Acoustic-Enhanced vs. Text-Only Models Across Language Families on MMSU Benchmarks

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does the performance gap between acoustic-enhanced and text-only models vary across different language families on MMSU tasks, and can this gap be reduced with cross-lingual acoustic feature. 10 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.1/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MMSU: A Massive Multi-task Spoken Language Understanding and Reasoning Benchmark. Research question: How does the performance gap between acoustic-enhanced and text-only models vary across different language families on MMSU tasks, and can this gap be reduced with cross-lingual acoustic feature adaptation?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.1/10.

3 Results

13 papers retrieved. 10 claims extracted; 1 independently verified. Quality review score: 5.1/10.

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
MMSU encompasses a wider range of acoustic features spanning 47 distinct tasks.	×	0.08
MMSU is the first benchmark to systematically incorporate linguistically grounded phenomena into spoken language underst	✓	0.21
MMSU requires models to integrate paralinguistic, phonetic, and semantic information for tasks such as sarcasm detection	×	0.09
MMSU includes 47 distinct tasks covering various linguistic phenomena and acoustic features.	×	0.11
MMSU includes tasks related to prosody, intonation, phonetics, rhetoric, syntactics, non-verbal, disfluency, and audio p	×	0.11
MMSU evaluates 22 models, including 12 Speech-LLMs and 10 Omni Large Language Models (OmniLLMs) with audio processing ca	×	0.08
The evaluation strategy for MMSU involves an audio clip and a text prompt, with the model choosing one of four options (	×	0.03
Answer options in MMSU are randomly ordered and balanced across the dataset to avoid positional bias.	×	0.02
All models in MMSU are evaluated with the same optimized instruction-following prompts to ensure fairness and minimize p	×	0.01
Qwen2.5-Omni-7B incorrectly identified the intonation of the sentence 'It's nice to meet you' as Fall-Rise Intonation in	×	0.02

References

- http://arxiv.org/abs/2506.04779v3

- http://arxiv.org/abs/2310.10378v5
- http://arxiv.org/abs/2207.08179v1