Technology, Czech Republic

Digital vocal biomarkers have shown remarkable research progress over the past decades. Machine learning models routinely achieve impressive performance, new biomarkers are published every month, and the scientific evidence continues to grow. Yet, despite this progress, only a handful of these innovations have reached routine clinical practice.

In this keynote, I share the fifteen-year journey of our research group, the Brain Diseases Analysis Laboratory at Brno University of Technology, from academic research to founding the start-up Scicake and developing the medical software platform Fonafix, which is now being deployed in major hospitals. Along the way, we made many mistakes, learned difficult lessons, and gradually realized that scientific excellence alone is not enough to create real clinical impact.

The largest gap is not between different machine learning algorithms, but between research and implementation. Bridging this gap requires addressing challenges that are often overlooked in academia, including regulatory compliance, ISO 13485 quality management, MDR certification, clinical validation, commercialization, reimbursement, sales, marketing, and user adoption. These topics are rarely discussed in research papers, yet they frequently determine whether an innovation will ever reach patients.

Perhaps our most important lesson was that researchers often optimize solutions for problems that clinicians do not consider their highest priorities. Speech-language pathologists repeatedly told us that their greatest challenge was not the lack of objective vocal biomarkers, but the overwhelming administrative burden. Before introducing advanced AI-based biomarkers, we first needed to digitize existing clinical processes, automate documentation, and save clinicians valuable time.

CORE CONFERENCE TAKEAWAY

The successful translation of digital vocal biomarkers requires much more than better algorithms. It begins with understanding clinicians' real needs and building solutions that integrate naturally into their daily workflow. Only then can digital biomarkers become a practical component of routine healthcare and clinical trials.

ORAL PRESENTATIONS

ORAL SESSION 1

Robust AI Methods, Generalisation & Signal Processing

ORAL SESSION 1 — ABSTRACT 1

ACOUSTIC CLASSIFICATION OF AMYOTROPHIC LATERAL SCLEROSIS FOR SAND CHALLENGE 2026: THE UNIVERSITY OF MARIBOR APPROACH

Gregor Donaj[1], Urh Kolarič[1], Damjan Vlaj[1], Andrej Žgank[1,*]

[1]*University of Maribor, Laboratory for Digital Signal Processing, Maribor, Slovenia*
[*]*Presented by*

Amyotrophic lateral sclerosis (ALS) belongs to a group of progressive neurodegenerative diseases that significantly affect speech. Assessing speech impairment with AI approaches thus enables monitoring the severity of ALS. The purpose is to present the University of Maribor's contribution to the SAND Challenge 2026 on the topic of ALS. We performed multi-class acoustic classification of ALS stage from speech using traditional machine learning techniques and a modern self-supervised learning framework. The experiments were performed using the VOiCe dataset, which was part of the SAND challenge. Speech recordings are labeled with the ALSFRS-R parameter, which indicates the ALS stage. Several approaches to acoustic classification were used, including hidden Markov models (HMM), support vector machines (SVM), and a self-supervised Wav2Vec 2.0 model. For acoustic features, the following approaches were applied: mel-frequency cepstral coefficients (MFCC), wavelet scattering (WS), and pitch-based features. Various approaches were tested to optimize the classification performance. The results highlighted the differences between traditional and self-supervised approaches. The baseline average F-1 score on the validation set was 0.6060. The proposed methods achieved average F-1 scores ranging from 0.2953 (pitch-based HMM) to 0.7576 (Wav2Vec) on the validation set. The results demonstrate a successful approach to acoustic classification of ALS from speech using different modeling approaches. The results from the validation and evaluation datasets indicate the need for improvements in robustness and generalization.

ORAL SESSION 1 — ABSTRACT 2

DEVELOPMENT AND VALIDATION OF AN AIRFLOW-ARTIFACTS MEASUREMENT TOOL AS A SUPPORT FOR REAL-WORLD AUDIO DATA COLLECTION CAMPAIGNS

Fulvio Cordella[1,*], Daniele Mascali[1], Silvia LoSardo[1], Luca Panzarella[1], Arianna Arienzo[1]

[1] *VoiceMed*
[*] *Presented by*

Airflow Artifacts (AAs), induced by bad positioning during respiratory maneuvers, could severely corrupt audio quality. The aim of this work was the development and validation of an AAs measurement tool for the preventive acceptance or rejection of respiratory audio recordings. A set of 10,707 sound events recorded from 1,649 subjects was manually labelled with the presence of AAs on a three-level scale: absence, moderate, and intense (example in Figure). Data were subject-wise split into training, validation, and test sets using a dedicated stratified-splitting algorithm that took into account the difference in per-subject numerosity, as each subject recorded a different number of samples. Stratification categories were chosen based on the combination of the sound identity and the AA levels. A small auto-encoder, previously trained on the training set in a self-supervised fashion, was used to extract features for the ordinal classification of the events with a two-layer multi-layer perceptron as the final classifier. On test data, the classifier reached a macro F1-score of 82.8%, with a sensitivity and specificity of 91% and 94% in recognizing events with and without AAs, and a sensitivity and specificity of 76.5% and 94.5% in discriminating moderate and severe AAs. The model consisted of a total of 1.5M parameters, with a total file footprint size of 6MB. Our model demonstrated to be a valid tool for AA measurement and for its deployment in preventive real-time checks, as an alternative to audio-restoration techniques, with the final aim of collecting less corrupted data or filtering existing datasets.

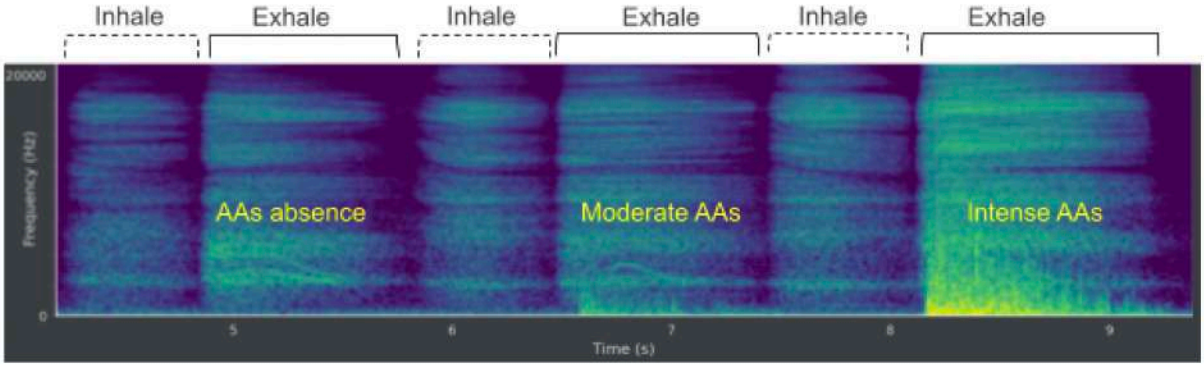


Figure: Examples of AAs intensity

ORAL SESSION 1 — ABSTRACT 3

RA-QA: BENCHMARKING HETEROGENEOUS AUDIO BIOMARKERS FOR RESPIRATORY HEALTH QUESTION ANSWERING

Gaia A. Bertolino[1,*], Yuwei Zhang[1], Tong Xia[2], Domenico Talia[3], Cecilia Mascolo[1]

[1]*University of Cambridge, United Kingdom*
[2]*Tsinghua University, China*
[3]*University of Calabria, Italy*
[*]*Presented by*

Respiratory and voice-related audio signals provide a non-invasive source of information for health assessment. Recent research has explored conversational respiratory health question answering (QA), where models process audio recordings and textual queries to produce clinically relevant answers. However, current research does not capture realistic respiratory assessment heterogeneity across modalities (speech, coughing, and breathing), devices (smartphones and electronic stethoscopes), acquisition settings (clinical vs self-recorded), question formats (verification and open-ended), and prediction targets (categorical labels and continuous outcomes). Existing studies cover only a subset of these dimensions, and no standardized benchmark enables consistent evaluation. We present RA-QA, a framework that standardizes eleven public respiratory health datasets into nine million QA pairs spanning modalities, devices, attributes, and question formats. We benchmark classical machine learning and multimodal audio-language baselines, establishing a reproducible reference setting for respiratory health QA. Our strongest benchmark baseline, based on CareAQA, shows modest but consistent performance differences across audio types. Breath achieved the highest accuracy (0.5536), followed by speech (0.5448), while cough was slightly lower (0.5295). Among speech-like recordings, counting outperformed sustained vowels (0.4118 vs 0.3769), where /o/ achieved the highest accuracy (0.3856) and /e/ was the weakest (0.3660), and normal counting was slightly stronger than fast counting (0.4183 vs 0.4052). These findings suggest that audio modality influences respiratory-health QA performance, with breath and speech achieving higher accuracy than cough, and counting outperforming sustained phonation.

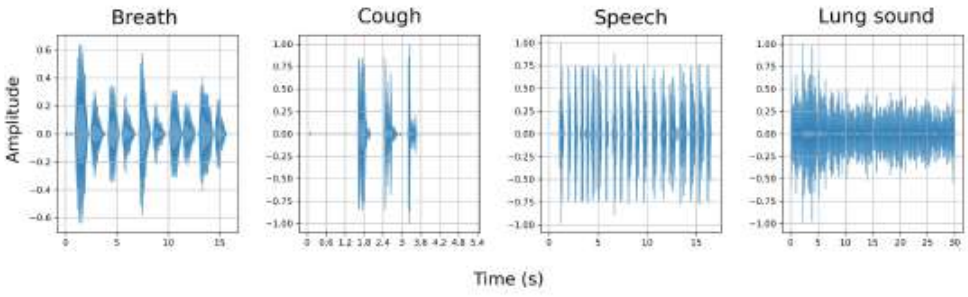


Figure: Examples of sound waves in the RA-QA collection

ORAL SESSION 1 — ABSTRACT 4

FAST INTEGER FILTERBANK FEATURES FOR EDGE SPEECH BIOMARKER EXTRACTION VIA FUSED MATRIX TRANSFORMS

Luís Pinto-Coelho[1,*]

[1]*ISEP - Polytechnic of Porto, Porto, Portugal*
[*]*Presented by*

The proliferation of speech-enabled medical devices creates strict constraints on feature extraction algorithms. Battery life, real-time latency requirements, and the frequent absence of floating-point units make conventional pipelines impractical on edge hardware. Traditional Mel-Frequency Cepstral Coefficients (MFCC) extraction is a major bottleneck in edge computing due to its reliance on floating-point FFT operations, which are poorly suited to resource-constrained hardware. This work aims to develop an integer-native, memory-efficient, and low-latency speech feature extraction method suitable for resource-constrained medical devices. We reformulate the entire MFCC pipeline as a single fused matrix-vector operation applied directly to the raw signal frame. Triangular mel filters are replaced by rectangular bandpass filters, yielding a weight matrix with Dirichlet kernel structure that enables accurate offline quantization to int8. The approach is further optimized for SIMD integer MAC units on modern edge processors (ESP32-S3). A multiplication-free variant based on Walsh-Hadamard Transform decomposition is also introduced for microcontrollers lacking hardware multipliers. For frame lengths $N \leq 512$ and output dimensionality $M \leq 24$, the proposed method requires fewer operations than fixed-point FFT while reducing feature extraction to a single memory pass. Experiments on standard speech benchmarks show that the approximate filterbank preserves classification accuracy within 92% of conventional MFCC while achieving up to 3.8$\times$ reduction in cycle count on target edge hardware. The proposed fused-matrix approach offers a practical, hardware-efficient alternative to traditional MFCC computation, enabling robust vocal biomarker extraction on ultra-low-power medical devices.

ORAL SESSION 1 — ABSTRACT 5

AGE IMBALANCE CAN BIAS VOICE-BASED COPD CLASSIFICATION: NEED FOR CONTROLLING AGE-RELATED CONFOUNDING IN VOICE-BASED BIOMARKERS

Maxime Verwoert[1,*], Loes van Bemmel[1], Sami O. Simons[1,2]

[1]*NUTRIM Institute of Nutrition and Translational Research in Metabolism, Maastricht University, Maastricht, the Netherlands*
[2]*Department of Respiratory Medicine, Maastricht University Medical Center (MUMC+), Maastricht, the Netherlands*
[*]*Presented by*

Vocal biomarkers have shown promise for detecting chronic obstructive pulmonary disease (COPD), but demographic confounders such as age may strongly influence model performance. This study investigates the extent to which age contributes to COPD detection from voice features. A crowdsourced dataset (SPEAKtoCOPD) containing 1,580 voice recordings from healthy individuals and people with respiratory diseases was analyzed. Acoustic features were extracted from sustained vowel /a/ recordings and used to train Random Forest classifier models for COPD detection. Model performance was evaluated using 10-fold cross-validation, stratified across group and age-bins. Three models were evaluated: (1) voice-only, (2) age-only, and (3) voice+age. The results were compared to a balanced age-matched subset, created using stratified sampling across 10-year age-bins and gender. Age alone achieved high classification performance (78%), exceeding the voice-based model (72%), indicating strong confounding. After age matching, performance of the voice-based model decreased substantially (66%), suggesting that part of the predictive signal was driven by age rather than disease-specific vocal characteristics. These findings highlight the importance of controlling for age in vocal biomarker development. Failure to account for demographic effects can lead to overestimation of model performance. Careful study design and evaluation are essential to ensure clinically meaningful and generalizable vocal biomarkers.

Dataset	COPD (Age / N)	Healthy (Age / N)	Model	Acc (%)	Sens (%)	Spec (%)
Original	68.5 ± 7.9 (271)	49.9 ± 14.1 (243)	Voice	72.2	72.0	72.4
			Age	78.3	87.5	69.1
			Voice+Age	78.9	80.1	77.8
Balanced	62.9 ± 7.8 (110)	61.8 ± 8.1 (110)	Voice	65.9	64.6	67.3
			Age	50.0	57.3	42.7
			Voice+Age	65.9	66.4	65.5

ORAL SESSION 1 — ABSTRACT 6

ROBUST VOCAL BIOMARKERS IN THE WILD: ADDRESSING THE IMPACT OF NOISE AND REVERBERATION

Mingchi Hou[1,2,*], Mahdi Amiri[1,2], Ina Kodrasi[1]

[1]*Idiap Research Institute, Martigny, Switzerland*
[2]*École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland*
[*]*Presented by*

Vocal biomarkers have emerged as a promising non-invasive indicator for health assessment, disease detection, and longitudinal monitoring. However, existing biomarker extraction frameworks are developed and evaluated using high quality recordings, which do not accurately reflect real-world conditions. In practical scenarios, speech signals are frequently corrupted by background noise and reverberation. These factors introduce significant distortions in the speech signal, considerably affecting the extracted vocal biomarkers. In this work, we systematically characterize the impact of noise and reverberation on commonly used acoustic and paralinguistic features. Our results show that even moderate degradations result in substantial shifts in feature distributions and decrease downstream performance in conventional machine learning models. We further show that standard pre-processing approaches such as off-the-shelf speech enhancement algorithms do not adequately address this issue. While these algorithms can improve perceptual quality, they often introduce artifacts or remove subtle pathological cues, thereby compromising the integrity of extracted biomarkers. To address these limitations, we propose an end-to-end learning framework that integrates a pathology preserving speech representation model with an enhancement objective. It jointly optimizes for environmental distortion suppression while preserving clinically relevant information. The proposed approach accounts for the mismatch between controlled training data and real-world recordings and improves generalization in the presence of noise and reverberation. Our findings highlight the importance of reconsidering traditional biomarker extraction frameworks and suggest that end-to-end methods are a critical step toward reliable vocal biomarker deployment in real-world healthcare applications.

ORAL SESSION 2

Neurocognitive & Mental Health Vocal Biomarkers

ORAL SESSION 2 — ABSTRACT 1

GENDER-SPECIFIC VOCAL BIOMARKERS FROM AN IMAGE DESCRIPTION TASK FOR SCREENING FATIGUE, SLEEP QUALITY, AND DEPRESSIVE SYMPTOMS: FINDINGS FROM THE COLIVE VOICE STUDY

Abir Elbéji[1,*], Mégane Pizzimenti[1], Guy Fagherazzi[1]

[1]*Deep Digital Phenotyping Research Unit, Department of Precision Health, Luxembourg Institute of Health, Strassen, Luxembourg*

[*]*Presented by*

Spontaneous speech features can reflect multi-level physiological and cognitive processes. Because these patterns may differ by gender, we tested separately whether feature combinations from an image description task could serve as vocal biomarkers of fatigue, poor sleep, and depressive symptoms. Colive Voice participants completed the modern Cookie Theft image description task, Fatigue Severity Scale, Pittsburgh Sleep Quality Index, and PHQ-9 (depressive symptoms). Principal component analysis was performed on 29 standardized features. Gender-stratified linear regressions, adjusted for age and education, tested associations between each component and the outcomes. Analyses included 298 US English-speaking participants (54% women ; mean age 51±9 years). The first two principal components explained 51% of the variance. PC1 reflected verbal productivity and narrative richness (more words, syllables, and content units), while PC2 captured speech tempo and efficiency (faster production, higher content-unit rate, and shorter sentence durations). Gender-stratified models showed associations only in women, where higher PC2 scores were associated with greater fatigue (β= 2.22, p = 0.03), poorer sleep quality (β= 1.35, p = 0.001), and more depressive symptoms (β = 0.71, p = 0.02). These findings show that rapid and efficient speech patterns relate to fatigue, poor sleep, and depressive symptoms in women. Such high-tempo, information-dense speech may reflect compensatory effort or heightened arousal. The lack of similar findings in men suggests gender-related variation, and further investigation in the field of voice for women's health is warranted. More diverse samples and acoustic features will strengthen the performance of such digital screening.

ORAL SESSION 2 — ABSTRACT 2

SPEECH BASED DEPRESSION AND ANXIETY ESTIMATION IN FEMALES WITH COMORBIDITY USING SELF-SUPERVISED MODELS AND LOW-RANK ADAPTATION

Asli Beşirli[1], Tuğrahan Karakadıoğlu[1], Cenk Demiroğlu[1,*]

[1]*Department of Electrical and Electronics Engineering, Ozyegin University, Istanbul, Turkey*
[*]*Presented by*

Major depressive and anxiety disorders represent significant global disability burdens, occurring more frequently in women. Their frequent co-occurrence often complicates clinical diagnosis and treatment. This study investigates how comorbidity affects speech-based digital biomarkers. We specifically aimed to evaluate the performance of state-of-the-art self-supervised learning models in estimating the severity of depression and anxiety from natural conversational speech in females. We collected a dataset from 130 Turkish women diagnosed with depression, anxiety, or comorbid depression and anxiety. Speech was recorded during clinical interviews covering both positive and negative emotional topics. We utilized three pretrained models—HuBERT, WavLM, and wav2vec2—and implemented Low-Rank Adaptation for parameter-efficient finetuning. These models were compared against traditional acoustic features, including Melfrequency cepstral coefficients and the AVEC-2013 feature set, using five distinct regression heads to predict Beck Depression Inventory and Beck Anxiety Inventory scores. The fine-tuned wav2vec2 model achieved superior accuracy, yielding a depression RMSE of 8.44 and an anxiety RMSE of 12.49. Notably, speech elicited by negative-topic questions consistently provided more predictive information than positive topics. Furthermore, depression in comorbid cases were assessed with higher accuracy (RMSE 7.34) than depressiononly cases (RMSE 8.03), suggesting that co-occurring anxiety may enhance the acoustic manifestations of depressive symptoms. Our findings demonstrate that depression estimation accuracy improves when separate models are trained for comorbid and depression-only subjects, rather than using a single unified model.

ORAL SESSION 2 — ABSTRACT 3

MIXED-REALITY HEAD–VOICE BIOMARKERS FOR NEURODEGENERATION ACROSS STRUCTURED SPEECH TASKS

Daria Hemmerling[1], Justyna Krzywdziak[1,*], Milosz Dudek[1]

[1]*AGH University of Krakow, Krakow, Poland*
[*]*Presented by*

Voice biomarkers in neurodegeneration are promising, but reproducibility remains limited by uncontrolled elicitation and sparse multimodal context. We evaluated whether mixed reality can standardise task delivery and reveal coupled vocal and head-motion signatures across clinically meaningful speech paradigms. We analysed five mixed-reality tasks in a cohort of 191 participants (92 healthy controls, 99 patients with neurodegenerative disease): **picture description**, **verbal response to a question, story recall, prolonged vowel phonation**, and **diadochokinetic syllable repetition** (/pa-ta-ka/, /pe-ta-ka/, /pa-ka-ta/). Speech activity was detected with Silero VAD, silence was removed, and openSMILE eGeMAPSv02 descriptors were summarised per task. Head pose was quantified using nine pitch/yaw/roll metrics. Group differences were assessed with two-sided Mann–Whitney U tests. Patients showed significantly greater head deviation across all tasks and axes, with the strongest and most consistent effect in roll; mean absolute roll was 2.5–4.0 times higher than in controls (all $p<0.001$). Speech occupancy was markedly lower during **picture description**, **question answering**, and **story recall** (46.9–47.2% vs 72.8–80.7% in controls; all $p<0.0001$), similar **during prolonged vowel phonation**, and higher during **diadochokinetic repetition** (40.4% vs 32.6%, $p=0.0003$). Acoustic separation was strongest in the connected-speech tasks—**picture description**, **question answering,** and **story recall**—with 103/150, 109/150, and 112/150 significant comparisons, respectively; fewer differences were observed for **prolonged vowel phonation** (30/150) and **diadochokinetic repetition** (48/150). Recurrent discriminative features included MFCC2, loudness variability, alpha ratio, spectral flux, jitter, and F0-related dispersion. Mixed reality enables standardised multimodal phenotyping that captures both motor instability and speech-production changes in neurodegeneration. A head–voice coupling biomarker derived from structured connected-speech and motor-speech tasks may support scalable screening and longitudinal monitoring.

ORAL SESSION 2 — ABSTRACT 4

SPEECH-BASED APHASIA SEVERITY DETECTION

Monica Gonzalez-Machorro[1,2,*], Uwe Reichel[1], Dorothea Peitz[2], Pascal Hecker[1], Felix Burkhardt[1], Christian Kohlschein[3], Cornelius Werner[2], Rahul Swaminathan[1], Björn Schuller[1,2]

[1]*audEERING GmbH, Germany*
[2]*CHI – Chair of Health Informatics, TUM University Hospital, Germany*
[2]*Department of Neurology, RWTH Aachen University*
[3]*Accenture GmbH, Germany*
[*]*Presented by*

Aphasia is a language disorder caused by an acquired neurological lesion. Speech and language therapy is the most effective intervention, but it requires a detailed assessment of the patient's communication abilities. In German-speaking countries, the Aachener Aphasie Test (AAT) is the standard tool for diagnosing and monitoring aphasia, evaluating all language modalities and classifying severity as mild, moderate, or severe. Most speech-based machine learning (ML) approaches predict aphasia type which supports diagnosis but is less suited for monitoring, as aphasia type changes less frequently. Our work aims to predict the severity of aphasia, offering better insight into disease progression and therapy outcomes. We analyse speech samples from 333 aphasia patients collected at University Hospital Aachen. We extracted acoustic features (voice quality, articulation, prosody) using the eGeMAPS feature set and deep embeddings using Whisper. Data were split into speaker-disjoint train dev, and models evaluated include eXtreme Gradient Boosting (XGB), Support Vector Machine (SVM), and a feed-forward neural network. The best model–SVM with Whisper embeddings–achieves 57% macro F1 (CI: 46–68%) on the test set, compared to a 33% chance level. Our findings show that speech alone holds promise for automatic aphasia severity assessment, enabling continuous monitoring rather than simple classification and supporting long-term care. While linguistic features may further improve performance, reliable ASR for aphasic speech remains challenging. Future work should examine clinical usability and model interpretability to ensure practical integration.

ORAL SESSION 2 — ABSTRACT 5

DETECTION OF COVERT HEPATIC ENCEPHALOPATHY FROM VOICE AND SPEECH USING STABILITY-DRIVEN FEATURE SELECTION AND XGBOOST

Máté Hireš[1], Peter Drotár[1,*], Jakub Gazda[1], Juan Carlos García-Pagan,[1], Sylvia Drazilova[1], Martin Janicko[1], Anna Baiges[1], Peter Jarcuska[1]

[1]*Dept. of Computers and Informatics, FEI, Technical University of Košice, Slovak Republic*
[*]*Presented by*

Covert hepatic encephalopathy (CHE) is a mild neurocognitive complication of liver disease, which often remains underdiagnosed due to the lack of simple, standardized screening methods. This study investigates the use of voice and speech features for CHE detection and aims to improve classification performance using stability-based feature selection. Voice recordings were collected from 93 Slovak-speaking participants, including 57 healthy controls and 36 CHE patients. Tasks included sustained vowels, diadochokinetic tasks (DDK), and continuous speech. A total of 246 features were derived from 29 acoustic descriptors using statistical functionals. Data preprocessing included mean imputation and standard scaling. Classification was performed using XGBoost with hyperparameter tuning. Evaluation used 10-fold stratified cross-validation, while ensuring subject-level separation. Feature importance was calculated separately for each task category and re-ranked using stability weighting across cross-validation folds. These rankings were then aggregated by averaging, and models were evaluated using the top K selected features. Baseline performance on the combined tasks achieved AUC of 81.35%. Using stability-based feature selection with K = 50, performance improved to 84.54%, with higher sensitivity (74.57%) and specificity (90.8%). This represents an improvement of more than 3% compared to the baseline without feature selection. Statistical analysis confirmed statistically significant calibration of the model predictions (p = 7.4 × 10^{-10}). The results with and without feature selection are presented in Table 1. The highest performance was observed for DDK tasks (AUC = 84.39%). Using stability-based feature selection, optimal performance was achieved with K = 50 features. Speech-derived acoustic features represent informative biomarkers for detecting CHE. Incorporating stability-based feature selection improves their consistency, further supporting the use of voice analysis as a reliable tool for CHE identification.

Table : Baseline performance (XGBoost without feature selection)

Task	Accuracy	Sensitivity	Specificity	AUC
Vowels	81.44	69.29	89.10	79.20
DDK	85.89	77.92	90.86	84.39
Speech	83.59	77.27	87.54	82.40
All	81.35	71.88	87.29	79.59
ALL (+FS)	84.54	74.57	90.80	82.68

ORAL SESSION 2 — ABSTRACT 6

CROSS-DATASET DOMAIN GENERALIZATION FOR PARKINSON DISEASE DETECTION USING FINE-TUNED WAV2VEC MMS ON SUSTAINED VOWELS

Sebastián Hreško[1,*], Peter Drotár[1]

[1]*Dept. of Computers and Informatics, FEI, Technical University of Košice, Slovak Republic*
[*]*Presented by*

Voice-based biomarkers offer a non-invasive approach for Parkinson's disease (PD) detection; however, cross-dataset variability remains a major barrier to generalization. In this study, we investigate the domain generalization capabilities of wav2vec Multilingual Massive Speech (MMS), a self-supervised foundation model trained on large-scale multilingual speech data to learn robust acoustic representations across languages and recording conditions. Compared to standard wav2vec models, MMS extends training to a massively multilingual setting, improving cross-lingual and cross-domain transfer. We employ a leave-one-dataset-out (LODA) evaluation strategy, where the model is trained on three datasets and tested on the remaining unseen dataset. Experiments are conducted on four corpora: CzechPD, Italian-PVS, PCGITA, and RMITPD, each containing multiple speech tasks. In this work, only sustained vowel recordings are used to ensure consistency across datasets. The wav2vec MMS encoder (48 transformer layers) is fine-tuned together with a classification head, updating the last six transformer layers to adapt high-level representations to the PD detection task. Feature-level aggregation is performed by computing the mean and standard deviation of frame-level embeddings from the final transformer layer, which are then used as input to the classification head. For subject-level prediction, probabilities are averaged across samples per subject, followed by thresholding at 0.5. Results demonstrate consistent cross-dataset performance, with improved robustness observed at the subject level compared to sample-level evaluation. These findings indicate that fine-tuned wav2vec MMS representations can effectively capture PD-related vocal characteristics across different recording conditions and languages, while highlighting the importance of aggregation strategies for reliable subject-level inference.

Table : Leave-one-out evaluation: Fine-tuning classification head + last 6 transformer layers

Test Dataset	UAR	F1	AUC	UAR(Subj)	F1(Subj)	AUC(Subj)
CzechPD	0.615	0.684	0.685	0.719	0.769	0.750
ItalianPVS	0.623	0.688	0.698	0.661	0.742	0.819
PCGITA	0.543	0.534	0.553	0.580	0.553	0.577
RMITPD	0.787	0.878	0.893	0.923	0.966	0.923
Average	**0.642**	**0.696**	**0.707**	**0.721**	**0.757**	**0.767**

ORAL SESSION 3

Clinical Vocal Biomarkers for Physical Health & Diseases Monitoring

ORAL SESSION 3 — ABSTRACT 1

SYSTEMATIC VOICE CHANGES IN MILD URTIS: EVIDENCE FOR VOICE-BASED BIOMARKERS

Anabell Hacker[1,2,*], Ingo Siegert[1,2]

[1]*Mobile Dialog Systems, Otto von Guericke University Magdeburg, Germany*
[2]*Intelligent Assistive Systems for Psychotherapy, University Hospital Magdeburg, Germany*
[*]*Presented by*

Upper respiratory tract infections (URTIs) induce physiological changes (e.g., vocal fold swelling) that affect speech production. However, most datasets focus on severe cases and lab recordings, lacking within-speaker variation. We present the Common Cold Corpus (CCC), an ecologically valid, longitudinal dataset comprising paired recordings from 178 participants (133 German, 45 English) in healthy and acutely ill states, collected remotely via personal devices. CCC includes read speech, spontaneous speech, and sustained vowels, enabling within-speaker comparisons under realistic conditions across varying symptom severities on the Wisconsin Upper Respiratory Symptom Survey (WURSS; mean: 3.18/7, SD = 1.65). Analyses of state-of-the-art speaker embeddings (ECAPA-TDNN, ECAPA2, TitaNet, ReDimNet) on a subset of 85 German speakers reveal that URTIs induce systematic domain shifts. Although verification performance remains high (EER 1.51–3.10%, illness increases error rates by 1–2 percentage points and significantly destabilizes representations, evidenced by substantial standardized displacement (Z_{shift} up to 8.83) and a 6.38% identity crossover rate. Directional analyses suggest that physiological changes trigger partially consistent transformations in representation space, indicating that common cold biomarkers systematically distort identity encoding. Ongoing extensions of the CCC include a large-scale perception experiment to assess whether human listeners can reliably detect illness in a forced-choice paradigm, enabling direct comparison between human and machine sensitivity. Additionally, we prepare a supplementary dataset of acted illness speech to distinguish genuine from simulated symptoms. These developments build a foundation for robust, non-invasive detection of URTI conditions and for studying the limits of voice analysis under health-related domain shifts and “fake illness” scenarios.

ORAL SESSION 3 — ABSTRACT 2

DATA AUGMENTATION FOR ROBUST VOCAL BIOMARKERS IN HEART FAILURE DECOMPENSATION

Aurèle Goetz[1,*], Valerio Mastrianni[1], Mariam Fouad[1], Leonhard Riehle[1], Manel Sabaté Tenas [2], Serhat Kabak[3], Friedrich Koehler[4], Hans-Peter Brunner-La Rocca[3], Marcus Hott[1]

[1]*Noah Labs, Berlin, Germany*
[2]*Hospital Clinic Barcelona, Cardiology, Barcelona, Spain*
[3]*Maastricht UMC+, Maastricht, Netherlands*
[4]*Deutsches Herzzentrum der Charité (DHZC), Centre for Cardiovascular Telemedicine, Berlin, Germany*
[*]*Presented by*

Vocal biomarkers offer a non-invasive, scalable approach for detecting heart failure (HF) decompensation by capturing effects of fluid accumulation, including lung congestion and laryngeal edema. However, recording conditions in remote patient monitoring vary across patients and over time due to environmental and device-related factors, introducing variability that interferes with learning meaningful disease progression patterns. Limited dataset sizes further constrain performance. This work introduces a dedicated data augmentation pipeline to improve generalization to unseen patients for early detection of HF decompensation. The proposed pipeline includes additive noise, room impulse responses, gain scaling, mild timestretching, and pitch perturbations. It was evaluated against a non-augmented baseline using two independent datasets: PRE-DETECT-HF (NCT07443969; 123 patients, 14 with usable pre-decompensation recordings as of April 2026; 1,517 recording sessions) and TIM-HF3 (DRKS00028195; 92 patients, 15 included; 402 recording sessions), differing in language, recording frequency, and acquisition conditions. Using OpenSMILE features from sustained vowels, an XGBoost model identified recordings within four weeks preceding HF hospitalization (presumed congestion) versus all other recordings. Evaluation used leave-one-patient-out cross-validation, with testing performed only on original recordings. The augmentation pipeline improved the mean patient-level ROC-AUC on individual recordings by 6.7±3.1%, 6.0±3.5% above baseline, for PRE-DETECT-HF, TIM-HF3, respectively. Improvements were consistent across settings and statistically significant (paired Wilcoxon test, Bonferroni-corrected, p=0.01). The proposed method consistently improves generalization across datasets, with gains further amplified in longitudinal monitoring via temporal aggregation of repeated predictions.

ORAL SESSION 3 — ABSTRACT 3

STENOVOICE-HICD: VOICE RECORDINGS AND DIABETES-RELATED COMPLICATIONS IN THE HEALTH IN CENTRAL DENMARK COHORT

Manuel Mounir Demetry Thomasen[1,2,*], Lars Schmidt Hansen[3], Kasper Norman[1], Guy Fagherazzi[4], Stefan Rahr Wagner[3], Adam Hulman[1,2]

[1]*Steno Diabetes Center Aarhus, Aarhus University Hospital, Aarhus N, Denmark*
[2]*Department of Public Health, Aarhus University, Aarhus, Denmark*
[3]*Department of Electrical and Computer Engineering, Aarhus University, Aarhus, Denmark*
[4]*Deep Digital Phenotyping Research Unit, Department of Precision Health, Luxembourg Institute of Health, Strassen, Luxembourg*
[*]*Presented by*

StenoVoice-HICD was established to collect voice recordings in a subcohort of Health in Central Denmark (HICD). The aim is to create a dataset for developing and validating vocal biomarkers for diabetes-related complications and disease progression. Individuals in the HICD cohort who had consented to further contact were invited in 2026 (Q1) to participate in StenoVoice-HICD by completing a 41-item questionnaire, covering self-reported diabetes status, diabetes complications, cardiovascular disease (CVD), and problem areas in diabetes (PAID-20), and a set of predefined Danish voice tasks, including /a/ phonation, counting, diadochokinesis, text reading, and an open-ended question. An open-source, web-based data-collection software was developed to collect high-quality voice recordings. Results represent the status on April 9th, 2026, and data collection is expected to end April 30th, 2026. We invited 11,460 participants, of whom 3,027 (26%) completed the questionnaire and voice recordings. Among them, 150 (4.9%) had type 1 diabetes, 985 (33%) had type 2 diabetes, and 1,892 (63%) were without diabetes. Among participants with diabetes, the prevalence of CVD was 18% and neuropathy 13%, while the median PAID-20 score was 6 (IQR: 2-19). The StenoVoice-HICD cohort is linked to nationwide registries, enabling extensive information on diagnoses, medications, socioeconomic status, and laboratory measurements. StenoVoice-HICD will serve as a novel resource for vocal biomarker research in diabetes and related complications. The cohort enables long-term follow-up, allowing for the development of predictive models to evaluate the utility of vocal biomarkers in disease risk assessment.

ORAL SESSION 3 — ABSTRACT 4

VOICE-BASED REMOTE PATIENT MONITORING IN CLINICAL TRIALS USING VOCAL BIOMARKERS: A TYPE 2 DIABETES EXAMPLE FROM THE COLIVE VOICE STUDY

Guy Fagherazzi[1], Charline Bour[1], Valeriya Sheyko[1,*], Abir Elbéji[1]

[1]*Deep Digital Phenotyping Research Unit, Department of Precision Health, Luxembourg Institute of Health, 1 A-B rue Thomas Edison, L-1445 Strassen, Luxembourg*

[*]*Presented by*

Remote patient monitoring (RPM) is emerging as a key lever to improve clinical research, particularly in chronic conditions such as type 2 diabetes (T2D), but scalable, passive, non-invasive tools remain limited. The human voice, accessible via smartphones, reflects physiological and psychological states and can be analysed to extract digital biomarkers for clinical endpoints. Colive Voice is an international clinical study conducted by the Luxembourg Institute of Health (LIH), involving 10,000 participants across fi ve continents. Participants completed health questionnaires and provided voice recordings (standardised and non-standardised tasks) reflecting their health perceptions. Audio recordings were harmonised and cleaned using proprietary LIH technology. Acoustic features were extracted and analysed using statistical and machine learning methods to develop voice-based endpoints. The picture description task was completed by 499 individuals with T2D (F=255/M=144), with an 85% usability rate. A total of 37 vocal features in men and 17 in women were significantly associated (p<0.05) with stress (PSS), fatigue (FSS), diabetes-related distress (PAID), cognitive decline, fear of hypoglycemia (HFS-II), and sleep quality (PSQI). A common vocal signature of psychological distress across all scales was identified. Voice-based RPM enables real-world, remote monitoring of psychological complications and disease-burden-related risks in T2D, providing clinically relevant insights between visits. This scalable approach supports more granular and responsive clinical trial endpoints and is transferable across therapeutic areas.

ORAL SESSION 3 — ABSTRACT 5

VOICE CHANGES DURING EXACERBATIONS OF COPD AND ASTHMA: FINDINGS OF THE TACTICAS STUDY

Loes van Bemmel[1,*], Lauren G Reinders[1], Marieke Spreeuwenberg[3], Hanneke van Helvoort[4], Visara Urovi[5], Julia Hoxha[6], Sami O. Simons[1,2]

[1]*NUTRIM Institute of Nutrition and Translational Research in Metabolism, Maastricht University, Maastricht, the Netherlands*
[2]*Department of Respiratory Medicine, Maastricht University Medical Center+, Maastricht, the Netherlands*
[3]*CAPHRI Care and Public Health Institute, Maastricht University, Maastricht, the Netherlands*
[4]*Radboud University Medical Center Nijmegen, Nijmegen, the Netherlands*
[5]*Institute of Data Science, Maastricht University, Maastricht, the Netherlands*
[6]*Zana Technologies GmbH, Karlsruhe, Germany*
[*]*Presented by*

Vocal biomarkers are a novel, unintrusive and promising way to potentially detect disease worsening. Exacerbations of Chronic Obstructive Pulmonary Disease (COPD) and asthma are periods of sharp symptom increase, likely exhibiting voice changes. In the TACTICAS study, we aimed to identify vocal biomarkers for exacerbations. For 12 weeks, 73 participants with COPD or asthma recorded daily voice and symptoms using the TACTICAS app and the EXACT® questionnaire at home. A multilevel analysis was used to identify significant fixed effects in 34 Praat features extracted from sustained vowel /a/ on the relevant days: the first (onset), worst (peak) and recovery days of exacerbation compared to stable baseline. 38 exacerbations were captured during the study. Several voice features had significant fixed effects, showing change in voice over the course of an exacerbation compared to baseline. Notably, 10 effects of features relating to tone (pitch, periods) and voice quality (jitter) were significant already at onset of an exacerbation. Voice changes during the course of an exacerbation, showing the potential of vocal biomarkers for disease monitoring at home. Early detection of exacerbations at onset of symptoms could accelerate medication initiation and enhance patient outcomes.

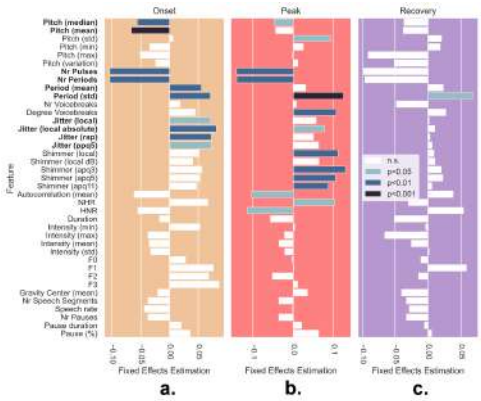


Figure: Fixed Effects Estimations of three exacerbation timepoints for speech features from sustained vowel /a/. Significant features at onset indicated in bold.

ORAL SESSION 4

Data Infrastructure, Validation & Voice Assessment

ORAL SESSION 4 — ABSTRACT 1

EXPLAINABLE SPEECH-BASED DIGITAL PHENOTYPING FOR SIBILANT MISPRONUNCIATION SCREENING IN POLISH-SPEAKING CHILDREN

Milosz Dudek[1,2], Daria Hemmerling[1,2,*], Kamil Kwarciak[1,2], Maria Pensko[2], Mateusz Kowalewski[2], Leonid Pavlovskyi[2], Sebastian Jurczak[2], Anna-Mariia Vitkovska[2], Natalia Mocko[3], Zuzanna Miodonska[3], Michal Krecichwost[4]

[1]*AGH University of Krakow, Cracow, Poland*
[2]*SoftServe, Cracow, Poland*
[3]*Department of Biomedical Engineering, Silesian University of Technology, Gliwice, Poland*
[4]*Institute of Linguistics, University of Silesia in Katowice, Katowice, Poland*
[*]*Presented by*

Early identification of speech sound errors in children is often limited by access to specialists. We present an explainable AI-based screening pipeline for Polish-speaking children that uses short prompted recordings to detect likely sibilant mispronunciations. Although intended for screening rather than diagnosis, the approach is relevant to voice-based digital phenotyping and clinically deployable speech AI. We used a proprietary corpus of 201 children aged 4–8 years, comprising 12,830 prompted word/logotome segments and 53,951 phoneme segments. The system combines a Polish wav2vec2 encoder, a 6-layer Transformer post-encoder, and a CTC head extended with bracketed substitution tokens. Recognized token sequences are aligned to canonical prompts to derive interpretable error types and confidence signals, which are translated into caregiver-facing feedback by a template-grounded assistant with explicit safety rules. On a speaker-disjoint test set of 10 unseen children and 559 utterances, the recognizer achieved 88.7% exact sequence match, 95.0% token accuracy, 5.95% WER, and 4.09% CER. A conservative screening rule based on bracketed substitution evidence reached 72.9% precision, 61.4% recall, F1=0.67, and a 2.7% false-alarm rate. When a mismatch was flagged, the predicted bracket class matched the reference in 85.7% of true-positive cases. Explainable token-level modeling from short prompted speech can support low-burden pediatric speech screening outside specialist settings. This work contributes a clinically aligned AI and audio-processing framework for voice-based phenotyping, while highlighting the need for larger-scale validation, caregiver studies, and clinician-in-the-loop deployment. This work was supported by the National Centre for Research and Development (NCBR), Poland, under research project No. 0179/L-15/2024, entitled Speech therapy computer system for recording and analyzing multimodal articulation data using the 4D speaker model and deep learning (SpeechCAD).

ORAL SESSION 4 — ABSTRACT 2

CLINICAL AND ACOUSTIC CHARACTERISTICS OF LARYNGEAL DYSTONIA, DYSTONIC TREMOR AND VOCAL TREMOR

Eleftheria Iliadou[1,2,*], Tania Rose[1], Evangelos Anagnostou[1], Jan Hlavnicka[3,4], Elina Tripoliti[1,5]

[1]*School of Health Sciences, School of Medicine, National and Kapodistrian University of Athens, Athens, Greece*
[2]*Department of Speech and Language Therapy, University of Peloponnese, Kalamata, Greece*
[3]*Department of Neurology and Center of Clinical Neuroscience, First Faculty of Medicine, Charles University and General University Hospital in Prague, Prague, Czech Republic*
[4]*Department of Circuit Theory, Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic*
[5]*Institute of Neurology, Department of Clinical and Movement Neurosciences, and National Hospital for Neurology and Neurosurgery, UCLH NHS Trust, London, UK*
[*]*Presented by*

Voice breaks are a well-recognized clinical phenomenon in laryngology and may be associated with laryngeal dystonia (LD), vocal tremor (VT), or dystonic tremor (DT), often posing a diagnostic challenge. Our objective was to assess and compare the clinical, endoscopic and acoustic characteristics of LD, VT and DT, in order to identify distinguishing voice features that may support biomarker development and validation. This is a prospective, consecutive study of patients diagnosed with LD, VT or DT evaluated at the Neurolaryngology Clinic of two Athens University Hospitals over 18 months. Demographic data, clinical, endoscopic and perceptual voice characteristics were recorded, and all patients underwent standardized acoustic voice analysis. A subgroup of patients who consented to treatment with Botulinum toxin type A (BTA) was re-evaluated one month after injection in order to assess post-treatment clinical, perceptual, and acoustic changes. Results: Thirteen patients were included (8 female), with a median age of 69 years (IQR 65–73). Eight patients were diagnosed with LD, 3 with VT, and 2 with DT. The severity of dysphonia ranged from Grade 1 to 3, with Strain and Instability representing the most prominent perceptual voice characteristics. Endoscopic and acoustic data, such as tremor frequency and Voice Onset Time, were correlated to identify which characteristics best corresponded to the diagnosis. LD, VT, and DT represent three distinct clinical entities with overlapping perceptual features. Detailed acoustic analysis may contribute to a better understanding of their underlying voice characteristics and support the identification of objective vocal biomarkers for improved differential diagnosis, phenotyping, and treatment monitoring.

ORAL SESSION 4 — ABSTRACT 3

VOCAL BIOMARKER DATA ANALYSIS PORTAL: DESIGN AND PRELIMINARY RESULTS

Cagla Nur Sen[1], Eralp Dogu[1,*]

[1]*Mugla Sitki Kocman University, Department of Statistics, Mugla, Türkiye*
[*]*Presented by*

The clinical integration of vocal biomarkers remains limited by high inter-subject variability, absence of patient-level statistical modeling, and the lack of open-source tools capable of real-time longitudinal tracking. Existing approaches fail to account for individual baseline differences, reducing their clinical reliability across heterogeneous cohorts. This study aimed to develop and validate an interactive digital health portal employing Linear Mixed-Effects Models (LMM) to monitor vocal biomarker progression in neurodegenerative disorders, with Parkinson's disease as a primary use case. An open-source analysis portal was developed using the R Shiny framework, integrating in real time with the UCI Machine Learning Repository (Parkinson's Dataset). LMMs were implemented via the 'lme4' package to address the hierarchical structure of repeated-measures vocal data: individual patients were modeled as random effects to control for baseline variability, while disease stage and visit were fixed effects. Acoustic features extracted included Jitter (%), Shimmer (dB), and Mel-Frequency Cepstral Coefficients (MFCCs) as markers of pathological vocal deviation. Z-score normalization was applied to reduce hardware-related variance across simulated smartphone recordings. Preliminary analyses demonstrated that vocal degradation slopes in the Parkinson's group significantly diverged from healthy controls ($p < 0.05$), with LMM fixed-effect estimates revealing disease stage as the primary driver of acoustic deterioration. The portal's dynamic reporting module translated model coefficients into interpretable clinical progression curves, supporting non-statistician end users. Cross-device signal consistency was confirmed following Z-score normalization. The proposed framework demonstrates that vocal biomarkers meet the criteria for clinical explainability and statistical reliability when modeled within a patient-level mixed-effects architecture — a combination not previously implemented in an open-source, real-time portal. This scalable tool directly supports eVoiceNet's mission by establishing a replicable methodology for harmonizing vocal data across multi-institutional trials, with clear potential for extension to federated learning and regulatory-grade biomarker validation.

ORAL SESSION 4 — ABSTRACT 4

BUILDING PERCEPTUAL GROUND TRUTH FOR VOCAL BIOMARKERS: A SCALABLE INFRASTRUCTURE FOR STANDARDISED AND REPRODUCIBLE AUDITORY-PERCEPTUAL VOICE DATA

Neus Calaf[1,*]

[1]*Biophysics Unit, School of Medicine, Autonomous University of Barcelona, Barcelona, Spain*
[*]*Presented by*

Vocal biomarker research has advanced through artificial intelligence and signal processing; however, progress remains limited by the lack of large-scale, standardised auditory-perceptual datasets required for reproducible and generalisable model development and validation. Auditory-perceptual voice evaluation is central to clinical and research practice, yet its variability and context dependence—particularly across languages, cultures, and expertise levels—remain insufficiently characterised. This work proposes a digital infrastructure to systematically capture and analyse perceptual voice data at scale. All-Voiced is a web-based platform designed to support auditory-perceptual voice evaluation through standardised tasks (e.g., Consensus Auditory-Perceptual Evaluation of Voice, CAPE-V), structured response capture, and integrated feedback mechanisms. Perceptual ratings across predefined attributes (e.g., overall severity, roughness, breathiness, strain) are collected alongside user and contextual metadata, enabling systematic and comparable data acquisition across users and settings. Aggregation approaches (e.g., median-based consensus) support modelling of agreement, variability, and calibration. The proposed infrastructure enables scalable and standardised generation of structured perceptual datasets while supporting training and calibration within the same system. By capturing ratings across diverse participants, it allows systematic investigation of perceptual variability, including cross-cultural and cross-linguistic differences in voice evaluation. Auditory-perceptual data constitute a critical yet underdeveloped component of the vocal biomarker pipeline. Infrastructure-based approaches can support standardisation, enhance reproducibility, and improve the generalisability and validation of vocal biomarker models.

ORAL SESSION 4 — ABSTRACT 5

EXTENDING AVQI-BASED VOICE ASSESSMENT TO REMOTE HOME PHONE CALLS IN VOICE DISORDER CARE

Saska Tirronen[1,*], Farhad Javanmardi[1], Mikhail Silaev[1], Carlos Patino[1], Manila Kodali[1], Mikko Petäjä[1], Paavo Alku[1]

[1]*Department of Information and Communications Engineering, Aalto University, Espoo, Finland*
[*]*Presented by*

The Acoustic Voice Quality Index (AVQI) is a widely used clinical measure of voice quality. However, its use outside controlled recording conditions is limited, reducing reliability across environments. This pilot aimed to extend AVQI-based assessment to voice samples recorded by phone in patients' homes. The goal was to support partial transfer of follow-up from clinics to remote settings, enabling more frequent measurements with lower resource needs. In collaboration with two Finnish voice clinics, the system was piloted with patients receiving care for voice disorders of various etiologies. Voice samples were collected during phone calls in the home environment. A proprietary machine learning algorithm adapted the recorded signals to a baseline target environment suitable for AVQI analysis. Technical performance was assessed as mean absolute error (MAE) between laboratory and remote telephony environments, for both AVQI and within-patient AVQI changes (delta AVQI). Patient and clinician experiences were evaluated with questionnaires. The adaptation reduced AVQI MAE between laboratory and remote telephony environments from 1.737 to 0.523. In addition, the MAE of within-patient AVQI changes decreased from 0.632 to 0.366. Questionnaire responses obtained so far were encouraging, although data collection is ongoing. Mean ratings across key usability and usefulness items were 4.3/5 among patients (n=10) and 4.3/5 among clinicians (n=4), supporting feasibility in clinical collaboration. This pilot provides preliminary evidence that AVQI-based voice assessment can be conducted from voice samples recorded remotely by phone in patients' homes, supporting more flexible and scalable follow-up in voice disorder care.

ORAL SESSION 4 — ABSTRACT 6

COMPARISON OF THE MULTIPARAMETRIC ACOUSTIC VOICE INDICES IN DIFFERENTIATING BETWEEN NORMAL AND DYSPHONIC VOICES

Virgilijus Ulozas[1,*], Kipras Pribuišis[1], Nora Ulozaite-Staniene[1], Tadas Petrauskas[1]

[1]*Department of Otorhinolaryngology, Lithuanian University of Health Sciences, Kaunas, Lithuania*
[*]*Presented by*

The study aimed to investigate and compare the accuracy and robustness of the multi-para metric acoustic voice indices (MAVIs), namely the Dysphonia Severity Index (DSI), Acoustic Voice Quality Index (AVQI), Acoustic Breathiness Index (ABI), and Voice Wellness Index (VWI) measures in differentiating normal and dysphonic voices. The VWI application combining the AVQI and Glottal Function Index GFI data was developed. The study group consisted of 129 adult individuals including 49 with normal voices and 80 patients with pathological voices. The diagnostic accuracy of the investigated MAVI in differentiating between normal and pathological voices was assessed using receiver operating characteristics (ROC). Moderate to strong positive linear correlations were observed between different MAVIs. The ROC statistical analysis revealed that all used measurements manifested in a high level of accuracy (area under the curve (AUC) of 0.80 and greater) and an ac ceptable level of sensitivity and specificity in discriminating between normal and pathological voices. However, with AUC0.99, the VWI demonstrated the highest diagnostic accuracy. The highest Youden index equaled 0.93, revealing that a VWI cut-off of 4.45 corresponds with highly acceptable sensitivity (97.50%) and specificity (95.92%). All MAVIs used in this study, namely the DSI, AVQI, ABI, and VWI, displayed good accuracy in distinguishing between normal and dysphonic voices. However, the VWI was found to be beneficial in describing differences in voice quality status and discriminating between normal and dysphonic voices based on clinical diagnosis, i.e., dysphonia type, implying the VWI's reliable voice screening potential.


PITCH
PRESENTATIONS

PITCH SESSION 1

Methods, Infrastructure, Ethics & Privacy

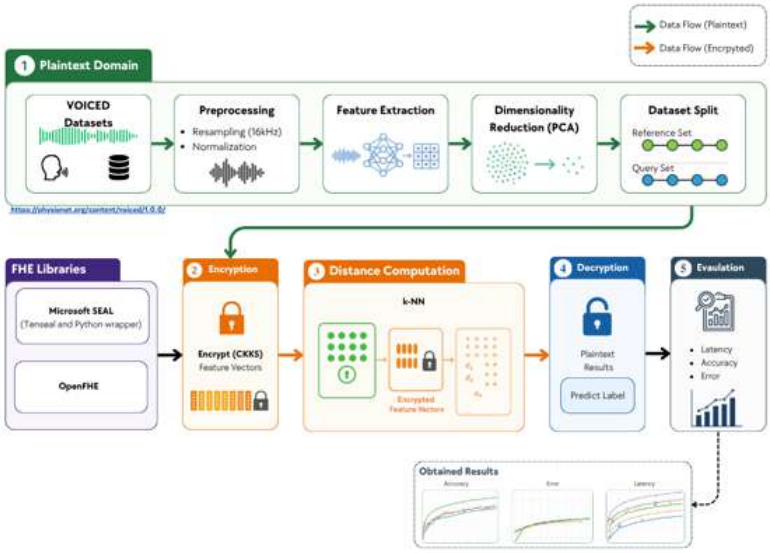
PITCH SESSION 1 — ABSTRACT 1

PRIVACY AND SECURITY IN VOICE-BASED HEALTH DATA PROCESSING USING FULLY HOMOMORPHIC ENCRYPTION (FHE)

Ahmet Faruk Dursun[1], Kübra Seyhan[1], Sedat Akleylek[2,*]

[1]*Department of Computer Engineering, Ondokuz Mayıs University, Samsun, Türkiye*
[2]*Chair of Security and Theoretical Computer Science, University of Tartu, Tartu, Estonia*
[*]*Presented by*

Secure processing of sensitive voice-based biometric data in third-party systems and the prevention of unauthorized access are crucial. In this paper, a security-processing framework (presented below) is constructed that combines fully homomorphic encryption (FHE) and machine learning (ML) to secure voice biomarker data. The voice signals in the VOICED dataset are processed at a sampling frequency of 16 kHz, with each sample converted into a 3-second segment containing 48,000 samples. High-dimensional vectors extracted from the voice signals are reduced to lower dimensions using principal component analysis (PCA) to reduce computational cost. Since the feature vectors contain decimal values, the CKKS FHE scheme is used for encryption and decryption. Classification of the encrypted data is performed using k-nearest neighbours (KNN), and results obtained with different FHE libraries are compared using several metrics. A framework for processing and encrypting voice biomarker data was constructed using ML and FHE techniques. CKKS encryption of VOICED data with Microsoft SEAL offered lower latency, while OpenFHE provided higher numerical accuracy and lower error rates. This study analyses how sensitive voice biomarker data can be processed using ML and how FHE can be used to provide security and privacy. The proposed model serves as a roadmap for the steps and models required to ensure the security of voice data.



PITCH SESSION 1 — ABSTRACT 2

ASSESSING VALIDITY OF ACOUSTIC BIOMARKERS: AN EVALUATION FRAMEWORK

Felix Burkhardt[1,*], Anna Derington[1], Monica Gonzalez-Machorro[1], Pascal Hecker[1], Uwe Reichel[1], Hagen Wierstorf[1], Dagmar Schuller[1,2], Rahul Swaminathan[1], Julia Weigl[1], Björn Schuller[1,3]

[1]*audEERING GmbH, Germany*
[2]*Hochschule Landshut, Germany*
[3]*CHI – Chair of Health Informatics, TUM University Hospital, Germany*
[*]*Presented by*

This contribution describes the audEERING approach to validating the effectiveness of acoustic biomarkers. Building on an existing framework for testing the correctness, fairness, and robustness of speech-emotion-recognition models, we adapt the framework to acoustic biomarkers. The framework evaluates correctness, fairness, robustness, and efficiency and is designed to model the verification and analytical-validation stages of the V3 framework. It distinguishes among features, such as acoustic biomarkers or models predicting them; tests, such as comparisons between original and augmented data for robustness or stability across languages; and conditions, such as GRBAS data used for validation. Correctness is assessed using clinical databases with ground-truth labels. Robustness is evaluated by comparing feature predictions for original and slightly augmented data. Fairness is assessed by measuring the statistical relevance of confounders, while efficiency quantifies real-time capability and required computing resources. We also introduce general evaluation concepts such as statistical test-signature distance, which can be applied to both robustness and fairness assessment. The framework is still being adapted, but initial work indicates that only a small number of changes to the previous framework are needed to replace speech-expression models with acoustic biomarkers. We therefore introduce a framework for testing the verification and validation of acoustic biomarkers and report initial results.

PITCH SESSION 1 — ABSTRACT 3

A SCALABLE END-TO-END FRAMEWORK FOR VOICE BIOMARKER DEVELOPMENT

Lars Nippert[1], Prashanth Pombala[1,*], Julia Hoxha[1]

[*]*Presented by*

Voice biomarkers represent a promising non-invasive modality for quantifying physiological and pathological states across clinical domains. However, translation into clinical practice is hindered by fragmented processing pipelines, limited reproducibility, and a lack of scalable analytical frameworks. We present VOICE-BIOME, a fully integrated end-to-end platform for standardized voice-biomarker development that enables reproducible processing from raw audio signals to validated machine-learning models. The platform consolidates signal preprocessing, feature extraction, model training, validation, and visualization within a unified architecture. Its no-code workflow interface, BIOME Studio, enables the configuration of complete analytical pipelines while ensuring traceability and reproducibility. VOICE-BIOME supports extraction of comprehensive acoustic and prosodic feature sets, including spectral, formant, glottal, perturbation, and pitch-related descriptors. Built-in process management facilitates scalable experimentation and collaborative research. In addition to other studies, the platform was evaluated in a heart-failure cohort comprising patients across the acute heart-failure spectrum. In a longitudinal study with 30 participants, models were trained to classify decompensated states relative to individual stable baselines. Performance was assessed using true-positive rate, false-positive rate, balanced accuracy, and area under the curve. The best-performing configurations achieved a balanced accuracy of up to 91.9% for free speech, with a true-positive rate of 90.0% and a false-positive rate of 6.3% under baseline conditions. Task-dependent performance differences indicated trade-offs between sensitivity and specificity. These results demonstrate the robustness and scalability of VOICE-BIOME™ for voice-biomarker modelling and support its applicability as a disease-agnostic framework for clinical research and deployment.

PITCH SESSION 1 — ABSTRACT 4

ON ADAPTIVE FEDERATED LEARNING FOR VOCAL BIOMARKER PROCESSING UNDER PRIVACY CONSTRAINTS

René Glitza[1,*], Luca Becker[1], Rainer Martin[1]

[1]*Institute of Communication Acoustics, Ruhr University Bochum, Bochum, Germany*
[*]*Presented by*

Federated learning (FL) offers a promising framework for collaborative machine learning when sensitive data cannot be centrally pooled. This is especially relevant for voice-based health technologies and vocal-biomarker research, where speech data may contain identifiable, clinically sensitive, and institution-specific information. By keeping raw data local and sharing only model updates, FL can support privacy-preserving collaboration across hospitals, clinics, and distributed devices while aligning with data-governance and regulatory requirements. However, FL alone does not solve the practical challenges of real-world deployment. Heterogeneous data distributions, variable data quality, fairness across sites, robustness to faulty or adversarial participants, and residual privacy risks in model updates remain important concerns. These issues are particularly relevant for vocal biomarkers, where datasets often differ across languages, devices, clinical populations, and recording conditions. This contribution presents an overview of FL as a privacy-aware and governance-oriented paradigm for medical AI, with a focus on its relevance to vocal biomarkers. Building on previous work, we discuss pFedMARL, a cooperative multi-agent reinforcement-learning framework for adaptive aggregation and personalization in FL. In prior experiments, pFedMARL improved federated training under heterogeneous and potentially adversarial conditions by dynamically adapting server-side aggregation weights and client-side personalization behaviour. Although it was not developed specifically for vocal biomarkers, these findings suggest that adaptive FL is a promising methodological direction for privacy-preserving multi-site training in voice-based health technologies.

PITCH SESSION 1 — ABSTRACT 5

ON PRIVACY-PRESERVING EDGE-CLOUD LEARNING FOR VOCAL BIOMARKERS USING THE PRIVACY FUNNEL

Luca Becker[1,*], René Glitza[1], Rainer Martin[1]

[1]*Institute of Communication Acoustics, Ruhr University Bochum, Bochum, Germany*
[*]*Presented by*

Voice-based health technologies increasingly rely on edge-cloud processing, where only a small front end of a deep neural network (DNN) can be executed on the local device while the larger back end remains in the cloud. Although this design is computationally attractive, it creates a critical privacy risk: the intermediate representations transmitted to the cloud often leak sensitive information such as speaker identity or other personal attributes. We suggest a privacy-preserving add-on for vocal-biomarker systems based on the privacy funnel (PF). The PF formalizes network optimization as the problem of minimizing information leakage about private variables while preserving sufficient task-relevant information for downstream inference. This trade-off is expressed in terms of mutual information, yielding a loss function that can guide a DNN towards representations that remain informative about the intended biomarker task while concealing private attributes. As in our prior work on privacy-preserving DNNs, we consider a partitioned neural architecture in which a lightweight encoder transforms speech into embeddings before transmission to a cloud decoder containing most of the parameters. We discuss how the PF can be realized through a variational information bottleneck or contrastive representation learning, and how privacy can be evaluated under strong speaker-verification threat models. Based on our prior work, we suggest that PF-based learning is a promising privacy-by-design framework for vocal-biomarker systems because it directly links deployable edge-cloud inference with information-theoretic control of sensitive-information leakage.

PITCH SESSION 2

Neurocognitive, Mental Health & Vocal Biomarkers

PITCH SESSION 2 — ABSTRACT 1

CLASSIFICATION OF DEPRESSION USING SPEECH VARIABILITY FEATURES

Gábor Kiss[1], Attila Zoltán Jenei[1], Dávid Sztahó[1,*]

[1]*Department of Telecommunications and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, Budapest, Hungary*

[*]*Presented by*

Major depressive disorder (MDD) is one of the most prevalent mental illnesses today and represents a significant global burden. Diagnosis is currently performed by psychiatrists, which limits access to a relatively narrow group of professionals and entails substantial costs. To make screening more widely available, machine-learning approaches based on acoustic biomarkers are receiving increasing attention. The pathophysiology of depression is closely related to speech production: psychomotor retardation and cognitive impairments directly affect speech planning and execution, and the speech of depressed individuals often becomes monotonous, with reduced articulatory variety. Speech-continuity features may therefore be effective markers of depression. We examined spontaneous speech samples from 135 depressed and 89 healthy native Hungarian speakers. Speech variability was measured automatically using 10 ms derivatives of speech intensity, the three highest-intensity Mel-band values, and spectral slope. Variability features were extracted separately from fluent and disfluent speech segments, including restarts and hesitations, and were used for classification with support vector machines and nested cross-validation. Speech from depressed participants showed significantly lower overall variability. At the same time, it contained more pauses, restarts, and speech impediments, which typically produced higher local variability. Classification accuracy reached 86%. These findings suggest that the extracted variability features may serve as vocal biomarkers for depression.

PITCH SESSION 2 — ABSTRACT 2

FROM MIXED REALITY TO CLINICAL INSIGHT: AI-BASED PREDICTION OF PARKINSON'S DISEASE SEVERITY

Joanna Stępień[1], Martyna Baran[1,*], Daria Hemmerling[1]

[1]*AGH University of Krakow, Krakow, Poland*
[*]*Presented by*

Parkinson's disease is a progressive neurological condition affecting movement, speech, eye function, and posture. This study explores multimodal data from voice, eye activity, hand movements, and posture collected using mixed-reality systems, with the aim of predicting clinical Parkinson's disease scores using artificial intelligence and deep learning. Features were selected through SHAP-based importance analysis, and missing values were imputed using k-nearest neighbours. We implemented a custom neural architecture with a Masked Feature Attention mechanism in which input features are concatenated with binary availability masks to compute attention weights dynamically. The core model consists of a shared multilayer-perceptron encoder and a task-specific decoder. The applied preprocessing and modelling methods enabled effective prediction of several clinical scales. The highest performance was achieved for MDS-UPDRS II (MAE = 0.154, $R^2 = 0.30$), followed by the Timed Up and Go test (MAE = 0.157, $R^2 = 0.43$) and MDS-UPDRS III (MAE = 0.177, $R^2 = 0.28$). Other scales, including MoCA and the Berg Balance Scale, showed moderate performance ($R^2 \approx 0.29$–0.31), whereas more heterogeneous measures such as MDS-UPDRS IV yielded lower predictive accuracy. The proposed multimodal approach shows that clinical scores in Parkinson's disease can be predicted with moderate accuracy, particularly for motor assessments, and suggests that mixed-reality-based multimodal analysis may support objective and scalable monitoring of disease progression.

PITCH SESSION 2 — ABSTRACT 3

VARIATIONS IN THE ASSESSMENT OF PARKINSON'S DISEASE ACROSS DIFFERENT MODALITIES

Attila Zoltán Jenei[1,*], Dávid Sztahó[1], Gábor Kiss[1]

[1]*Department of Telecommunications and Artificial Intelligence, Budapest University of Technology and Economics, Budapest, Hungary*

[*]*Presented by*

Parkinson's disease (PD) is one of the most common neurological disorders, alongside Alzheimer's disease. Diagnosis is difficult because of the wide range of symptoms and the absence of a single specific diagnostic test. The goal of this work is to develop a procedure that assists physicians in diagnosis. A dataset containing speech, drawing, and movement signals was collected from 33 individuals with PD and 47 healthy controls. Features were extracted from the three recording modalities using x-vectors, MobileNet, and time-series feature-extraction methods. Classification was performed using support vector machines, random forests, and k-nearest neighbours within a nested cross-validation procedure. A subsequent post-model fusion stage combined the estimates produced by the individual modalities. The combined use of all three modalities achieved the best performance, with 94.6% balanced accuracy and a standard deviation of 7.0%, comparable to the speech-only model, which achieved 91.7% balanced accuracy with a standard deviation of 9.3%. The three-modality model significantly outperformed drawing or movement alone, although its performance was not significantly different from that of the speech models. These results indicate that speech alone may serve as a strong biomarker for PD detection; nevertheless, adding complementary modalities is recommended to capture the variability of early symptoms.

PITCH SESSION 2 — ABSTRACT 4

A CLINICAL SPEECH DATASET FOR COGNITIVE IMPAIRMENT AND MILD DEMENTIA SCREENINGS

Antonio Perna[1,2], Maria Santina Ler[1], Gennaro Cordasco[3], Anna Esposito[1,2,4,*]

[1]*Università degli Studi della Campania "Luigi Vanvitelli", Department of Psychology, Caserta, Italy*
[2]*SIREN – Italian Society of Neural Networks, Vietri sul Mare (SA), Italy*
[3]*University of Salerno (UNISA), Department of Computer Science, Salerno, Italy*
[4]*Corvinus Institute for Advanced Studies, Corvinus University of Budapest, Budapest, Hungary*
[*]*Presented by*

Dementia and mild cognitive impairment (MCI) are major clinical challenges, and early diagnosis is often hindered by reliance on time-consuming specialist-administered neuropsychological evaluations that are not uniformly available. Automatic speech analysis has emerged as a promising digital biomarker capable of identifying cognitive changes through fluency, prosody, and pause organisation, but standardized data resources remain limited. We introduce an Italian clinical speech corpus designed to support scalable, non-invasive screening tools. The acquisition protocol included two tasks: reading aloud the Italian version of Aesop's fable *The North Wind and the Sun*, and producing autobiographical narratives about positive and negative life events. Recordings were collected under supervised clinical conditions in a quiet environment and stored as uncompressed single-channel WAV files at 44.1 kHz and 16-bit resolution. Multiple repetitions of the reading task were collected for some participants to capture within-subject variability. The dataset comprises 132 Italian speakers aged 45–89 years, all assessed using the Montreal Cognitive Assessment and the Mini-Mental State Examination. The clinical cohort ($n = 78$) includes 35 patients with Alzheimer's disease and 43 with MCI; it comprises 33 men and 45 women, with a mean age of 71.8 ± 9.3 years and 8.7 ± 4.03 years of education. Mean MoCA and MMSE scores were 17.6 ± 6.4 and 23.3 ± 4.6, respectively, and the corpus also includes descriptors of apathy, anxious-depressive symptoms, vascular burden, and other clinical comorbidities. The healthy-control cohort ($n = 54$) comprises 23 men and 31 women, with a mean age of 69.1 ± 8.5 years, 11.6 ± 4.6 years of education, and mean MoCA and MMSE scores of 24.6 ± 3.6 and 27.7 ± 2.05, respectively. Descriptive statistics show that cognitive deterioration affects the temporal organisation of speech: patients with Alzheimer's disease exhibit significantly longer and more variable productions than participants with MCI and healthy controls. The dataset supports extraction of vocal biomarkers based on both low-level and high-level speech descriptors and provides a transparent empirical foundation for training advanced computational architectures.

PITCH SESSION 2 — ABSTRACT 5

VOICE SIGNATURES OF MOMENTARY PSYCHOLOGICAL STRESS IN REAL-LIFE ENVIRONMENTS: RESULTS FROM THE COLIVE VOICE STUDY

Noémie Topalian[1,2,*], Abir Elbéji[1,*], Charline Bour[1,3], Hanin Ayadi[1], Mégane Pizzimenti[1], Camille Perchoux[2], Guy Fagherazzi[1]

[1]*Deep Digital Phenotyping Research Unit, Department of Precision Health, Luxembourg Institute of Health (LIH), Strassen, Luxembourg*
[2]*Urban Development and Mobility, Luxembourg Institute of Socio-Economic Research (LISER), Esch-sur-Alzette, Luxembourg*
[3]*Université Grenoble Alpes, Fonds de Dotation Clinatec, Grenoble, France*
[*]*Presented by*

Voice is hypothesized to be modulated by stress and may therefore provide a means of stress detection and monitoring. This study investigates the effect of momentary psychological stress on voice in real-life recordings across languages, genders, and vocal tasks. Participants in the Colive Voice study reported their stress level on a five-point Likert scale and performed two tasks: reading a text and sustaining the vowel /a/. Vocal features were extracted using the DisVoice library, and ordinary least-squares regressions were used to evaluate associations between vocal features and stress. Models were stratified by gender and language (French or English), adjusted for demographic and lifestyle factors and chronic health conditions, and corrected for multiple testing. The analysis included 4,155 participants: 2,011 completed the protocol in French (1,621 women and 390 men), and 2,144 in English (1,105 women and 1,039 men). During text reading, stress was associated with two articulatory features among English-speaking women. Among French-speaking women, stress was associated with pitch and shimmer, while pause duration and one glottal feature were also related to stress. During sustained /a/ phonation, pitch and variability of the pitch perturbation quotient decreased with stress among English-speaking men, whereas voice intensity and loudness increased with stress among French-speaking women. The findings confirm associations between momentary stress and several vocal features in real-life settings, but these associations were not consistent across languages, tasks, or genders.

PITCH SESSION 2 — ABSTRACT 6

UNDERSTANDING NEURAL CORRELATES OF SPEECH AND LANGUAGE CHANGES IN NEUROLOGICAL CONDITIONS

Melanie Jouaiti[1,*], Fatima Al Haji[1], Magda Chechlacz[1]

[1]*School of Computer Science, University of Birmingham, Birmingham, UK*
[*]*Presented by*

Speech and language are increasingly recognized as promising biomarkers of dementia. Understanding their neural underpinnings is therefore fundamental to improving diagnosis and developing effective pathology-specific biomarkers. This project combines two neuroimaging techniques: magnetic resonance imaging (MRI), a well-established component of diagnostic pathways, and functional near-infrared spectroscopy (fNIRS), a less invasive technique with good spatial resolution. The aim is to link speech and language changes to alterations in language-specific neural networks in neurodegeneration, including structural network changes measured by MRI and functional network changes measured by fNIRS. The project will map dementia-subtype-specific speech impairments to structural and functional changes in the connectivity of language-supporting neural networks using tasks that engage different cognitive processes. Simultaneous speech and fNIRS data will be collected during a panel of speech and language tasks, including picture description, verbal fluency, and reading. For each task, extracted speech features will be correlated with anomalies in brain structure and connectivity. Participants will include people with mild cognitive impairment, healthy older adults, and individuals with neurological conditions. Pilot data from healthy participants already show distinct activation patterns across the different speech and language tasks. This ongoing work will investigate how observable speech and language changes relate to neural activation and connectivity, potentially supporting a more accurate understanding and diagnosis of neurological conditions.

PITCH SESSION 2 — ABSTRACT 7

MACHINE LEARNING-BASED PREDICTION MODELS FOR COGNITIVE DECLINE PROGRESSION: A COMPARATIVE STUDY IN MULTILINGUAL SETTINGS USING SPEECH ANALYSIS

B. Ceyhan[1], S. Bek[2], Tuğba Önal-Süzek[3,*]

[1]*Department of Software Engineering, Yaşar University, İzmir, Türkiye*
[2]*Department of Neurology, Faculty of Medicine, Muğla Sıtkı Koçman University, Muğla, Türkiye*
[3]*Department of Bioinformatics, Graduate School of Natural and Applied Sciences, Muğla Sıtkı Koçman University, Muğla, Türkiye*
[*]*Presented by*

Mild cognitive impairment (MCI) is commonly associated with dementia, making early prediction of progression from MCI to dementia important for preventing or alleviating cognitive decline. Because dementia affects cognitive functions such as language and speech, speech analysis may provide a convenient and cost-effective way to detect disease progression. We examined spontaneous speech and written Mini-Mental State Examination scores from a 60-patient dataset collected at the Muğla University Dementia Outpatient Clinic and a 153-patient dataset from the Alzheimer's Dementia Recognition through Spontaneous Speech challenge. To our knowledge, this is the first study to analyse the contribution of acoustic features extracted from spontaneous speech to the distinction between different degrees of cognitive impairment using both an Indo-European language and a Turkic language, which differ substantially in word order, agglutination, noun cases, and grammatical marking. When each machine-learning model was evaluated on the language on which it had been trained, a random-forest classifier achieved 95% accuracy on the English-language ADReSS dataset, while a multilayer-perceptron neural network achieved 94% accuracy on the Turkish-language clinical dataset. In a second experiment, each language-specific model was evaluated on the other language's dataset, yielding accuracies of 72% and 76% for English and Turkish, respectively. These findings demonstrate the cross-language potential of acoustic features for automated monitoring of cognitive-impairment progression in patients with MCI and support their use as a cost-effective option for clinicians and patients.

PITCH SESSION 2 — ABSTRACT 8

AWKWARD ACOUSTIC ALLIANCES: HOW VOCAL BIOMARKER DEVELOPERS NAVIGATE CONTRADICTIONS IN MENTAL HEALTHCARE DEBATES

Martijn Logtenberg[1,*]

[1]*School of Governance, Utrecht University, Utrecht, The Netherlands*
[*]*Presented by*

Vocal biomarker developers working in mental health occupy a contradictory position. They present their technologies as moving psychiatry towards more objective, transdiagnostic approaches that transcend established diagnostic categories, yet the same algorithms are commonly trained on datasets labelled using categorical DSM diagnoses. This study explains why this expectation gap exists and how developers navigate the resulting paradox. It combines interviews with key stakeholders, including business leaders, researchers, and healthcare professionals, with participant observation at industry conferences and analysis of corporate materials. Together, these methods reveal the coalitional pressures that shape technological choices and expectations. Regulators, insurers, and clinical partners demand conformity with established diagnostic standards as a condition for validation, reimbursement, and adoption. This creates a practical bind: dimensional or transdiagnostic approaches may better suit the technology and expand addressable markets, but categorical approaches are currently rewarded by the healthcare system. Although developers can temporarily defer system-level change, doing so intensifies tensions between investor expectations and clinical practice. Developers respond through several strategies, including pivoting from diagnosis towards symptom-severity monitoring, embedding vocal biomarkers in broader technology ecosystems to reduce regulatory exposure, or leaving clinical markets for applications with lower evidentiary requirements. The expectations raised by developers are therefore not purely technical or scientific; they are political and coalition-dependent. Recognizing this helps explain a broader pattern in healthcare AI, in which the gap between technological promise and institutional reality shapes the plausibility of success.

PITCH SESSION 2 — ABSTRACT 9

SPEECH EMOTION RECOGNITION BASED ON CONVOLUTIONAL NEURAL NETWORKS

Rajko Jotanović[1], Jovan Galić[1,*], Gordana Gardašević[1]

[1]*Department of Telecommunications, Faculty of Electrical Engineering, University of Banja Luka, Banja Luka, Bosnia and Herzegovina*

[*]*Presented by*

Recent advances in speech emotion recognition highlight the growing importance of paralinguistic speech analysis in domains extending beyond human–computer interaction to biomedical engineering and digital health. In particular, extracting and interpreting acoustic and prosodic features from speech signals provides a promising basis for vocal biomarkers associated with physiological and psychological conditions. This work presents a convolutional-neural-network approach for classifying five emotional states: neutral, anger, fear, sadness, and happiness. The model uses automatic feature learning from spectrogram representations to capture complex, nonlinear acoustic patterns. Experiments were conducted on the GEES database of acted emotional speech, using Mel-frequency cepstral coefficients as the primary acoustic features. To assess robustness and generalization, two experimental scenarios were evaluated: speaker-dependent and speaker-independent classification. The model achieved an overall accuracy of 93.8% in the speaker-dependent scenario and 84.4% in the speaker-independent scenario. These findings support the eVoiceNet initiative by demonstrating how established speech-emotion-recognition techniques can be adapted to vocal-biomarker frameworks. The work also promotes collaboration between the speech-processing and biomedical-research communities and encourages the integration of engineering methods into emerging digital-health solutions.

PITCH SESSION 3

Clinical Health Applications & AI for Vocal Biomarkers

PITCH SESSION 3 — ABSTRACT 1

EXPLAINABLE MACHINE LEARNING FRAMEWORK FOR VOCAL BIOMARKER IDENTIFICATION IN NEUROLOGICAL AND SYSTEMIC DISEASES

Veysel Alcan[1,*], Bothanya Sakkini[2]

[1]*Department of Electrical and Electronics Engineering, Engineering Faculty, Tarsus University, Mersin, Türkiye*
[2]*Department of Electrical and Electronics Engineering, Graduate Education Institute, Tarsus University, Mersin, Türkiye*
[*]*Presented by*

Human voice is increasingly recognized as a non-invasive digital biomarker reflecting interactions among respiratory, laryngeal, and neural systems. However, most studies focus on binary classification (healthy vs pathological), while the ability of voice to differentiate between diseases remains limited. We propose an explainable cross-disease vocal biomarker framework to distinguish Parkinson's disease, lung disease, and thyroid disease using a publicly available multi-task voice dataset. Subject-level representations were constructed by aggregating approximately 40 recordings per individual, including phonetic utterances and cough signals, using an open-source dataset provided by the Advanced Biometric Research Center. A comprehensive handcrafted feature extraction pipeline was implemented, covering pitch–voicing, perturbation, harmonicity, acoustic, spectral, and cepstral domains. The framework integrates non-parametric statistical analysis, feature selection, multiclass and pairwise machine learning, and explainable AI methods, including permutation importance and feature block ablation. A compact biomarker subset of 15 features outperformed the full 65-feature space, achieving a balanced accuracy of 0.662 in multiclass classification. Pairwise analysis revealed heterogeneous separability across diseases: Parkinson vs Thyroid (AUC = 0.877) and Lung vs Thyroid (AUC = 0.834) showed strong discrimination, whereas Lung vs Parkinson remained near chance level (AUC = 0.558). Explainability analyses consistently identified voice breaks, cepstral peak prominence (CPP), MFCC11, and perturbation-related features as dominant biomarkers. These findings demonstrate that voice signals encode disease-specific acoustic signatures beyond generic pathology detection. The proposed framework supports a shift toward explainable disease differentiation and provides a reproducible methodology aligned with emerging standards in vocal biomarker research and AI-driven digital health.

PITCH SESSION 3 — ABSTRACT 2

THE EVOLVING LANDSCAPE OF VOICE AND SPEECH BIOMARKERS: A BIBLIOMETRIC ANALYSIS

Hilal Berber Çiftci[1], Ferhat Alkan[1,*], Namık Yücel Birol[2]

[1]*Tarsus University, Faculty of Health Sciences, Department of Speech and Language Therapy, Mersin, Türkiye*
[2]*Cappadocia University, School of Health Sciences, Department of Speech and Language Therapy, Nevşehir, Türkiye*
[*]*Presented by*

Voice and speech biomarkers have gained increasing attention as non-invasive indicators for screening, diagnosis, and monitoring across neurological, psychiatric, and medical conditions. This study aimed to map the intellectual and thematic structure of voice and speech biomarker research through bibliometric analysis. Records indexed in the Web of Science Core Collection were searched on March 7, 2026, using Topic terms related to vocal, voice, speech, acoustic, and digital voice biomarkers. A total of 349 documents published between 2009 and 2025 were retrieved and analyzed using bibliometric performance indicators and science mapping techniques. The dataset comprised 203 sources, 1,879 authors, and 10,528 references, with an annual growth rate of 35.02%, an average of 10.13 citations per document, and an international co-authorship rate of 33.81%. Annual scientific production increased markedly after 2021, rising from 20 publications in 2021 to 122 in 2025. The most productive authors were Quatieri, Fagherazzi, and Tröger, while Harvard University was the leading affiliation. Trend topics increasingly centered on machine learning and vocal biomarkers. Thematic mapping identified machine learning, artificial intelligence, and voice biomarkers as motor themes. The most locally cited references were Fagherazzi et al. (2021) and Cummins et al. (2015). The field has shifted from limited early output to a rapidly expanding research area, with recent studies clustering around computational methods and disorder-specific applications, indicating a more structured and clinically focused knowledge base.

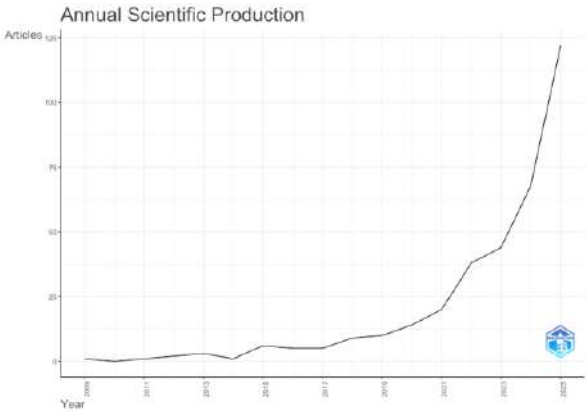


Figure: Annual scientific production in voice and speech biomarker research from 2009 to 2025, showing a marked increase after 2021

PITCH SESSION 3 — ABSTRACT 3

SAMU VOICE: TOWARD AI-ASSISTED DETECTION OF ACUTE MYOCARDIAL INFARCTION THROUGH VOCAL BIOMARKERS IN FRENCH EMERGENCY CALLS

Elio Stasica[1,*], Clément Joly[2,3], Romain Serizel[1], Vincent P. Martin[1], Emmanuel Vincent[1], Tahar Chouihed[2,3]

[1]*Université de Lorraine, CNRS, Inria, LORIA, 54000 Nancy, France*
[2]*Emergency Medical Services, CHRU-Nancy, Université de Lorraine, 54000 Nancy, France*
[3]*Emergency Medical Services, CHRU-Nancy, INSERM, Université de Lorraine, CIC, 54000 Nancy, France*
[*]*Presented by*

In France, the SAMU (Emergency Medical Services) handles 32 million calls annually. Half of these involve chest pain, among which medical regulators have to identify Acute Myocardial Infarction (AMI), which requires emergency care: one-year mortality risk rising 7.5% for every 30-minute diagnostic delay. The SAMU VOICE project, conducted through the ENACT consortium bridging the Nancy University Hospital and INRIA, investigates whether vocal and linguistic features extracted from emergency call audio can support physicians in faster decision-making when faced with AMI-looking clinical pictures. First, we designed a clinically grounded corpus collection methodology capturing both correctly triaged and initially missed cases, as well as different sub-phenotypes of AMI, planning to include a total of 10,000 patients. Second, we launched an ongoing perceptual study measuring physicians' ability to identify AMI from call recordings. Finally, we developed a pipeline to automatize this task, including diarization, transcription, speaker identification, acoustic features extraction and symptom networks computation. On a pilot dataset of 100 patients, classification experiments yielded inconclusive results, as expected at this scale. These experiments nonetheless validated the data collection and annotation methodology, established baseline performance for diarization ($\sim$18% DER) and transcription ($\sim$33% WER), and identified the directions to pursue on the full corpus. Full-scale corpus collection, perceptual study and model fine-tuning are underway. Once these objectives are achieved, SAMU VOICE will be fully equipped to lay the ground for the first systematic evaluation of AI-assisted AMI detection from emergency call audio.

PITCH SESSION 3 — ABSTRACT 4

LANGUAGE INDEPENDENT VOICE/SPEECH BIOMARKERS

Ilhana Setic-Avdagic[1,*]

[1]*ENT Department, ASA Hospital Sarajevo, SSST, Sarajevo Medical School, Bosnia and Herzegovina*
[*]*Presented by*

Many non-laryngeal diseases – neurological, respiratory and metabolic conditions – can affect voice production due to their impact on respiratory support, laryngeal muscle control and neuromuscular coordination. These voice changes are often subtle or occur before traditional symptoms, making voice analysis a valuable tool for early screening, disease staging and remote monitoring, and thus can serve as a digital biomarker for disease screening. It is particularly interesting to mark those voice/speech features that can function independently of the language or specific vocabulary the speaker uses. A literature review of language-independent voice biomarkers in voice-affected disorders published in the last 5 years was conducted using Scopus, PubMed and Web of Science databases. Among the many disorders affecting the voice, Parkinson's disease has been the most studied in relation to vocal biomarkers. Research has identified several key acoustic features that remain robust regardless of the language spoken: Fundamental Frequency (F0/Pitch), Jitter, Shimmer, Harmonic-to-Noise Ratio, Speech Pause Duration and Frequency, Formants (F1, F2), Spectral Flux and Centroid. Recent research into language-independent voice and speech biomarkers has shifted toward utilizing acoustic, paralinguistic, and prosodic features – such as pitch, rhythm, and vocal fold movements – that are universal across human populations, rather than relying on language-specific semantic content. These biomarkers allow AI models to detect neurological, mental, or physical impairments without requiring translation or language-specific training. In order to determine the most relevant, sensitive and applicable language-independent voice/speech biomarkers, further prospective clinical trials in demographically and linguistically diverse populations are needed.

PITCH SESSION 3 — ABSTRACT 5

STANDARDIZATION OF MEASURES AS A PREREQUISITE FOR APPLICABILITY IN CLINICAL PRACTICE

Katarina Pavičić Dokoza[1,*], Zdravko Kolundžić[2], Anja Liović[2]

[1]*SUVAG Polyclinic, Zagreb, Croatia*
[2]*Faculty for Speech and Language Disorders, University of Rijeka, Croatia*
[*]*Presented by*

Data related to the prevalence of voice disorders vary significantly in different studies. The Voice Handicap Index (VHI) questionnaire is the most commonly used questionnaire. Research conducted in the Republic of Croatia showed that the cut-off point of 18.32 with 93.44% sensitivity and 98.08% specificity for VHI-HR was determined. The point prevalence of perceived voice disorders (PVD) was 45.68% for teachers, and 21.11% for non-teachers. The odds of having PVD for teachers are 2.83 times higher than for non-teachers. After determining the prevalence of voice disorders in the general population and in the teacher population, the question arose as to what is associated with the occurrence of vocal fatigue that significantly impacts vocal function. The Croatian adaptation of the Vocal Fatigue Index (VFI-C) was used to quantify the degree of perceived vocal fatigue. Structural equation modelling and path analysis confirmed the significant direct relationship between teaching communication scenario and VFI-C and fatigue and perceived voice disorders, which emphasized the need for development of prevention-oriented programs for voice disorders in teachers. Work in progress is the validation of the Acoustic Voice Quality Index (AVQI), Versions 2.06 and 3.02, in the Croatian Language. AVQI has proven to be a valid measure for discriminating healthy from dysphonic voice comprised of six acoustic measures. Validation of AVQI for the Croatian language together with previous research will enable a better combination of subjective and objective measures in clinical settings both in terms of diagnostics and therapy.

PITCH SESSION 3 — ABSTRACT 6

VOICE BIOMARKERS FOR MULTI-CONDITION HEALTH SCREENING USING A SHORT NATURAL SPEECH SAMPLE: A REAL-WORLD VALIDATION

Lara Gervaise[1,*], Pierre Schweitzer[1], Julian Fritsch[1]

[1] *Virtuosis, Lausanne, Switzerland*
[*] *Presented by*

Voice contains physiological information reflecting neurological, respiratory, and mental states. We aim to develop scalable, language-agnostic vocal biomarkers for multi-condition screening using a short, natural speech sample. We developed machine learning models using acoustic features extracted from $\geq$30-second speech recordings collected from $N = 13,266$ participants across diverse populations, devices, languages, and real-world conversational setups (e.g., clinical interviews vs self-recordings). Models were trained and validated against both diagnoses and clinical reference standards, including PHQ-9, GAD-7, UPDRS, and MOCA. Performance was evaluated using ROC-AUC, sensitivity, specificity, and F1-score, with subgroup analyses across demographics and recording conditions. Models demonstrated strong generalization across languages and devices, with ROC-AUC > 0.80 for depression, Parkinson's disease, and Alzheimer's disease screening tasks. Performance variability remained acceptable across conversational setups, supporting the robustness of a unified acoustic-only approach. Longer recording durations were associated with increased performance variability depending on the conversational setup. Acoustic voice biomarkers enable accessible, non-invasive, and scalable health screening across multiple conditions. Their language-agnostic nature and low data requirements support integration into global care pathways. Future work will focus on prospective clinical validation in larger populations (N> 30,000).

PITCH SESSION 3 — ABSTRACT 7

ACOUSTIC VOCAL BIOMARKERS FOR OBJECTIVE SPEECH MONITORING IN GLIOMA PATIENTS UNDERGOING AWAKE SURGERY: A MULTI-MODAL VALIDATION STUDY

Luca Ducceschi[1,*], F.C. van Ierschot[2], M.J.E. van Zandvoort[2]

[1]*Eurac Research, Bolzano, Italy*
[2]*Department of Neurology & Neurosurgery, University Medical Center Utrecht / Experimental Psychology, Utrecht University, Utrecht, The Netherlands*
[*]*Presented by*

During awake glioma surgery, preserving communicative ability is paramount. Yet, intra-operative monitoring relies on subjective, time-consuming manual scoring that often fails to detect subtle neuromotor linguistic degradation. We evaluated an AI/NLP-driven acoustic pipeline to objectively measure spontaneous speech, shifting assessment from subjective observation to rapid, quantifiable digital biomarkers. Spontaneous speech samples via a picture description task were collected longitudinally from 11 glioma patients across four surgical phases: pre-operative (T0), intra-operative (T1), direct post-operative (T2), and four-month post-operative follow-up (T3), yielding 44 total recordings. Audio files were processed using DisVoice, an open-source pathological speech analysis Python module, extracting more than 600 acoustic features spanning from phonation stability, articulation kinematics, to prosodic trajectories. Trained clinicians manually scored samples for Mean Length of Utterance (MLU), noun and verb production accuracy, and syntactic complexity. Convergent validity was assessed via Spearman's rank correlations between automated acoustic proxies and manual clinical scores. The pipeline proved robust against ambient intra-operative noise. Acoustic metrics significantly decoupled distinct clinical impairments: increasing syntactic complexity (MLU) destabilized laryngeal endurance (Period Perturbation Quotient/Jitter, rho = 0.46, $p < .01$), while cognitive hesitation during noun retrieval was powerfully predicted by degradation in articulatory voice-offset precision (Bark Band Energies, rho = -0.62, $p < .0001$). Furthermore, articulation rate (VRate) robustly mapped overall neuromotor fluency (rho = 0.38, $p < .05$). Global fluency naming tasks remain coupled to respiratory-laryngeal control stability (Shimmer, rho = -0.33, $p < .05$). Acoustic vocal biomarkers sensitively capture subclinical neuromotor and linguistic deteriorations during glioma surgery. By quantifying critical cognitive-motor crosstalk, this automated approach provides a scalable, objective paradigm for real-time intra-operative mapping, urgently needed to prevent permanent post-surgical deficits and optimize patient quality of life.

PITCH SESSION 3 — ABSTRACT 8

VOICE BIOMARKERS OF TURNER SYNDROME: INITIAL EVIDENCE FROM SUSTAINED VOWELS

Marc Freixes[1,*], Jordi Sanz[1], Joan Claudi Socoró[1], Jordi Margalef[1], Maria Esther Esteban[2], Aroa Casado[2], Isabella Monlleó[3], Debora Michelatto[3], Francesc Alías-Pujol[1], Alejandro González[1], Álvaro Heredia-Lidón[1], Alba Llauró[1], Gema Beltran-Serrano[1], Mar Coch-Alcina[1], Neus Martínez-Abadías[2], Xavier Sevillano[1]

[1]*HER Human Environment Research Group, La Salle - Universitat Ramon Llull (URL), Barcelona, Spain*
[2]*Universitat de Barcelona, Barcelona, Spain*
[3]*Universide Federal de Alagoas, Maceió, Brazil*
[*]*Presented by*

Turner syndrome is a rare chromosomal disorder affecting females, characterised by marked phenotypic variability and frequent diagnostic delays. The BeNeXT project aims to develop non-invasive, scalable diagnostic and prognostic tools integrating facial, body, gait and voice biomarkers with genomic data and machine-learning. This study focuses on the identification of voice biomarkers. Nineteen Brazilian participants with Turner syndrome and forty-three controls were recorded using a smartphone. Acoustic features were extracted from sustained /a/ vowels using Praat and openSMILE and age-related effects were residualized. Group differences were analysed using non-parametric statistical tests, and supervised machine-learning models were trained to evaluate discriminative performance. Significant differences between Turner syndrome and control participants were observed mainly in formant-related features and shimmer-based measures. Supervised models achieved fair discrimination between participants (AUC 0.7-0.8), with tree-based ensemble methods showing better performance. These findings indicate that sustained vowels capture interpretable acoustic differences associated with Turner syndrome. Sustained vowels provide a feasible basis for identifying vocal biomarkers of Turner syndrome. Both statistical and machine-learning analyses highlight the relevance of resonance and amplitude perturbation features. Despite the limited sample, these preliminary findings support sustained phonation as a promising starting point for voice-based characterisation of Turner syndrome.

PITCH SESSION 3 — ABSTRACT 9

BREATHING AND SPEECH PATTERNS FOR PREDICTIVE MODELLING OF EXACERBATIONS OF ASTHMA AND COPD

Mikołaj Najda[1,*], Yuyang Yan[1], Loes van Bemmel[2], Frits M.E. Franssen[2], Sami O. Simons[2], Visara Urovi[1]

[1]*Institute of Data Science, Maastricht University, The Netherlands*
[2]*Department of Respiratory Medicine, Maastricht University Medical Centre, The Netherlands*
[*]*Presented by*

COPD and asthma cause significant morbidity and mortality across the globe. The development of non-invasive telemonitoring solutions offers the potential to improve patient care. The aim of the study is to examine whether breathing and speech patterns extracted from audio recordings enhance model generalizability of distinguishing stable from exacerbation periods in COPD and asthma. Patterns related to breathing and speech were extracted from a dataset that combined daily audio speech recordings with a patient-reported outcome measure (EXACT). Medical experts categorized patient health status as stable or exacerbation. A tree-based model was investigated for exacerbation prediction. The model was trained on signal-related features, breathing and speech patterns, and combination of both. Results were examined for variability among all people and feature importance. The study included 21 people (mean age 62.7 years; 66.7% female), 61.9% diagnosed with COPD and 38.1% with asthma. Minimum breath duration and speech duration differed between stable and exacerbation states ($p < 0.05$). The model, trained on combined features sets achieved Sensitivity of 0.78, 0.10 SD exceeding models trained on acoustic-only (0.70, 0.17 SD) and patterns-only sets (0.68, 0.13 SD). Inconsistencies were observed in patient-level results across exacerbation severity. Speech Duration, Spectral Contrast, Harmonic to Noise Ratio, Breath Group Duration and 26 Mel-Frequency Cepstral Coefficient were found to have the highest importance for the best performing model. Combining signal derived features with breathing and speech patterns helps models to balance performance across metrics.

PITCH SESSION 3 — ABSTRACT 10

ADVANCED VOICE BIOMARKERS FOR HEART FAILURE DETECTION USING MACHINE LEARNING AND EXPLAINABLE ARTIFICIAL INTELLIGENCE

Justyna Krzywdziak[1], Sławomir Pawłowski[2], Miłosz Dudek[1], Guido Caluori[3], Paweł Kurzelowski[2], Monika Kalicka[2], Wojciech Wojakowski[2], Tomasz Jadczyk[2,*], Daria Hemmerling[1]

[1]*AGH University of Krakow, Poland*
[2]*Medical University of Silesia, Poland*
[3]*University of Bordeaux, IHU Liryc, France*
[*]*Presented by*

Voice-based biomarkers represent a non-invasive approach for heart failure (HF) screening. This study evaluates whether novel vowel space metrics and explainable artificial intelligence can improve HF detection and interpretability. The study included 122 participants (72 with HF and 50 healthy controls, HC). A total of 386 phonation samples of five Polish vowels (/a/, /e/, /i/, /o/, /u/) were analyzed. Acoustic feature extraction included standard and novel parameters – Vowel Space Area (VSA) and Formant Centralization Ratio (FCR). Machine learning models were trained using 5-fold cross-validation. Model interpretability was assessed using SHapley Additive exPlanations. Models achieved strong discrimination, with AUC values of 0.845 ± 0.046 (logistic regression), 0.879 ± 0.057 (random forest), 0.883 ± 0.032 (gradient boosting), and 0.890 ± 0.034 (multilayer perceptron), with corresponding accuracies up to 0.821 ± 0.049 and F1 score up to 0.756 ± 0.056. Significant group differences were observed for alpha ratio (-22.00 ± 10.56 vs -16.53 ± 10.80; $p < 0.001$), Hammarberg index (30.30 ± 10.60 vs 23.55 ± 10.97; $p < 0.001$), and F0 stability index (4.04 ± 39.85 vs 54.62 ± 46.70; $p = 0.0035$). In contrast, VSA (84350 ± 39760 vs 84326 ± 35991, $p = 0.7259$), FCR (2.02 ± 0.17 vs 2.05 ± 0.14, $p = 0.8402$) and vowel articulation index (0.50 ± 0.04 vs 0.49 ± 0.03; $p = 0.8402$) were not significantly different. Advanced acoustic analysis combined with machine learning and explainable artificial intelligence offers a promising HF screening. Further studies are needed to confirm these preliminary findings.

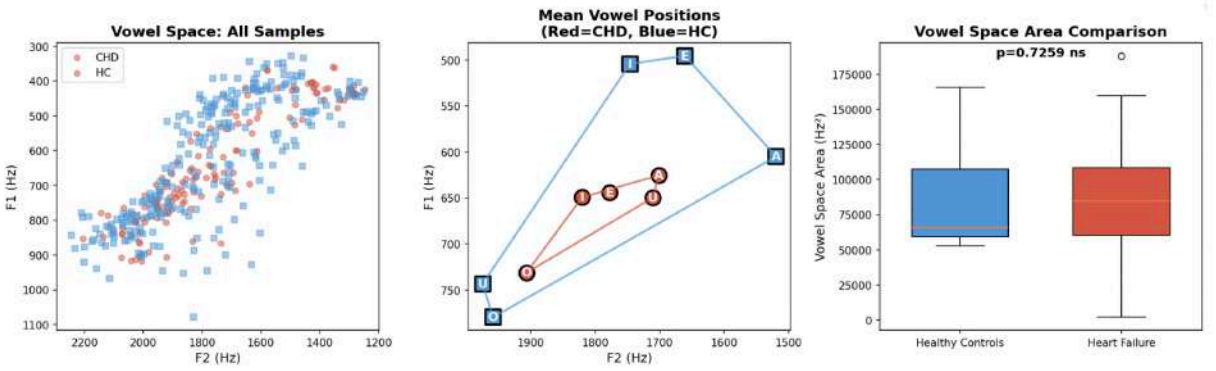


Figure: Results

POSTER PRESENTATIONS

POSTER 1

VOCAL BIOMARKERS AND AI-BASED VOICE TECHNOLOGIES - KNOWLEDGE, ATTITUDES, AND READINESS OF SPEECH AND LANGUAGE THERAPISTS

Anna Andreeva[1]

[1,*]*Department of Logopedics, South-West University "Neofit Rilski", Blagoevgrad, Bulgaria*
[*]*Presented by*

Artificially intelligence-based voice technologies use subtle changes in speech as digital biomarkers for screening, diagnosis, and monitoring across many conditions. Research is rapidly growing, but translation into routine speech and language therapy practice is still early and raises practical, ethical, and training questions. Despite the potential of AI-based technologies, little is known about the readiness of speech and language therapists to adopt such technologies. This study aims to evaluate students' and practitioners in speech and language therapy knowledge, attitudes, and perceived barriers regarding vocal biomarkers, with a focus on AI integration, clinical validation, ethical concerns, and implementation challenges. For the purpose of the study an online survey was developed. The participants are 25 students in speech and language therapy and 25 professionals (speech and language therapists). The data will be analyzed with SPSS tool. Findings will inform training, policy, and clinical integration strategies about the knowledge, attitudes and the readiness of the participants to use AI-based technologies in their practice aligned with emerging digital health frameworks.

POSTER 2

EXPLOITING LARGE FOUNDATIONAL MODELS FOR EXTRACTION OF LINGUISTIC BIOMARKERS FROM TELEFONIC SPEECH

Dijana Petrovska Delacretaz[1,*], Taha Yaou[1], Sarah Eddahbi[1], Danouma Alpha Ouattara[1], Graziella Mangone[2], Sara Sambin[2], Alizé Chalançon[2], Manon Gomes[2], Stéphane Lehéricy[2], Jean-Christophe Corvol[2], Marie Vidailhet[2], Isabelle Arnulf[2], Nesma Houmani[1], Laetitia Jeancoals[2]

[1]*SAMOVAR, Telecom SudParis, Institut Polytechnique de Paris, Evry, France*
[2]*Sorbonne Université, Paris Brain Institute, CM, Inserm, CNRS, APHP, Hôpital Pitié-Salpêtrière, 75013 Paris, France*
[*]*Presented by*

Parkinson's disease (PD) is a progressive neurodegenerative disorder affecting millions worldwide, characterized by a wide range of motor and non-motor symptoms. Among these symptoms, alterations in speech and voice quality stand out as early and prominent indicators of the disease. The emergence of powerful speech foundation AI models has revolutionized the field by providing powerful tools for speech processing and feature extraction. Our goal is to exploit publicly available large language models (LLMs) to analyze early Parkinson and healthy controls speech data in order to extract vocal biomarkers from the speech to text transcriptions. The database at our disposal is the ICEBERG dataset, a longitudinal observational study carried out at the Clinical Investigation Center for Neurosciences within the Paris Brain Institute (ICM). 334 subjects were followed during five years and underwent clinical evaluations, neurological examinations, motor and cognitive evaluations, biological sampling, brain magnetic resonance imaging (MRI), as well as yearly audio-visual recordings at the hospital and monthly telephonic recordings calling a telephonic server. In our previous study we extracted embeddings with wav2vec2.0, Whisper and SeamlessM4T models from the hospital recordings for PD detection. In this study we are analyzing the telephonic data using transcriptions with OpenAI's Whisper Large V3 and FrWhisper open-source models. FrWhisper is a fine-tuned version of Whisper Large V3 model, specifically optimized for French speech recognition with enhanced capabilities for transcribing interjections, hesitations, word repetitions, and interrupted words in conversational French. These transcriptions provide complementary information that are relevant for the early Parkinson disease detection.

POSTER 3

ARTIFICIAL INTELLIGENCE DIAGNOSTICS FOR SMALL HEALTH SYSTEMS: OPPORTUNITIES IN OTORHINOLARYNGOLOGY

Elvir Zvrko[1,*]

[1]*Department of Otorhinolaryngology, University of Montenegro, Podgorica, Montenegro*
[*]*Presented by*

Small health systems face limited specialist availability, restricted access to advanced diagnostics, and delayed referral pathways. In otorhinolaryngology, these constraints may reduce timely detection of laryngeal and other head and neck diseases. This abstract explores the potential of artificial intelligence (AI)-driven diagnostics, including vocal biomarkers, to improve efficiency, accessibility, and early detection in small health systems. A narrative translational overview was conducted of current AI applications relevant to otorhinolaryngology, with emphasis on voice analysis, endoscopic image interpretation, and diagnostic support models applicable to low-resource or small-scale healthcare settings. AI-based diagnostic tools show strong potential to support earlier triage, non-invasive screening, and more efficient referral of patients with voice disorders and suspected laryngeal pathology. Voice-based models are particularly promising because they are low-cost, scalable, and suitable for remote use. In small health systems, such tools may help compensate for workforce shortages, improve access for geographically dispersed populations, and support more equitable care delivery. AI diagnostics, especially voice-based approaches, may become highly valuable in small health systems by enabling earlier detection, better resource allocation, and broader access to specialist-oriented assessment. Their successful implementation will depend on clinical validation, workflow integration, and appropriate regulatory and data protection frameworks.

POSTER 4

NKULULEKO: A SOFTWARE TOOL FOR RAPID VOICE DATABASE PROCESSING

Felix Burkhardt[1,*], Bagus Atmaja[2], Björn Schuller[1,3]

[1]*audEERING GmbH, Germany*
[2]*NAIST, Japan*
[3]*CHI Chair of Health Informatics, TUM University Hospital, Germany*
[*]*Presented by*

Nkululeko [Burkhardt and Atmaja, 2025] is a Python toolkit for audio-based machine learning that uses a command-line interface and configuration files, eliminating the need for users to write code. Built on sklearn and PyTorch, it enables training and evaluation of speech databases with state-of-the-art machine learning approaches and acoustic features. Key capabilities include model demonstration, database storage with predicted labels, and bias detection through correlation analysis of target labels (e.g., depression) with speaker characteristics (age, gender) or signal quality metrics. Nkululeko targets novice users interested in speaker characteristics detection (emotion, age, gender) without programming expertise, focusing on education and research. Core design principles include exploring combinations of acoustic features, models, and preprocessing for optimal performance, database analysis with visualizations, and inference on audio files or streams. Users run experiments via a single command. Open-source tools accelerate science through security, customizability, and transparency. While several open-source tools exist for audio analysis, none specialize in speech analysis with high-level interfaces for novices. Nkululeko fills this gap with key principles: minimal programming skills (CSV data preparation and command-line execution); standardized data formats (CSV and AUDFORMAT); replicability through shareable configuration files; high-level INI-file interface requiring no Python coding; and transparency via comprehensive debug output and automated reporting. Nkululeko has been used in several research projects since 2022. We believe it to be a useful tool to the community of acoustic biomarker practitioners to perform initial experiments on databases in a fast and efficient way, including advanced deep learning based end-to-end models without the need to become a dedicated Python programmer.

POSTER 5

SELF-SUPERVISED SPEECH REPRESENTATIONS FOR VOCAL BIOMARKER DEVELOPMENT: OPPORTUNITIES, CHALLENGES, AND VALIDATION NEEDS

Gül Tahaoğlu[1,*]

[1]*Department of Computer Engineering, Karadeniz Technical University, Türkiye*
[*]*Presented by*

Vocal biomarkers are emerging as promising tools for non-invasive, scalable, and repeatable health assessment. Recently, self-supervised learning (SSL) models such as wav2vec 2.0, HuBERT, and WavLM have enabled the extraction of rich speech representations without requiring large-labelled datasets. This abstract examines the role of SSL-based audio features in vocal biomarker development and discusses their methodological potential and validation needs. It is presented a focused methodological review of recent studies using SSL representations for voice-based analysis of neurological, psychiatric, and respiratory conditions. The review compares SSL-based approaches with conventional handcrafted acoustic features in terms of representation capacity, robustness, transferability, and data efficiency. Key methodological dimensions including model selection, layer choice, feature aggregation, clinical interpretability, privacy, and external validation are considered. Recent findings indicate that SSL features often outperform traditional acoustic descriptors, particularly in low-resource and cross-task settings. They capture clinically relevant paralinguistic and phonetic information and support improved generalization across datasets. However, performance remains sensitive to dataset bias, recording variability, language mismatch, and limited explainability. Insufficient external validation and weak clinical benchmarking still restrict real-world deployment. SSL-based speech representations provide a strong foundation for next-generation vocal biomarkers. Future work should prioritize clinically grounded validation, explainable modelling, privacy-aware design, and standardized benchmarking to translate these methods into trustworthy healthcare applications.

POSTER 6

SEPARATING ARTICULATORY AND PHONATORY INFORMATION IN PARKINSON'S DISEASE SPEECH USING SPECTRAL IMAGE REPRESENTATIONS AND NEURAL NETWORKS

Iván Alba Elhazaz[1,*], Juan Ignacio Godino-Llorente[1]

[1]*Bioengineering and Optoelectronics Group, Dept. of Signals, Systems and Radiocommunication, ETSI Telecomunicación — Universidad Politécnica de Madrid (UPM), Ciudad Universitaria, Madrid, Spain*

[*]*Presented by*

Dysarthria in Parkinson's Disease (PD) affects both articulation and phonation, yet speech-based biomarkers may merge these two sources of pathological information. This study aims to disentangle articulatory from phonatory contributions to speech degradation in PD using decomposed spectral image representations, evaluated in a Diadochokinetic (DDK) task context — a controlled, clinically meaningful interpretable paradigm. DDK sequences (speech recordings of /pa/, /ta/, /ka/) were drawn from four multilingual datasets: GITA, NEUROVOZ, CZECH, and GERMAN, providing cross-lingual validation. Two parallel spectral image streams were derived: (1) a spectrogram of the Linear Prediction Coding (LPC) residual signal, capturing predominantly phonatory (glottal source) information; and (2) a spectrogram reconstructed from autoregressive spectral envelope coefficients, capturing articulatory filter dynamics. Both representations were used to independently train Convolutional Neural Network (CNN) classifiers for PD versus healthy control discrimination. Additionally, formant trajectories extracted via the KARMA algorithm were explored as a complementary articulatory representation, enabling both image-based and multivariate time-series modelling approaches. Results were evaluated separately and compared to evaluate the individual and synergic contributions of the glottal source residual and the articulation. Both glottal and articulatory spectral representations yield discriminative performance above change level across the datasets evaluated. Furthermore, the results show their respective discriminative power as well as their specific generalisation capabilities across multiple datasets which has been a limiting factor in recent studies concerning the presented corpora. By decomposing the speech signal into phonatory and articulatory image streams, this framework quantifies the relative discriminatory power of each subsystem for the screening of PD. The multi-dataset design supports generalisability claims. This approach offers a principled methodology for interpretable vocal biomarker development in PD.

POSTER 7

FROM ALGORITHM TO CERTIFIED SOLUTION & STRATEGY FOR AI SPEECH BIOMARKERS ON THE EUROPEAN MARKET: A FONAFIX CASE STUDY

Ján Mucha[1,*], Jiří Mekyska[1], Zoltán Galáž[1]

[1]*Scicake s.r.o., Purkyňova 648/127, 61200, Brno, Czech Republic*
[*]*Presented by*

The integration of vocal biomarkers into clinical practice faces two primary barriers: strict medical device regulations (MDR) and a lack of proven go-to-market models. The Fonafix system presents an approach for transforming AI-based speech analysis into a standardized diagnostic tool for dysarthria within the European context. The system is being developed in strict compliance with the technical requirements for medical devices under EU Regulation 2017/745 (MDR). The current strategy involves transitioning from technical preparation for Class I to the certification process for Class IIa, including clinical evaluation according to MEDDEV 2.7/1 and the implementation of a vigilance system. The research component relies on multilingual validation and an analysis of clinical specialists' needs in the CEE region and Southern Europe. Clinical validation has confirmed the system's ability to objectify speech parameters and reduce diagnostic time by 80%, from 90 to 10 minutes. Market research identified a critical unmet need, as up to 95% of surveyed professionals currently do not use any dedicated software for dysarthria assessment. The market entry strategy targets a significant potential within the EU, where approximately 2.4 million patients suffer from dysarthria, utilizing a hybrid B2B model for clinical facilities and integration into public health insurance systems. Successful implementation of vocal biomarkers requires the parallel development of technical documentation and a scalable business strategy. Fonafix demonstrates that bridging academic research with regulatory rigor and a clear business model is key to the sustainable development of digital healthcare in the European Union.

POSTER 8

END-TO-END SECURE VOICE BIOMARKER PIPELINE FOR POST-QUANTUM ERA

Kübra Seyhan[1,*], Sedat Akleylek[2]

[1]*Department of Computer Engineering, Ondokuz Mayıs University, Samsun, Türkiye*
[2]*Chair of Security and Theoretical Computer Science, University of Tartu, Tartu, Estonia*
[*]*Presented by*

Voice biomarkers contain sensitive biometric data, raising both short-term and long-term privacy and security concerns. The static nature of voice biomarker data has also put it at risk of "store now, decrypt later" with the emergence of post-quantum security. In this paper, a quantum-resistant voice biomarker pipeline is proposed to address the security and privacy of voice biomarkers. We propose a voice biomarker pipeline framework that integrates post-quantum cryptography (PQC) across multiple stages to provide long-term security and privacy. In the proposed pipeline, the data acquisition, modeling and analysis, storage and management, and transmission phases are designed to incorporate PQC. The security concerns are identified, and how they can be addressed using specific PQC are explained. The proposed end-to-end framework defines the security and privacy-focused problems, requirements, and necessary precautions in voice biomarker processing pipelines. The details of key security concerns and potential cryptographic primitives and integration with PQC standards are analyzed from the data acquisition to transmission phases. The proposed end-to-end secure voice biomarker processing pipeline framework is designed to provide long-term protection for the security and privacy of sensitive data by addressing potential risks and solutions at the pipeline level.

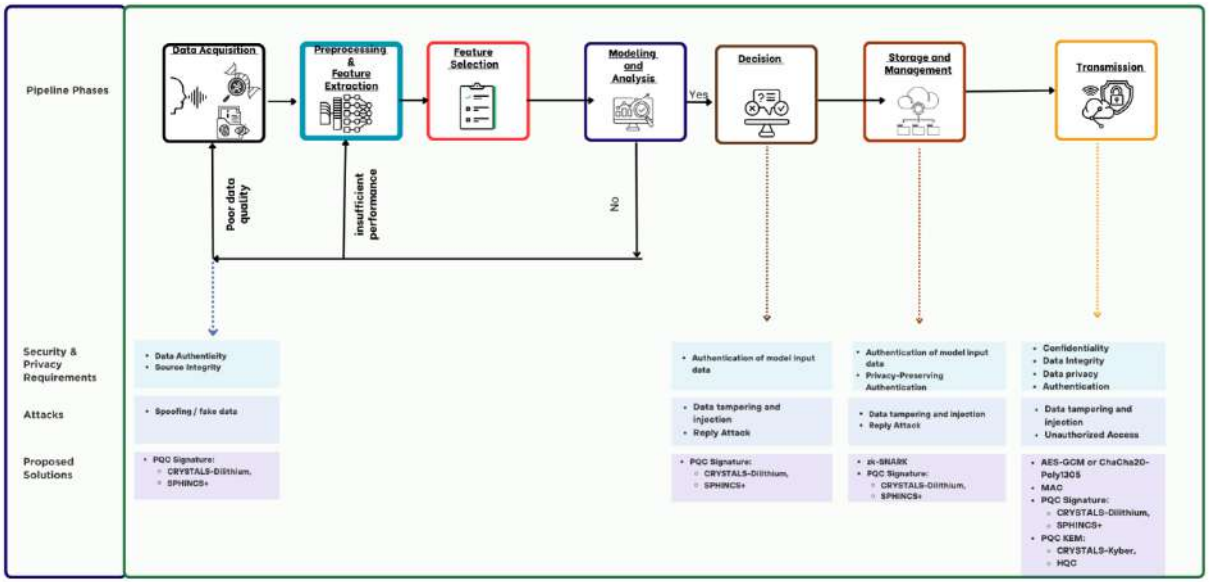


Figure: Proposed framework

POSTER 9

VOCAL BIOMARKERS FOR EMPHYSEMA DETECTION IN PATIENTS WITH COPD: DATA FROM THE ENBED-STUDY

Lauren G. Reinders[1,2,*], Simona Schäfer[3], Loes van Bemmel[1,2], Louisa Schwed[3], Alexander Mackay[4], David Nobbs[4], Julia Gomez[4], Ragnar Lunde[5], Martijn Cuijpers[5], Frits M.E. Franssen[2,6], Hester A. Gietema[7], Sami O. Simons[1,2]

[1]*NUTRIM Institute of Nutrition and Translational Research in Metabolism, Maastricht University, Maastricht, the Netherlands*
[2]*Department of Respiratory Medicine, Maastricht University Medical Center+, Maastricht, the Netherlands*
[3]*ki:elements GmbH, Saarbrücken, Germany*
[4]*Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd., Basel, Switzerland*
[5]*Laurentius Hospital, Roermond, the Netherlands*
[6]*Department of Research and Development, Ciro, Horn, the Netherlands*
[7]*Department of Radiology and Nuclear Medicine, Maastricht University Medical Centre, Maastricht, The Netherlands*
[*]*Presented by*

Emphysema represents a distinct phenotype within Chronic Obstructive Pulmonary Disease (COPD), characterized by alveolar wall destruction and static and dynamic lung hyperinflation. Diagnosis is made by chest CT-scans, however, these are costly and not widely available. Yet, accurate identification of emphysema phenotype is essential for clinical decision making. Voice could be a low-cost, accessible alternative. We investigated the association between speech features and emphysema percentage in patients with COPD. Patients with COPD performed a sustained vowel /a/ and paced reading task before and after exercise. Emphysema on chest CT-scans were quantified using Syngo.via (Siemens Healthineers®) and voice features were extracted using Sigma library. Data was analyzed using a Pearson regression analysis, covariates age, sex and BMI were regressed out for each feature. 196 patients were included, of whom 91 females, with a mean age of 68 years (SD8.6). In the paced reading task before exercise, the H1-A3 Harmonic difference (SD and mean) and H1-H2 harmonic difference (mean) were significantly correlated with emphysema percentage (respectively rho = 0.26, -0.24, 0.22). After exercise, the rate loudness peaks and H1-A3 harmonic difference (SD) were significantly correlated (both rho = 0.30). The sustained vowel /a/ duration after exercise was negatively correlated with emphysema percentage (rho = -0.23). Only findings after exercise remained significant after correction for multiple testing. Speech features, particularly those reflecting airflow instability obtained after exercise, showed consistent associations with emphysema percentage. This supports the hypothesis that vocal biomarkers are sensitive to underlying pathophysiological changes in COPD and may contribute to its phenotypic characterization.

POSTER 10

VOICE BIOMARKERS TO PREDICT COPD EXACERBATIONS IN PRIMARY CARE: THE VOCES PROSPECTIVE STUDY

Brenda Leon-Gomez[1], Pere Toran[1], Anfal Boulaayoun[2], Marcos Faundez-Zanuy[2,*]

[1]*Institut d'Investigació en Atenció Primària de Salut Jordi Gol, Barcelona, Spain*
[2]*TecnoCampus Mataró, Universitat Pompeu Fabra, Barcelona, Spain*
[*]*Presented by*

Chronic obstructive pulmonary disease (COPD) exacerbations are associated with functional decline, hospitalization, and mortality. Voice analysis is a promising non-invasive digital biomarker, but prospective real-world evidence in primary care remains limited. The VOCES study aims to evaluate the prognostic value of acoustic voice analysis for predicting COPD exacerbations over 24 months. VOCES is an observational, longitudinal, prospective study in five primary care centers in Mataró, Spain. Adults aged 35–65 years with COPD will be recruited into two age- and sex-matched cohorts: exacerbators (n=200) and non-exacerbators (n=67). At baseline and month 24, voice will be recorded using a standardized protocol at rest and after a 6-minute walk test. Acoustic measures will be integrated with spirometry, oxygen saturation, heart rate, respiratory rate, inflammatory biomarkers, and patient-reported outcomes. Primary outcomes are time to first exacerbation, total number of exacerbations, and annualized exacerbation rate. Survival, count, and mixed-effects models will be used. Recruitment is planned in a real-world primary care setting. We expect to determine whether baseline and post-exertion vocal biomarkers improve exacerbation risk stratification and reflect physiological and clinical status. VOCES will generate prospective evidence on the feasibility and prognostic utility of voice as a digital biomarker for COPD monitoring in primary care. We welcome collaborations, methodological feedback, and partnership opportunities to strengthen study development and translational impact.

POSTER 11

A SPEAKING SKILLS TRAINING AND ASSESSMENT PLATFORM FOR PROFESSIONAL COMMUNICATION DEVELOPMENT

Maria Koutsombogera[1,*]

[1]*Endurae Voice Technology OÜ*
[*]*Presented by*

We introduce Interveu, a tool that measures personalised speech performance, delivered through an online course. It is a solution that teaches employees, learners or any adult to speak better and sound more professional and confident. It is ideal for professionals whose voice is critical (salespeople, interviewees, teachers, etc.) and adds value to teams by unlocking individual speaking potential. Interveu is built on speech processing and personality computing techniques. Its core features include measuring perceived "Big 5" personality traits (openness, conscientiousness, extraversion, agreeableness, neuroticism) and voice skills (volume, speed, pitch, pauses, voice stability). The pedagogical framework selected (Understanding by Design) is modular, allowing the adaptation of the course to a variety of training scenarios. The platform has been carefully designed to include built-in anti-bias measures. The prototype has promising results in the prediction accuracy of each of the Big 5 personality traits. Interveu has been successfully validated through trials with VET and Higher Education learners in 8 EU countries, with > 300 learners giving a 95% positive recommendation rate. Interveu offers added value to users by providing personalised feedback about their perceived personality together with observations related to speaking style and tips on how to improve their speaking performance. Most importantly, it educates before measuring, through an online course designed with strong pedagogy principles. Interveu seeks to empower its users by addressing them as learners of communication skills, offering professional development and not being a mere decision-making application.

Figure: Stages of the Interveu tool functionality, excluding the online course

POSTER 12

SAFEGUARDING VOCAL BIOMARKERS: DETECTING DEEPFAKE AUDIO USING COCHLEAGRAM REPRESENTATIONS AND FUSED TRANSFORMER ARCHITECTURES

Metehan Bulut[1], Gül Tahaoğlu[1,*], Guzin Ulutas[1], Beste Ustubioglu[1], Arda Ustubioglu[1], Mustafa Ulutas[1], Samet Dincer[1,*], Tekin Altun[1]

[1]*Department of Computer Engineering, Karadeniz Technical University, Trabzon, Türkiye*
[*]*Presented by*

The integration of vocal biomarkers in telehealth relies on the authentic physiological characteristics of the patient's voice. With the significant advancements in deepfake voice technologies in recent years, distinguishing fake voices from real voices has become quite difficult. Malicious actors can use advanced synthesis methods to artificially mimic pathological voice features (e.g., tremor or shortness of breath), leading to serious clinical misdiagnoses. This study aims to protect voice-based health technologies by developing a robust approach to distinguish misleading synthetic speech from real clinical voice. To overcome the limitations of traditional acoustic features in capturing synthetic anomalies, we utilized cochleagram representations. These simulate the human cochlea's bio-physical processes, yielding detailed time-frequency features. For classification, we employed an ensemble deep learning strategy combining Vision Transformers (ViT) for global context modeling and Cross-Covariance Image Transformers (XCIT) for efficient feature-dimension attention. A late mean score fusion was applied and evaluated on the ASVspoof 2019 dataset. The proposed fusion model, combining ViT and XCIT with biologically inspired cochleagrams, achieved an Equal Error Rate (EER) of 6.94% and a minimum tandem Detection Cost Function (min t-DCF) of 0.11. This demonstrates significant efficacy in detecting both Text-to-Speech (TTS) and Voice Conversion (VC) spoofing attacks compared to baseline models. Safeguarding vocal biomarkers is critical for telehealth security. Fusing ViT and XCIT architectures with cochleagram features highly increases manipulated voice detection rates. The ability to detect spoofed voices ensures the diagnostic accuracy, data confidentiality, and reliability of AI-powered clinical voice analysis.

POSTER 13

CREATING AND SHARING SENSITIVE SPEECH DATA WHILE PROTECTING PRIVACY

Sebastian Stober[1,*], Suhita Ghosh[1]

[1]*Artificial Intelligence Lab, Otto-von-Guericke-University, Magdeburg, Germany*
[*]*Presented by*

Protecting the identity of speakers in audio recordings is critical when working with sensitive data, particularly in healthcare settings. Existing anonymization techniques struggle to preserve the natural rhythm, intonation, and distinctive speech characteristics common in elderly and pathological populations—qualities that are vital for effective remote health monitoring. We introduce DDSP-QbE, a voice conversion approach that combines differentiable digital signal processing with query-by-example to overcome these limitations. Through human perception-motivated designed training objectives, the model learns to separately handle linguistic content, prosodic features, and domain-specific speech patterns, allowing it to generalize to atypical and clinically relevant speech. Both objective metrics and human evaluations confirm that DDSP-QbE substantially surpasses current state-of-the-art voice conversion systems in preserving intelligibility, prosody, and speech-domain characteristics across a range of datasets, conditions, and speakers, without compromising audio quality or speaker anonymity. Clinical experts further validated the approach by assessing twelve domain-specific attributes relevant to medical practice.

POSTER 14

INFLUENCE-BASED EXPLAINABLE AI FOR DYSARTHRIA SEVERITY ASSESSMENT USING CASE-LEVEL EVIDENCE

Xiaoliang Wu[1,*], Qiyang Sun[2], Yupei Li[2], Erfan Loweimi[3], Jennifer Williams[1], Zhengjun Yue[4]

[1]*Department of Electronics and Computer Science, University of Southampton, Southampton, United Kingdom*
[2]*Imperial College London, London, United Kingdom*
[3]*University of Edinburgh, Edinburgh, United Kingdom*
[4]*King's College London, London, United Kingdom*
[*]*Presented by*

Dysarthria severity assessment is critical for clinical decision-making, yet manual perceptual evaluation is time-consuming and inconsistent. While deep learning models achieve strong performance, their lack of explainability limits clinical adoption. This study proposes an explainable framework that provides clinically meaningful, case-based evidence for model decisions. We introduce an influence-based, instance-level explainability framework that quantifies how each training sample affects predictions on a given test utterance. Using gradient-based influence estimation, we identify supportive and opposing training samples and aggregate their effects across severity levels. Experiments were conducted on the TORGO dysarthric speech dataset using a four-class classification setup. Explanation validity was evaluated through controlled deletion experiments. Controlled deletion experiments demonstrate that removing high-influence samples significantly degrades model performance, while removing low-influence samples improves it, confirming the faithfulness of influence estimates. Class-level analysis reveals a structured pattern where predictions rely primarily on same-level and adjacent severity samples. Influence strength decreases monotonically with ordinal distance, indicating that the model captures the ordered nature of dysarthria severity. The proposed framework provides explainable, instance-level evidence by linking predictions to perceptible reference speech samples. This enables direct verification by clinicians and supports model auditing. The approach offers a practical pathway toward clinically reliable and explainable speech-based assessment systems.

Acknowledgements

The first International Vocal Biomarkers Conference (IVBC2026) and this Book of Abstracts are the result of the collective effort of the entire eVoiceNet community.

We warmly thank the **Local Organising Committee** at the Luxembourg Institute of Health (LIH) for hosting the conference and making it free and open to all participants, the **Scientific Committee** and the reviewers for shaping a programme of high scientific quality, and the **Working Group leaders and Core Group** of COST Action CA24128 for their continuous work in building this community.

Above all, we thank the keynote speakers, presenters, pitchers, poster authors, and the 130 participants from 35 countries whose contributions fill these pages.

Book of Abstracts prepared by
Sybille Barvaux (Luxembourg Institute of Health, Luxembourg)
Ján Mucha (Scicake, Czech Republic)
Marcos Faundéz-Zanuy (Tecnocampus Mataró, Universitat Pompeu Fabra, Spain)
Dijana Petrovska-Delacrétaz (SAMOVAR, Télécom SudParis, France)
Eralp Doğu (Muğla Sıtkı Koçman University, Türkiye)

Local Organising Committee
The International Vocal Biomarkers Conference was locally organised by the Deep Digital Phenotyping Research Unit (DDP) within the Department of Precision Health at the Luxembourg Institute of Health (LIH).

This publication is based upon work from COST Action eVoiceNet, CA24128, supported by COST (European Cooperation in Science and Technology).

Author Registry

THANK YOU

See you at the next International Vocal Biomarkers Conference

eVoiceNet — COST Action CA24128

evoicenet.eu

DOI: 10.5281/zenodo.21641261