Architectural Modifications in M2S-AVSR for Low-Resource Audio-Visual Speech Recognition

Assignee Research

June 11, 2026

Abstract

Audio-Visual Speech Recognition (AVSR) enhances speech recognition robustness by leveraging visual cues, while real-world scenarios remain challenging due to viewpoint variation, audio distortion, and visual occlusion, which degrade modality quality and increase audio-visual asynchrony. In this paper, we propose a novel Modality-aware Multi-view Self-supervised representation framework for robust Audio-Visual Speech Recognition (M2S-AVSR). First, we introduce a multi-view representation learning encoder to learn view-invariant visual speech representations. Next, we employ a modality-aware mod

1 Introduction

This paper examines: M2S-AVSR: Modality-aware Multi-view Self-supervised Representation for Robust Audio-Visual Speech Recognition. Research question: What is the impact of architectural modifications in M2S-AVSR (e.g., attention mechanisms, fusion layers) on its scalability and accuracy in low-resource audio-visual speech recognition tasks, evaluated on AVSpeech and LRS3 benchmarks?.

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 7.4/10.

3 Results

15 papers retrieved. 11 claims extracted; 8 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
M2S-AVSR achieves up to 29.4% relative improvement under viewpoint perturbation and visual degradation settings on LRS3.	✓	0.28
M2S-AVSR achieves new state-of-the-art performance on the MISP2021-AVSR test set.	✓	0.26
M2S-AVSR achieves the best result in outdoor scenes on AISHELL8-RealScene.	✓	0.26
Deep learning has significantly advanced ASR systems, leading to strong performance under controlled conditions.	✓	0.17
Robust speech recognition in real-world environments remains challenging due to background noise, reverberation, competi	✓	0.22
Visual information, such as lip movements, can provide complementary cues when the acoustic signal is unreliable.	✓	0.22
AVSR systems have achieved substantial improvements over audio-only counterparts, particularly in noisy environments.	✓	0.18
Self-supervised learning (SSL) methods, such as wav2vec, WavLM, and Whisper, have substantially improved acoustic modeli	×	0.14
AV-HuBERT and related approaches learn robust visual speech representations from large-scale unlabeled audio-visual data	✓	0.24
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/2101.12729v1
- http://arxiv.org/abs/2606.05763v2
- http://arxiv.org/abs/2604.22203v1