FlashSpeech vs. Diffusion-Based Models in Zero-Shot Speaker Verification on VoxCeleb

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 8 peer-reviewed papers addressing the following research question: How does FlashSpeech's zero-shot speaker verification accuracy compare to diffusion-based models on the VoxCeleb benchmark across varying signal-to-noise ratios. 11 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. 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 FlashSpeech's zero-shot speaker verification accuracy compare to diffusion-based models on the VoxCeleb benchmark across varying signal-to-noise ratios?.

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 4.5/10.

3 Results

8 papers retrieved. 11 claims extracted; 3 independently verified. Quality review score: 4.5/10.

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 framework combines speaker embeddings extracted from both noisy and enhanced speech sources.	✓	0.25
The proposed framework utilizes a Siamese architecture to extract embeddings from noisy and enhanced speech.	✓	0.17
The framework employs a triplet loss function based on cosine distance defined as L(A, P, N) = max(0, d(A, P) - d(A, N)	×	0.09
Unlike method [19], the proposed framework does not employ a learning-based interpolation agent to determine the linear	×	0.03
The proposed framework is agnostic to specific speaker verification and speech enhancement techniques, allowing the use	✓	0.23
Generative DNNs used for speech enhancement can lead to significant distortions of the speaker's intrinsic characteristi	×	0.14
On the VoxCeleb1 dataset with babble noise at -15 dB SNR, the SpeakerNet ECAPA-TDNN model using the proposed method ('Ou	×	0.05
On the VoxCeleb1 dataset with babble noise at -15 dB SNR, the proposed method (25.21% EER) outperformed the 'Noisy' base	×	0.04
On the VoxCeleb1 dataset with music noise at -20 dB SNR, the proposed method achieved an EER of 44.05%.	×	0.02
The t-SNE visualization in Figure 1 uses embeddings extracted from two speakers in the VoxCeleb1 dataset with babble noi	×	0.04
DeepFilterNet3 was used to generate the enhanced speech embeddings shown in Figure 1(b).	×	0.05

References

- http://arxiv.org/abs/1706.08612v2
- http://arxiv.org/abs/2508.18913v1
- http://arxiv.org/abs/2404.14700v4