A Health-Aware and Noise-Robust Speaker Recognition Framework Using Adaptive Neural Architectures

Arundhati Niwatkar[1], Sai Kiran Oruganti [2]

[1]Lincoln University College, Petaling Jaya, Selangor, Malaysia;

[2] Lincoln University College, Petaling Jaya, Selangor, Malaysia

Email ID saisharma@lincoln.edu.my

Abstract: Speaker verification has increased in use as a biometric technology employed to verify users using a variety of authentication methods, including using voice recognition, digital assistants, and providing a secure method to interact with other machines. Contemporary speaker verification systems have made extensive use of deep neural network architectures for speaker verification, including using embeddings for speakers using x-vectors, ECAPA-TDNNs, self-supervised learning, and other related models; however, the performance of these systems is often degraded (sometimes drastically) from real-world impacts, such as environmental noise, mismatch of channels, and physiological influences of the voice due to factors like illness or fatigue. Over the past few years, researchers have made great advances concerning deep learning architectures and methodology, advantages of self-supervised learning for developing speaker representations based on speech and using voice biomarkers for determining health. This review provides an overview of these advances and of the limitations of previous forms of speaker verification (e.g., topography, vocal stability, etc.) in separating the identity of a speaker from health variations/changes caused by issues such as environmental impacts, illness, and other physiological influences. This document will discuss a framework for creating a health-aware and noise-robust speaker verification system using an adaptive learning architecture that utilizes advanced architecture, including dynamic attention mechanisms as well as adversarial trained systems. The goal of the proposed work is to create a speaker verification system that will recognize the identity of the speaker regardless of variations in noise and/or health; thus, it will function effectively as a biometric verification mechanism in a variety of real-world applications.

Keywords: Speaker Verification – Voice Biomarkers; Deep Learning; Noise Robustness; Adaptive Learning Architectures

---

Introduction

Speaker recognition is a popular biometric technology that identifies or verifies users by voice. This technology is widely used in biometric authentication, voice-activated assistance, banking systems, and smart home devices. With the advances in deep learning, the performance of speaker recognition systems has improved significantly over the years. Traditional speaker recognition methods have used manually created features such as Mel-frequency cepstral coefficients (MFCCs) and statistical models

like the Gaussian mixture model (GMM). However, the advent of deep learning means that we can now learn speaker embeddings in an end-to-end fashion using deep neural networks. The x-vector architecture presented by Snyder et al. [2] showed that deep neural networks could produce high-quality representations of speakers. In addition, the ECAPA-TDNN architecture put forth by Desplanques et al. [1] further improved the quality of the embeddings by using channel attention and feature aggregation mechanisms to improve the quality of the embeddings.

Recently, techniques such as wav2vec 2.0, HuBERT, and WavLM, known as self-supervised learning (SSL), have been very successful in how they learn speech representations from vast amounts of data without labels [3]–[5]. However, the majority of voice recognition systems assume that there are stable vocal conditions. In real life, however, various factors will negatively impact the performance of a voice recognition system. For example, speech signals may be negatively affected by background noise, a difference in how two different recording devices are able to capture the same person's voice, and health-related causes like a sore throat, exhaustion or a respiratory illness from listening to someone speak can impact their speech signal. All of these things together contribute to making it very difficult for voice recognition systems to perform correctly; therefore, making it difficult for voice biomarker researchers to rely on speech signals to make any type of medical diagnosis or any other form of analysis [6], [12]. Furthermore, most of the voice biomarker analyses and research conducted today mainly focus on how to diagnose a condition rather than on maintaining speaker identity during person identification when utilizing a voice signal as a method for identifying someone. Thus, there is a significant need for voice biometric/voice recognition systems, which can effectively maintain an individual's identity while also demonstrating resiliency when a person's voice changes due to illness/environment and/or when the voice signal needs to be captured under varying conditions.

Related work

The use of deep learning to increase accuracy in speaker recognition has been growing rapidly since the development of new methods of extracting embedding features from speech signals. Some of the initial works that developed this method include Snyder et al. [2], who introduced the x-vector model as a method of extracting speaker embeddings using Time Delay Neural Networks (TDNNs) as a model for parameterizing speech signals and were particularly successful in representing speakers from the VoxCeleb data set. Subsequent improvements were made with the introduction of the ECAPA-TDNN by Desplanques et al. [1], which utilized channel attention and feature propagation to enhance recognition performance using deep learning. Recent advancements in speaker recognition have also been made with NeXt-TDNN, which was developed by Heo et al. [8], to improve temporal modeling of speaker verification. Zhang et al. [10] developed MEConformer to further enhance the representation of speakers through the use of convolutional neural networks and transformers. These models were developed for use on clean speech processes and do not take into account the physical characteristics of each user’s speech.

Self-supervised learning (SSL) has quickly become popular in speech analysis and processing applications. Baevski et al. [3] introduced the wav2vec 2.0 model, which uses contrastive learning to learn the contextual properties of speech. HuBERT [4], developed by Hsu et al., improved feature

extraction through the use of masked predictions of hidden units in the model. WavLM also improved the modeling of speech by using a large-scale dataset of noisy speech to improve the knowledge of speakers and speech concepts through noise-robust learning [5]. While all three methods provide excellent results for applications in speech processing, none of them support speaker recognition that considers health.

A few different studies have recently looked into using speech signals as an indicator of a person's health state. For example, the Coswara dataset was created by Sharma et al. [6], to provide recordings of breathing, cough and speech in order to detect respiratory diseases. Similarly, Krongauz et al. [12] studied multi-phenotype classification via speech embeddings, while Annabestani et al. [13] applied acoustic biomarkers for detecting voice disorders. Finally, studies conducted by Kim et al. [14] explored how deep neural networks could be utilized to identify mental health conditions via analysis of voice data. As with most of these studies, the primary focus was on whether or not there are indicators of disease, rather than whether an individual's identity can be confirmed using speech signals.

Table 1. Comparison of the related work or previous research by other researchers

Author / Year	**Focus Area**	**Method / Model**	**Dataset**	**Key Contribution**	**Research Gap / Limitation**
Desplanques et al., 2020	Speaker Verification	ECAPA-TDNN	VoxCeleb	Channel attention; improved embeddings	No health-aware modeling; unstable under illness
Snyder et al., 2018	Speaker Embeddings	x-vector TDNN	VoxCeleb	Robust DNN speaker embeddings	Sensitive to noise & physiological variation
Baevski et al., 2020	Self-Supervised Learning	wav2vec 2.0	LibriSpeech	Contextual speech representations	Not optimized for health robustness
Hsu et al., 2021	SSL Speech Modeling	HuBERT	Libri-Light	Masked prediction learning	No multimodal or health feature fusion
Chen et al., 2022	SSL Speech Processing	WavLM	Large corpora	Noise-robust speech representations	Does not model health-induced variation
Sharma et al., 2020	Health Speech Analysis	Acoustic biomarkers	Coswara	Health detection from speech & breath	Identity verification not considered
Yamagishi et al., 2021	ASV Robustness	Spoof detection	ASVspoof	Standard robustness benchmark	Focus on spoofing, not illness variation

Key Contribution

A thorough literature review was completed to identify significant research gaps in the current speaker recognition research area. Most existing speaker recognition models are designed assuming that the vocal characteristics of an individual speaker will be stable across time periods. In reality, there are many occurrences of significant changes to the vocal characteristics of individuals (e.g., changes in voice) due to health-related conditions (e.g., colds, fatigue, respiratory disorders, etc.). Because of this, current speaker verification systems do not model these changes in an individual's voice. As a result, when an individual's voice becomes less stable due to a health-related condition, the ability of a speaker verification system to verify that individual is greatly diminished. Many of the noise reduction methods that are used in speech processing may not only focus on eliminating noise but may also inadvertently drop specific acoustic features that identify the speaker. This may result in a decreased ability to identify the speaker. In addition, research on voice biomarkers and research related to speaker verification are often conducted as two separate areas of research with very little connection between the two. In order to overcome the limitations associated with current speaker recognition systems, we propose the creation of a unique Health-Aware and Noise-Robust Speaker Recognition Framework. This framework integrates health-state modeling with speaker recognition in order to improve the robustness of systems when exposed to real-world conditions.

The key contributions of this work are threefold. First, the study identifies the limitations of existing speaker recognition systems when exposed to health-related voice variations and environmental noise. Second, it proposes the integration of voice biomarker analysis with traditional speaker verification techniques to better understand physiological changes in speech.

Method, Experiments and Results

To develop a speaker recognition system that works effectively to recognize a speaker, regardless of changes and variations in speaker identity because of noise or health, the proposed framework will consist of the following components:

Feature Extraction: The first step will be to process the speech signal using appropriate acoustic feature extraction techniques, such as mel-frequency cepstral coefficients (MFCCs) or deep speech embeddings.

A Deep Neural Encoder: The input audio signals will be processed through a ResNet-based encoder to learn hierarchical representations for all speech input.

A Dynamic Attention Layer: A dynamic mechanism will be used to dynamically assign weights to each of the frequency bands on the basis of the level of noise or the health of the speaker.

Multi-Branch Classification: Adversarial Examples will be used during training to ensure that the Speaker Embeddings are primarily focused on speaker identity features while decreasing the impact of health and noise-related variations. The datasets that will be used for training and testing will consist of VoxCeleb, Coswara, Saarbrücken Voice Database, and the MUSAN noise dataset.

Discussions

The proposed research framework solves many of the limitations found in previous research. By incorporating noise detection and health-state modeling into a speaker recognition architecture, the system is more capable of addressing the variations present in everyday speech. Utilizing dynamic attention mechanisms allows the model to pay attention to the spectral features of speech that are relevant and noise-free, while not considering both the noise and health-state affect the sounds produced when a speaker is speaking. Also, adversarial learning separates identity-related features from the other types of features present in an individual's speech. This research could greatly enhance the speaker recognition systems' ability to work reliably within applications such as secure bank authentication, healthcare monitoring systems and voice-activated smart devices.

Conclusions

Speaker recognition technologies (SRT) are rapidly evolving thanks to deep learning and self-supervised learning techniques. However, most of these systems have been created for use in controlled environments which will perform poorly when applying them to real-world use cases because of how various health states will affect the recorded speech signal and the interference noise from the environment. The current paper reviews the state-of-the-art in both speaker-embedded models and self-supervised learning from speech for considering the applicability of the current research into voice biomarker development. The review indicates a significant investigation gap exists in terms of incorporating health state awareness into SRT. In response to this investigation gap a demonstrates Health-Aware and Noise-Robust Speaker Recognition Framework via Adaptive Neural Architectures is presented. This framework combines dynamic attention mechanisms, adversarial training, and multimodal modeling to increase the performance and robustness of existing systems. Going forward, the next phase of our work is to implement this framework and evaluate its performance against large volumes of multi-condition speech dataset.

References

1. B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification,” in Proc. Interspeech, 2020.
2. D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust DNN Embeddings for Speaker Recognition,” in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.

3. A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020.
4. W.-N. Hsu et al., “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units,” in IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021.
5. S. Chen et al., “WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing,” IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 6, pp. 1505–1518, 2022.
6. N. Sharma et al., “Coswara — A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis,” in Proc. Interspeech, 2020.
7. J. Yamagishi et al., “ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection,” arXiv preprint arXiv:2109.00537, 2021.
8. H.-J. Heo, U.-H. Shin, R. Lee, Y.-J. Cheon, and H.-M. Park, “NeXt-TDNN: Modernizing Multi-Scale Temporal Convolution Backbone for Speaker Verification,” in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024.
9. J.-W. Jung et al., “ESPnet-SPK: Full Pipeline Speaker Verification Toolkit with Multiple Reproducible Recipes, Self-Supervised Front-Ends, and Off-the-Shelf Models,” in Proc. Interspeech, 2024.
10. X. Zhang et al., “MEConformer: Highly Representative Embedding Extractor for Speaker Verification via Selective Convolution,” Expert Systems with Applications, vol. 244, 2024.
11. G. Praveen Rajasekhar and J. Alam, “Less is Enough: Adapting Pre-trained Vision Transformers for Audio-Visual Speaker Verification,” in Proc. NeurIPS Efficient Natural Language and Speech Processing Workshop, 2024.
12. D. Krongauz, H. Pinto, S. Kohn, Y. Marmor, and E. Segal, “HPP-Voice: A Large-Scale Evaluation of Speech Embeddings for Multi-Phenotypic Classification,” arXiv preprint arXiv:2505.16490, 2025.
13. M. Annabestani et al., “AI-Driven Acoustic Voice Biomarker-Based Hierarchical Classification of Benign Laryngeal Voice Disorders,” arXiv preprint arXiv:2512.24628, 2025.
14. J.-W. Kim et al., “Deep Neural Network-Based Analysis of Voice Biomarkers for Monitoring Treatment Response in Adolescent Major Depressive Disorder,” Communications Medicine, vol. 6, 2026.
15. H. Nakamura and S. Tokuno, eds., Voice Biomarkers: Current Status and Issues in the Development. Singapore: Springer, 2025.