What is the comparative robustness of OpenPangu-MLA's multilingual paralinguistic feature extraction when eval

Assignee Research

June 10, 2026

Abstract

Speech inherently contains rich acoustic information that extends far beyond the textual language. In real-world spoken language understanding, effective interpretation often requires integrating semantic meaning (e.g., content), paralinguistic features (e.g., emotions, speed, pitch) and phonological characteristics (e.g., prosody, intonation, rhythm), which are embedded in speech. While recent multimodal Speech Large Language Models (SpeechLLMs) have demonstrated remarkable capabilities in processing audio information, their ability to perform fine-grained perception and complex reasoning in

1 Introduction

This paper examines: MMSU: A Massive Multi-task Spoken Language Understanding and Reasoning Benchmark. Research question: What is the comparative robustness of OpenPangu-MLA's multilingual paralinguistic feature extraction when evaluated against adversarial perturbations in the MMSU benchmark subsets?.

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 3.8/10.

3 Results

12 papers retrieved. 10 claims extracted; 1 independently verified. Quality review score: 3.8/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
MMSU encompasses a wider range of acoustic features spanning 47 distinct tasks.	×	0.10
MMSU increases reasoning complexity by requiring models to integrate paralinguistic, phonetic, and semantic information.	×	0.08
MMSU is the first benchmark to systematically incorporate linguistically grounded phenomena into spoken language underst	✓	0.21
MMSU was evaluated on 22 models, including 12 Speech-LLMs and 10 Omni Large Language Models (OmniLLMs) with audio proces	×	0.10
Each instance in MMSU consists of an audio clip and a text prompt, with the model choosing one of four options (A–D).	×	0.03
Answer options in MMSU are randomly ordered and balanced across the dataset to avoid potential positional bias.	×	0.02
All models in MMSU are evaluated with the same optimized instruction-following prompts to ensure fairness and minimize p	×	0.02
The sentence 'It's nice to meet you' is a common greeting that typically ends with a neutral or slightly falling intonat	×	0.01
The first part 'It's nice to meet you,' is spoken in a neutral tone, which is characteristic of a greeting.	×	0.02
The second part, 'you,' is spoken with a rising intonation.	×	0.05

References

- http://arxiv.org/abs/2306.07713v3

- http://arxiv.org/abs/2011.00577v3
- http://arxiv.org/abs/2506.04779v3