AI Classification of Respiratory Illness through Vocal Biomarkers and A Bespoke Articulatory Speech Protocol

Dr Tim Bashford, Mr Hok Shing Lau, Mr Mark Huntly
Mr Nathan Morgan, Mrs Adesua Iyenoma

Wales Institute of Digital Information (WIDI))
University of Wales Trinity Saint David
Swansea, Wales, United Kingdom.

Dr Tom Powell, Dr Biao Zeng

Wales Institute of Digital Information (WIDI))
University of South Wales
Treforest, Wales

Corresponding author: tim.bashford@uwtsd.ac.uk

***Abstract* - Speech biomarkers represent a powerful indicator for detecting, monitoring and categorising neurological, psychological, pathological and pulmonary conditions. Facilitated by advances in computational power and artificial intelligence (AI) techniques, we present a novel ecosystem for data acquisition, analysis and storage, using an articulatory speech task. By automatically segmenting, aligning and extracting features from the vocal recordings, we present a feature extraction pipeline toward the classification of pathological conditions, specifically respiratory disease through recorded voice. Data is stored within a Trusted Research Environment, for which this work also presents a range of ethical considerations.**

Keywords - Speech biomarkers, respiratory disease, artificial intelligence

I. BACKGROUND

A. Speech biomarkers

Computational analysis of human voice through artificial intelligence and machine learning allow for the prediction and monitoring of pulmonary function.

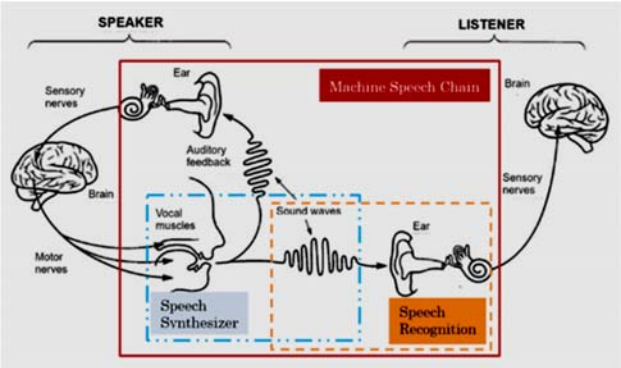

Figure 1. The Machine Speech Chain [1]

While communication is the primary purpose of speech, there is considerable additional information which can be gathered from the process (illustrated in figure 1). Speech generation involves complex coordination of organs (figure 2, meaning speech contains markers of the body's function, from cognitive and mental state to respiratory condition. Even a short sample of speech may provide a complex snapshot of cognition and functioning relevant to many disease areas. Therefore, a significant research effort has been applied towards the discovery of "speech biomarkers", the characteristics of voice (known as features) in speech that can be identified and validated, as associated with a clinical condition of the body.

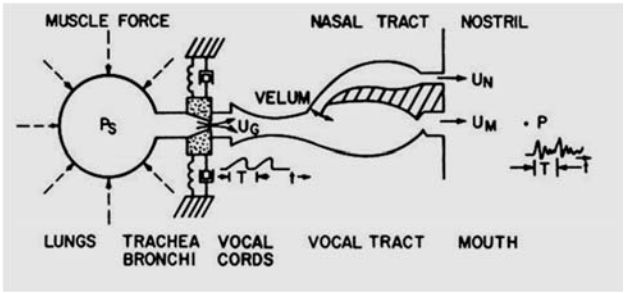

Figure 2. Linear Schematic Diagram of the Vocal Tract System. Sounds are created at this point through acoustic excitation when air is expelled from the lungs through the vocal tract, creating frequency pitch [2].

The potential of using speech to estimate respiration condition is significant and has been widely reported [3] [4], demonstrating that assessment of physical condition and respiratory health can be performed with speech.

B. Rationale for the Development Of Models Utilizing Speech Biomarkers

Previous discovery and utilization of digital biomarkers has been dominated by the traditional methods, which involve:

1) The careful design and selection of handcrafted features, ranging from spectral features to phonation features. Examples of feature sets include COMPARE and GEMAPS

2) The use of statistical models such as Support Vector Machines (SVMs) or Decision Trees as classifiers.

High recognition accuracy can be obtained with these traditional methods [3][5], however previous studies show that the speech features extracted from audio recorded under varied recording conditions, e.g. different recording devices and levels of background noise [6][7] may result in inconsistencies in the model and its decision making.

With the evolution and general success of deep learning strategy in many healthcare-related domains, there has been a trend of moving toward the use of deep learning by:

1) Directly identifying the pathological condition with only a limited number of handcrafted features such as a spectrogram [8].

2) Acting as a feature extractor to transform the data into a quantitative representation that is universal to many downstream tasks [9].

These models can potentially result in higher consistency toward realistic environments [8]. However, that these models and their results lack interpretability has been a frequent criticism. As a result, their utilization in the medical and healthcare sectors has been limited, where interoperability is an important factor in gaining trust from both clinicians and patients.

One of the key challenges that is specific to speech and less addressed by previous research, is the relative lack of knowledge of speech content that maximizes useful information. For example, speech tasks can vary widely, potentially including elements of sustained phonation, script reading, interviews, unstructured conversations, and/or phonemes being articulated at specified rates and frequencies.

C. Trusted Research Environments

Use of digital biomarkers within a machine learning model requires a large quantity of diverse samples from people with a range of backgrounds and clinically diagnosed medical conditions, for training.

Many such samples are recorded in laboratory conditions, using expensive recording equipment in a sound proofed environment to eliminate background noise and capture as accurate a sample as possible. While this ensures the standard and quality of the recording, it significantly raises the barrier for generating vocal samples and, due to the requirement for specialized equipment in a specialized room, does not scale well. Models trained from this clean data suffer degradation in performance when used with samples recorded outside of these conditions. There was significant interest in remotely collecting voice samples using consumer-grade equipment in non-soundproofed environments using web applications, during the COVID-19 pandemic. This approach permitted mass-collection of data, with lesser potential for influence from researcher bias. However, this approach is technically demanding and requires careful adherence to regulatory requirements around privacy of the collected data.

Per the General Data Protection Regulation (GDPR), speech is classified as ‘Special Category Personal Data’. It should be noted that voice, especially speech, can reveal demographic information of an individual, or potentially de-anonymize them entirely. There is no consensus on the suitable anonymization techniques, nor their impact on the successful obtainment of specific biomarkers. To protect participant privacy, work with recorded voice biomarkers requires robust mechanisms for the entire research data lifecycle, from collection, storage and analysis to eventual destruction of data.

In recent years, we have seen the healthcare research paradigm shifting toward to the use of Data Safe Havens (DSHs), or Trusted Research Environments (TREs), digital systems that securely hold and provide restricted access to sensitive data for approved researchers to conduct analysis. Because the data does not leave the TRE, strict security measures protect the privacy of the recorded individual from whom the sample was collected. TREs significantly reduce the potential for data to be misused or re-identified. Therefore, it was considered necessary to utilize such an approach to meet the strict legal requirements of this form of data-driven digital healthcare research.

Within a TRE, researchers are provided with granular levels of access to information based on assessed need. Any data imported to and exported from the TRE is protected at the system level and mechanisms are in place to prevent sensitive data from leaving the TRE. This results in significantly reduced potential for data to be misused or re-identified. Therefore, it is a good solution to meet the strict legal requirements of digital healthcare research. Existing TREs in the UK are generally designed for simple statistical analysis. In recent years, especially during the COVID-19 pandemic, there has been a need for the next generation of TREs that enables advanced analysis of sensitive data in an ethical and secure manner while meeting the needs of researchers, data controllers and the public.

II. OVERVIEW OF ARIA

The ARIA project was established to explore the relationship between speech and respiratory illness, and the tools and techniques necessary to discover and qualify the relationship. The project has a heavy emphasis on Artificial Intelligence and Machine Learning techniques. In this paper, we present our work on the project, including the following:

• ARIA Research Environment: a Trusted Research Environment designed for data acquisition, analysis and storage for voice biomarker discovery and pathological voice analysis.
• VoxLab: software for performing visual analysis of speech and voice biomarkers using a cloud-based machine learning pipeline.
• Helicopter protocol: a novel Speech protocol for assessing pulmonary function.

A. ARIA Research Environment

ARIA is a collaborative ecosystem that creates a secure and safe environment for the exploration of voice data. There is a specific focus on respiratory diseases such as asthma and Chronic Obstructive Pulmonary Disorder (COPD), however the fundamental 'building blocks' of ARIA can be applied to any chronic condition where relevant. Any organization that wishes to work within the ARIA-TRE ecosystem must meet the following criteria:

• Work to improve the health and wellbeing outcomes of people with chronic conditions.
• Adhere to strict established data protection and information governance processes.
• Embrace the collaborative, multi-organizational ethos of the ecosystem to develop UK-based development of voice analysis.

The ARIA ecosystem does not seek to be a single tool or process for the use of voice related analysis but will provide the environment for clinical, academic and industry colleagues (as appropriate) to explore.

B. VoxLab

VoxLab is designed to allow users to perform visual analytics on recorded voice samples, securely. VoxLab is a software application (depicted in figure 3) deployed in the ARIA trusted research environment to allow users to analyze and annotate voice audio data.

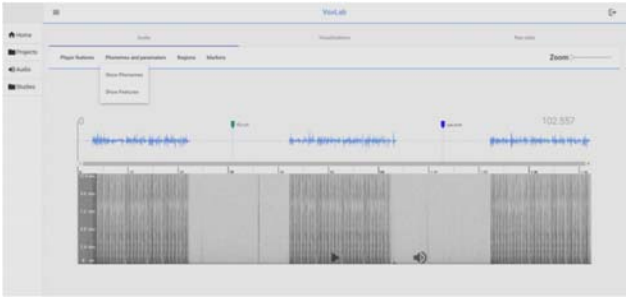

Figure 3. The VoxLab Interface

Fundamental to VoxLab is a cloud-based feature extraction pipeline [10][11][12]. The majority of pathological speech analysis focuses on analyzing the acoustic features of voice during sustained phonation, which makes it difficult to adapt to various speech tasks. We consider speech as segments of speech units, including words and phonemes. This involves extracting low-level acoustic and spectral features such as jitter, shimmer, and harmonic to noise ratio and segmenting higher-level features, such as the word or phoneme presented in the speech. With this structure, the pipeline can extract general voice parameters, by extracting the statistical characteristics of these low-level features, the statistical characteristics of high-level features, such as the duration of pause segments and distribution of phonemes in a running speech and combined features, extracting low level features based on the segment given a high-level feature e.g., the mean jitter of the phoneme 'a' over an articulated passage.

Applying a deep learning model in the context of healthcare requires understanding of its decision-making process. With the advent of explainable AI (XAI), a visualization module was developed to aid the current research effort in interpreting model decisions, in the context of speech biomarkers. As a prototype, we implemented the attention map aggregation [13] for visualizing the decision making of Audio Spectrogram Transformers [14] that have been employed by the studies of audio-based AI classifiers for COVID-19 infection status [15].

C. A Novel Speech Protocol

To accompany the computational model, we utilize a speech breathing task, originated by Zeng et al [16]. The protocol consists of 3 periods of repeating the word "helicopter" as quickly as possible for 20 seconds, followed each time by a 20 second break, for a total of 100 seconds. This swift repetition is intended to apply mild strain to the vocal chords and cause respiratory load, allowing for easier analysis of speech breathing. This is aided by the word containing a voiceless glottal fricative, requiring significant airflow to maintain in rapid succession.

This word optimizes the selection of prosodic and acoustic features, such as utterance length and vocal intensity respectively. Because of the aforementioned respiratory loading, prosodic information about, e.g., the number of syllables produced in a single breath is a significant indicator of respiratory function.

III. FUTURE WORK

Verification of the pipeline is ongoing. An important foundation to the efficacy and accuracy of the pipeline is the segmentation and alignment of words and phonemes, allowing the extraction of features (example in figure 4 from segmented frames of a speech passage. This is the

primary step in voice activated systems [17], and it is also true for the extraction, analysis and manipulation of data held within a speech signal. This is an important sub-problem within speech analysis and research [18], as seen in technologies such as Text To Speech (TTS) and Automatic Speech Recognition (ASR). Makowski and Hossa [19] describe the term as ”the process of dividing a speech signal into discrete non-overlapping fragments”, used in reference to the division into frames for parameterization.

As a result, consideration is given to the accuracy of the speech segmentation and alignment, measured through a Word Error Rate (WER). It is important to the whole extraction pipeline to ensure the most accurate WER, to establish the most accurate segmentation and alignment that will give the basis to the extraction of features for analysis. To this end, approaches to improve WER are considered, with a particular emphasis on accent and pronunciation.

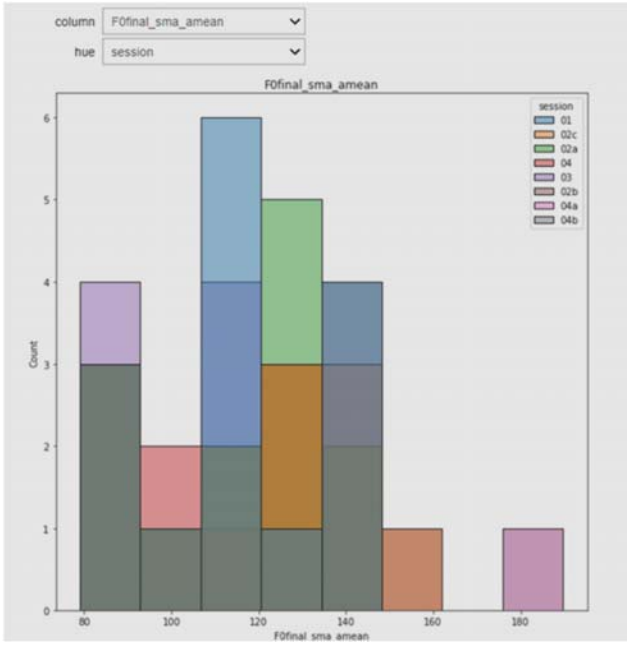

Figure 4. Feature extraction example

Previous work has considered such topics as the correlation between pitch accents and syllable tokens, mostly within content words [20].

An upcoming project named ”Breathing with Long Term COVID Study” (BLoCS) will run with the long COVID unit within Cwm Taf Morganwg University Health Board (CTMUHB). In this work, long COVID patients will, under clinician supervision, complete the speech protocol, with each sample simultaneously recorded on a range of devices to eliminate differences between them caused by, for example, high-pass filtering. Patients will then complete a spirometry test of their pulmonary function, which will be used to calibrate the vocal samples within the model.

REFERENCES

[1] A. Tjandra, S. Sakti, and S. Nakamura, “Machine speech chain,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 976–989, 2020.

[2] L. R. Rabiner and R. W. Schafer, “Introduction to digital speech processing,” Foundations and Trends® in Signal Processing, vol. 1, no. 1–2, pp. 1–194, 2007.

[3] V. Nathan, K. Vatanparvar, M. M. Rahman, E. Nemati, and J. Kuang, “Assessment of chronic pulmonary disease patients using biomarkers from natural speech recorded by mobile devices,” May 2019.

[4] M. Cooper and T. Bashford, “Ai and spotting the sound of illness,” ITNOW, vol. 66, no. 1, pp. 24–25, Feb. 2024.

[5] B. W. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, L. Stappen, S. Ottl, M. Gerczuk, P. Tzirakis, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, L. J. Rothkrantz, J. A. Zwerts, J. Treep, and C. S. Kaandorp, “The interspeech 2021 computational paralinguistics challenge: Covid-19 cough, covid-19 speech, escalation & primates,” in Interspeech 2021, ser. interspeech 2021. ISCA, Aug. 2021.

[6] B. van der Woerd, M. Wu, V. Parsa, P. C. Doyle, and K. Fung, “Evaluation of acoustic analyses of voice in nonoptimized conditions,” Journal of Speech, Language, and Hearing Research, vol. 63, no. 12, pp. 3991–3999, Dec. 2020.

[7] J. Penney, A. Gibson, F. Cox, M. Proctor, and A. Szakay, “A comparison of acoustic correlates of voice quality across different recording devices: A cautionary tale,” in Interspeech 2021, ser. interspeech 2021. ISCA, Aug. 2021.

[8] T. Dang, J. Han, T. Xia, D. Spathis, E. Bondareva, C. SiegeleBrown, J. Chauhan, A. Grammenos, A. Hasthanasombat, R. A. Floto, P. Cicuta, and C. Mascolo, “Exploring longitudinal cough, breath, and voice data for covid-19 progression prediction via sequential deep learning: Model development and validation,” Journal of Medical Internet Research, vol. 24, no. 6, p. e37004, Jun. 2022.

[9] E. Casanova, A. Candido Jr., R. C. Fernandes Jr., M. Finger, L. R. S. Gris, M. A. Ponti, and D. P. Pinto da Silva, “Transfer learning and data augmentation techniques to the covid-19 identification tasks in compare 2021,” in Interspeech 2021, ser. interspeech 2021. ISCA, Aug. 2021.

[10] Y. Maryn and N. Roy, “Sustained vowels and continuous speech in the auditory-perceptual evaluation of dysphonia severity,” Jornal da Sociedade Brasileira de Fonoaudiologia, vol. 24, no. 2, pp. 107–112, 2012.

[11] V. Nathan, M. M. Rahman, K. Vatanparvar, E. Nemati, E. Blackstock, and J. Kuang, “Extraction of voice parameters from continuous running speech for pulmonary disease monitoring,” in 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, Nov. 2019.

[12] Z. Liu, M. Huckvale, and J. McGlashan, “Automated voice pathology discrimination from continuous speech benefits from analysis by phonetic context,” in Interspeech 2022, ser. interspeech 2022. ISCA, Sep. 2022.

[13] H. Chefer, S. Gur, and L. Wolf, “Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers,” in 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Oct. 2021.

[14] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," in Interspeech 2021, ser. interspeech 2021. ISCA, Aug. 2021.

[15] H. Coppock, G. Nicholson, I. Kiskin, V. Koutra, K. Baker, J. Budd, R. Payne, E. Karoune, D. Hurley, A. Titcomb, S. Egglestone, A. Tendero Canadas, L. Butler, R. Jersakova, J. Mellor, S. Patel, ˜ T. Thornley, P. Diggle, S. Richardson, J. Packham, B. W. Schuller, D. Pigoli, S. Gilmour, S. Roberts, and C. Holmes, "Audio-based aiclassifiers show no evidence of improved covid-19 screening over simple symptoms checkers," Nature Machine Intelligence, vol. 6, no. 2, pp. 229–242, Feb. 2024.

[16] B. Zeng, E. M. Williams, C. Owen, C. Zhang, S. K. Davies, K. Evans, and S.-R. Preudhomme, "Exploring the acoustic and prosodic features of a lung-function-sensitive repeated-word speech articulation test," Frontiers in Psychology, vol. 14, Aug. 2023.

[17] A. E. Sakran, S. M. Abdou, S. E. Hamid, and M. Rashwan, "A review: Automatic speech segmentation," International Journal of Computer Science and Mobile Computing, vol. 6, no. 4, pp. 308–315, 2017.

[18] M. A. Al-Manie, M. I. Alkanhal, M. M. Al-Ghamdi, N. Mastorakis, A. Croitoru, V. Balas, E. Son, and V. Mladenov, "Automatic speech segmentation using the arabic phonetic database," in WSEAS International Conference. Proceedings. Mathematics and Computers in Science and Engineering, no. 10. World Scientific and Engineering Academy and Society, 2009.

[19] R. Makowski and R. Hossa, "Automatic speech signal segmentation based on the innovation adaptive filter," International Journal of Applied Mathematics and Computer Science, vol. 24, no. 2, pp. 259–270, Jun. 2014.

[20] S. Ananthakrishnan and S. S. Narayanan, "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence," IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 1, pp. 216–228, Jan. 2008.

.