Multimodal Model Audio Token Compression and Intent Classification Under Noise

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does varying audio token compression ratios in multimodal models impact intent classification accuracy on the SLUE-voicebank dataset under low signal-to-noise ratio conditions. 11 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.4/10. 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 varying audio token compression ratios in multimodal models impact intent classification accuracy on the SLUE-voicebank dataset under low signal-to-noise ratio conditions?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.4/10.

3 Results

15 papers retrieved. 11 claims extracted; 2 independently verified. Quality review score: 4.4/10.

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 framework combines speaker embeddings extracted from both noisy and enhanced speech using a Siamese archite	✓	0.32
The proposed framework utilizes a triplet loss function based on cosine distance defined as L(A, P, N) = max(0, d(A, P)	×	0.03
Unlike method [19], the proposed framework does not employ a learning-based interpolation agent to determine the optimal	×	0.04
The proposed framework is agnostic to specific speaker verification and speech enhancement techniques, allowing the use	✓	0.25
Generative DNNs used for speech enhancement can lead to significant distortions of the speaker's intrinsic characteristi	×	0.13
On the VoxCeleb1 dataset with babble noise at an SNR of -15 dB, the SpeakerNet ECAPA-TDNN model achieves an Equal Error	×	0.04
On the VoxCeleb1 dataset with babble noise at an SNR of -15 dB, the SpeakerNet ECAPA-TDNN model achieves an EER of 32.77	×	0.05
On the VoxCeleb1 dataset with babble noise at an SNR of -15 dB, the proposed method achieves an EER of 25.21% using the	×	0.03
On the VoxCeleb1 dataset with music noise at an SNR of -20 dB, the proposed method achieves an EER of 44.05% using the S	×	0.03
For the SpeakerNet ECAPA-TDNN model under babble noise, the proposed method yields lower EER values than both the 'Noisy	×	0.04
Figure 1 displays a t-SNE visualization of SpeakerNet embeddings for two speakers in the VoxCeleb1 dataset with babble n	×	0.03

References

- http://arxiv.org/abs/2505.07365v2
- http://arxiv.org/abs/2508.18913v1
- http://arxiv.org/abs/2111.10367v3