Multimodal Speaker Verification Robustness via Generative Audio Enhancement

Assignee Research

June 12, 2026

Abstract

Recent advancements in speaker verification techniques show promise, but their performance often deteriorates significantly in challenging acoustic environments. Although speech enhancement methods can improve perceived audio quality, they may unintentionally distort speaker-specific information, which can affect verification accuracy. This problem has become more noticeable with the increasing use of generative deep neural networks (DNNs) for speech enhancement. While these networks can produce intelligible speech even in conditions of very low signal-to-noise ratio (SNR), they may also sever

1 Introduction

This paper examines: A Framework for Robust Speaker Verification in Highly Noisy Environments Leveraging Both Noisy and Enhanced Audio. Research question: Does training multimodal speaker verification backbones with generatively enhanced audio improve robustness scores on the AVSpeech noisy test split compared to raw audio baselines?.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.4/10.

3 Results

8 papers retrieved. 24 claims extracted; 18 independently verified. Quality review score: 7.4/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The proposed method combines embeddings from noisy and enhanced audio to improve speaker verification in highly noisy en	✓	0.16
Generative DNNs for speech enhancement can produce superior speech quality but may distort speaker characteristics under	✓	0.20
The proposed framework uses a triplet loss function based on cosine distance for speaker verification.	✓	0.22
The proposed framework is lightweight and agnostic to specific speaker verification and speech enhancement techniques.	✓	0.29
The proposed framework outperforms other methods in severe noisy conditions where previous speaker verification methods	✓	0.20
The proposed framework reduces computation complexity compared to methods that employ a learning-based interpolation age	✓	0.18
The proposed framework delivers reliable speaker verification performance even in severe noisy conditions.	✓	0.15
The proposed framework uses a Siamese architecture to extract speaker embeddings from both noisy and enhanced speech.	×	0.13
The proposed framework combines speaker embeddings in a highly informative latent space.	✓	0.17
The proposed framework utilizes state-of-the-art speech enhancement techniques.	×	0.13
The proposed framework demonstrates superior performance in speaker verification under severe noise conditions.	✓	0.18
The proposed framework does not require training a dedicated speech enhancement module or speaker verification module.	✓	0.17
The proposed framework can use any pre-trained enhancement or verification module out-of-the-box.	✓	0.18
The proposed framework offers a more practical solution compared to methods that rely on training dedicated modules.	✓	0.17
The proposed framework achieves an EER of 13.17% for babble noise at 0 dB SNR, compared to 9.70% for noisy and 13.45% fo 4	×	0.12
The proposed framework achieves an EER of 15.67% for babble noise at -5 dB SNR, compared to 16.39% for noisy and 18.18%	×	0.13
The proposed framework achieves an EER of 19.31% for babble noise at -10 dB SNR, compared to 26.19% for noisy and 25.12%	✓	0.15
The proposed framework achieves an EER of	✓	0.16

References

- http://arxiv.org/abs/2508.18913v1
- http://arxiv.org/abs/2002.03562v2
- http://arxiv.org/abs/2304.03515v1