Noise-Robust Speaker Recognition System Using Deep 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 [1]pdf.arundhati@lincoln.edu.my; [2]saisharma@lincoln.edu.my

Abstract: Speaker verification is recognized as a highly effective biometric authentication method thanks to its simplicity and high application range. Nowadays, speaker verification systems are widely implemented in voice-activated personal assistants, enhanced security systems, remote identity verification systems, and other devices for man-machine communication. Furthermore, a lot of innovations and techniques from deep learning have allowed to significantly increase verification accuracy by means of architectures such as x-vector embeddings, ECAPA-TDNN, and self-supervised speech representation models. Nevertheless, the performance of modern systems mostly suffers due to unfavorable conditions of actual operation, which include background noise, transmission channel inconsistencies, recording environment conditions, and speaker's health problems such as illness or tiredness. Many researchers currently concentrate their efforts on improving the representation learning of the speakers by means of advanced neural networks, self-supervised learning methods and finding voice biomarkers.

Keywords: speech; voice; self-supervised

Introduction

Speaker verification has become an essential biometric identification technique offering easy, contactless method for confirming personal identity. Use of this technology has increased significantly in various applications such as voice activated assistants, smart devices, banks, access control or remote identification system. Recent improvements in deep learning technology led to better efficiency of introducing speaker identification methods in practice because this technology enables creation of speaker models based on speech data. Many researchers have worked on enhancing speaker verification systems' reliability Chen and Jeng [2] developed lightweight CNN system for identification purposes. As for text-dependent verification, Kameda et al. [3] presented a technique for speaker identification based on the use of the SSI-DNNs. Their experiments showed that speaker identification can be achieved even with short speeches.
Although these studies have significantly improved verification accuracy and robustness, most of them address either noise, model efficiency, or spoofing independently. Comparatively fewer works investigate the combined influence of environmental noise and health-related voice variations, highlighting the need for more adaptive and health-aware speaker verification frameworks.

Related work

Table 1. Details of the previous research

Author / Year	Research Focus	Method / Model	Dataset	Key Contribution	Research Gap / Limitation
Lyu et al., 2025	Speaker verification and speaker tracking	Fast adaptation of pretrained speaker verification model	Source speaker tracking dataset	Rapid adaptation of pretrained verification models with minimal retraining.	Does not address background noise or health-related vocal changes.
Chen & Jeng, 2024	Text-independent speaker verification	Lightweight 3D CNN	Speech dataset	Reduced model complexity while maintaining competitive performance.	Limited evaluation under noisy and physiological variations.
Kameda et al., 2024	Text-dependent speaker verification	SSI-DNN	Short-utterance speech dataset	Reliable verification from short utterances.	Limited to text-dependent scenarios; illness effects not studied.
Farsiani et al., 2022	Text-independent speaker identification	End-to-end CNN	Speech dataset	Improved identification without handcrafted features.	Did not focus on noise or health robustness.
JIST, 2024	Speaker recognition in noisy conditions	Autoencoder-based denoising	Noisy speech corpus	Noise suppression while preserving speaker information.	Health-related voice variation not considered.
Sharma et al., 2020	Health-related speech analysis	Acoustic biomarker analysis	Coswara	Introduced speech dataset for respiratory disease research.	Designed for health analysis, not speaker verification.
Yamagishi et al., 2021	Secure speaker verification	Spoof detection benchmark	ASVspoof 2021	Standard benchmark for spoof/deepfake detection.	Natural illness-induced voice changes not addressed.
Heo et al., 2024	Deep speaker verification	NeXt-TDNN	Speaker verification benchmarks	Improved temporal feature extraction and embeddings.	No health-aware modeling.

Jung et al., 2024	Speaker verification toolkit	ESPnet-SPK	Multiple benchmarks	Reproducible end-to-end verification framework.	Health-aware verification not included.
Zhang et al., 2024	Speaker embedding extraction	MEConformer	Speaker verification datasets	High-quality embeddings using selective convolution.	Does not address health-induced or noisy speech jointly.

Recent advances in speaker verification have been driven by deep learning models that learn discriminative speaker embeddings directly from speech. Lyu *et al.* [1] demonstrated that adapting pretrained speaker verification models can efficiently support speaker tracking without extensive retraining. Lightweight architectures have also received attention, with Chen and Jeng [2] proposing a 3D convolutional network for text-independent verification, while Kameda *et al.* [3] showed that deep neural networks can achieve reliable text-dependent verification even from short utterances.

End-to-end learning has further improved speaker representation. Farsiani *et al.* [4] reported that convolutional neural networks outperform conventional handcrafted feature-based methods for text-independent speaker identification. To address adverse acoustic conditions, autoencoder-based denoising techniques have been explored to preserve speaker characteristics while reducing background interference [5].

The availability of large public datasets has also accelerated research in robust speech processing. The Coswara dataset has enabled investigations into the impact of respiratory health on speech characteristics [6], whereas ASVspoof 2021 has become a standard benchmark for evaluating resilience against spoofing and synthetic speech attacks [7].

More recently, advanced architectures such as NeXt-TDNN [8] and MEConformer [10] have achieved improved speaker embeddings by enhancing temporal feature extraction and representation learning. In addition, ESPnet-SPK [9] provides a comprehensive and reproducible framework that integrates self-supervised front-ends with state-of-the-art speaker verification models.

Method, Experiments and Results

The proposed framework aims to develop a speaker recognition system that can reliably identify individuals even when speech is affected by environmental noise or temporary health-related voice variations. Initially, acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs) and deep speech embeddings are extracted to capture speaker-specific information. These features are then processed using a ResNet-based deep neural encoder, which learns robust representations of the speaker's identity while reducing the influence of unwanted acoustic variations.

To further improve performance, a dynamic attention mechanism is incorporated to emphasize the most informative regions of the speech signal based on the input characteristics. In addition, a multi-branch learning framework with adversarial training is employed to encourage the model to focus on identity-related features while suppressing the effects of noise and health-induced voice changes. The proposed system will be trained and evaluated using VoxCeleb, Coswara, Saarbrücken Voice Database, and MUSAN datasets to assess its effectiveness under diverse acoustic and physiological conditions.

Conclusions

The structure created has the goal of producing a system of identifying speakers that allows identifying speaking people well. At the beginning of the development stage, the features of voice (like Mel-Frequency Cepstral Coefficients and deep speech embeddings) are being extracted and analyzed. After extracting the features, they are being transformed using the ResNet method, allowing the system to learn about its speaker identity effectively. However, outside sounds have a strong influence on the accuracy of the system. The introduction of the attention mechanism should allow defining the city of the sound signals.

References

1. X. Lyu, Y. Wang, T. Zhao and H. Liu, "Fast Adaptation of Pretrained Speaker Verification System for Source Speaker Tracking," *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, 2025, pp. 1-2, doi: 10.1109/ICASSP49660.2025.10889147.

2. J. -Y. Chen and J. -T. Jeng, "Text-Independent Speaker Verification Using Lightweight 3D Convolutional Neural Networks," *2024 International Conference on System Science and Engineering (ICSSE)*, Hsinchu, Taiwan, 2024, pp. 1-5, doi: 10.1109/ICSSE61472.2024.10608879.

3. K. Kameda, S. Tsuge, S. Kuroiwa, Y. Horiuchi and M. Nishida, "Text-Dependent Speaker Verification Using SSI-DNN Trained on Short Utterance," *2024 IEEE 13th Global Conference on Consumer Electronics (GCCE)*, Kitakyushu, Japan, 2024, pp. 808-810, doi: 10.1109/GCCE62371.2024.10760742.

4. Shabnam Farsiani, Habib Izadkhah, Shahriar Lotfi, An optimum end-to-end text-independent speaker identification system using convolutional neural network, Computers and Electrical Engineering, Volume 100, 2022,107882, ISSN 0045-7906.

5. Application of autoencoders in speaker recognition system in noisy environment." *Journal of Integrated Science & Technology*, 2024, 12(2), 742.

6. N. Sharma et al., “Coswara — A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis,” in Proc. Interspeech, 2020.

6. J. Yamagishi et al., “ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection,” arXiv preprint arXiv:2109.00537, 2021.

7. 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.

8. 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.

9. X. Zhang et al., “MEConformer: Highly Representative Embedding Extractor for Speaker Verification via Selective Convolution,” Expert Systems with Applications, vol. 244, 2024.