Diffusion-Based Speech Enhancement for Zero-Shot Speaker Verification on VoxCeleb

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: How does the integration of diffusion-based speech enhancement impact zero-shot speaker verification accuracy on the VoxCeleb benchmark compared to traditional enhancement methods under varying SNR. 12 claims were extracted from source literature; 6 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: A Framework for Robust Speaker Verification in Highly Noisy Environments Leveraging Both Noisy and Enhanced Audio. Research question: How does the integration of diffusion-based speech enhancement impact zero-shot speaker verification accuracy on the VoxCeleb benchmark compared to traditional enhancement methods under varying SNR conditions?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.5/10.

3 Results

9 papers retrieved. 12 claims extracted; 6 independently verified. Quality review score: 6.5/10.

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	✓	0.23
Generative DNNs for speech enhancement can produce superior speech quality but may distort speaker characteristics under	✓	0.21
The proposed framework uses a triplet loss function based on cosine distance for speaker verification.	×	0.05
The proposed framework is lightweight and agnostic to specific speaker verification and speech enhancement techniques.	✓	0.31
The proposed framework outperforms other methods in severe noisy conditions where previous speaker verification methods	×	0.13
The proposed framework reduces computation complexity compared to methods that employ a learning-based interpolation age	×	0.04
The proposed framework delivers reliable speaker verification performance in severe noisy conditions.	×	0.11
The proposed framework uses a Siamese architecture to extract speaker embeddings from both noisy and enhanced speech.	✓	0.18
The proposed framework combines speaker embeddings in a highly informative latent space.	×	0.10
The proposed framework leverages complementary information from both noisy and enhanced sources to enhance the robustnes	✓	0.24
The proposed framework uses state-of-the-art speech enhancement techniques.	✓	0.17
The proposed framework outperforms other methods in terms of EER (Equal Error Rate) for various noise types and SNR leve	×	0.06

References

- http://arxiv.org/abs/1706.08612v2
- http://arxiv.org/abs/2407.01939v2

- http://arxiv.org/abs/2508.18913v1