Audio Token Compression and Intent Classification Accuracy in Multimodal Speech Models

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 1 peer-reviewed paper addressing the following research question: What is the correlation between audio token compression ratios and intent classification accuracy for multimodal models evaluated on the SLUE-voicebank dataset under varying signal-to-noise ratios. 9 claims were extracted from source literature; 8 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Lost in Transcription, Found in Distribution Shift: Demystifying Hallucination in Speech Foundation Models. Research question: What is the correlation between audio token compression ratios and intent classification accuracy for multimodal models evaluated on the SLUE-voicebank dataset under varying signal-to-noise ratios?.

2 Methodology

Systematic literature search across multiple databases yielded 1 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.8/10.

3 Results

1 papers retrieved. 9 claims extracted; 8 independently verified. Quality review score: 6.8/10.

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
Speech foundation models trained at a massive scale, both in terms of model and data size, result in robust systems capa	✓	0.41
Traditional metrics like word error rate (WER) and character error rate (CER) are commonly used to evaluate ASR performa	✓	0.43
Hallucination in ASR models is especially concerning in high-stakes domains such as healthcare, legal, and aviation, whe	✓	0.29
The hallucination error rate (HER) is introduced as a metric to quantify hallucinations in ASR models.	✓	0.20
High WERs can mask low hallucination rates, while low WERs may conceal dangerous hallucinations.	✓	0.27
Synthetic noise, both adversarial and common perturbations like white noise, pitch shift, and time stretching, increase	✓	0.29
Distribution shift correlates strongly with HER (= 0.91).	✓	0.21
The analysis includes over 20 ASR models.	×	0.12
The findings highlight the importance of incorporating HER alongside traditional metrics like WER to better assess ASR m	✓	0.33

References

- https://doi.org/10.18653/v1/2025.findings-acl.1190