Lightweight Speech Enhancement Integration in Multimodal LLMs: WER and Task Accuracy on Noisy Multi-Speaker Benchmarks

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does integrating lightweight speech enhancement modules like GTCRN into multimodal large language models affect word error rate and task accuracy on noisy multi-speaker benchmarks compared to. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Speech Enhancement for Wake-Up-Word detection in Voice Assistants. Research question: How does integrating lightweight speech enhancement modules like GTCRN into multimodal large language models affect word error rate and task accuracy on noisy multi-speaker benchmarks compared to end-to-end trained baselines?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.8/10.

3 Results

14 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 4.8/10.

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.

References

- http://arxiv.org/abs/2101.12732v1
- http://arxiv.org/abs/2409.16005v1
- http://arxiv.org/abs/2508.19583v2