Self-Supervised Audio Representations vs. End-to-End Phoneme-Free Models in Low-Resource Code-Switched Speech Recognition

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How do self-supervised audio representations compare to end-to-end phoneme-free models in low-resource code-switched speech recognition accuracy. 11 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.6/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: M2S-AVSR: Modality-aware Multi-view Self-supervised Representation for Robust Audio-Visual Speech Recognition. Research question: How do self-supervised audio representations compare to end-to-end phoneme-free models in low-resource code-switched speech recognition accuracy?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.6/10.

3 Results

12 papers retrieved. 11 claims extracted; 3 independently verified. Quality review score: 6.6/10.

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
M2S-AVSR achieves up to 29.4% relative improvement under viewpoint perturbation and visual degradation settings on LRS3.	✓	0.27
M2S-AVSR achieves new state-of-the-art performance on the MISP2021-AVSR test set.	✓	0.25
M2S-AVSR achieves the best result in outdoor scenes on AISHELL8-RealScene.	✓	0.25
Deep learning has significantly advanced ASR systems, leading to strong performance under controlled conditions.	×	0.05
Robust speech recognition in real-world environments remains challenging due to background noise, reverberation, competi	×	0.14
Visual information, such as lip movements, can provide complementary cues when the acoustic signal is unreliable.	×	0.08
AVSR systems have achieved substantial improvements over audio-only counterparts, particularly in noisy environments.	×	0.04
Self-supervised learning (SSL) methods, such as wav2vec, WavLM, and Whisper, have substantially improved acoustic modeli	×	0.11
AV-HuBERT and related approaches learn robust visual speech representations from large-scale unlabeled audio-visual data	×	0.11
M2S-AV 600 achieves a score of 21.95 on LRS3+Vox2(En).	×	0.03
M2S-AVROVER 600 achieves a score of 18.82 on LRS3+Vox2(En).	×	0.03

References

- http://arxiv.org/abs/2606.05763v2
- http://arxiv.org/abs/2303.02719v2
- http://arxiv.org/abs/2007.04134v1