SLAM-ASR Models with Whisper and HuBERT Encoders on Limited Data across Low-Resource Languages in the MIRACL Benchmark

Assignee Research

June 22, 2026

Abstract

Large language models (LLMs) have demonstrated potential in handling spoken inputs for high-resource languages, reaching state-of-the-art performance in various tasks. However, their applicability is still less explored in low-resource settings. This work investigates the use of Speech LLMs for low-resource Automatic Speech Recognition using the SLAM-ASR framework, where a trainable lightweight projector connects a speech encoder and a LLM. Firstly, we assess training data volume requirements to match Whisper-only performance, re-emphasizing the challenges of limited data. Secondly, we show th

1 Introduction

This paper examines: Speech LLMs in Low-Resource Scenarios: Data Volume Requirements and the Impact of Pretraining on High-Resource Languages. Research question: How does the performance of SLAM-ASR models with Whisper/HuBERT encoders vary when fine-tuned on a limited subset of speech data (e.g., 10K hours) across different low-resource languages in the MIRACL benchmark compared to high-resource languages like English?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.3/10.

3 Results

9 papers retrieved. 15 claims extracted; 12 independently verified. Quality review score: 7.3/10.

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
Increasing the quantity of training data consistently improves the overall performance of the SLAM-ASR model, regardless	✓	0.24
EuroLLM 1.7B consistently outperforms Salamandra 2B in the SLAM-ASR framework.	×	0.14
The performance gap between Salamandra and EuroLLM tends to close as more data becomes available.	✓	0.16
The SLAM-ASR framework with EuroLLM 1.7B and 200 or 252 hours of training data obtains a WER of 6.4% and 6.1% respective	✓	0.31
The SLAM-ASR framework does not outperform a Whisper-large- or Whisper-large-v3-turbo-only set-up on Fleurs IT (WER = 4.	✓	0.27
With 200 hours of CV IT training data with EuroLLM 1.7B, the WER on CV IT is 6.4% but 13.2% on FL IT.	✓	0.26
LoRA fine-tuning of the LLM can improve the alignment between speech and text tokens.	✓	0.17
The SLAM-ASR framework struggles with generalizing across domains.	×	0.11
The speech encoder may learn less robust representations in low-resource settings.	✓	0.18
The LLM may generate less accurate transcripts due to limited exposure to the target language in low-resource settings.	✓	0.23
The projector may struggle to effectively align the different modality embeddings in low-resource settings.	✓	0.18
In low-resource settings, there is often little to no data available for ASR and other speech-related tasks such as spok	✓	0.29
The research investigates the viability of Speech LLMs in low-resource scenarios using ASR as a case study.	×	0.13
The research questions (RQ1 and RQ2) are addressed by progressively increasing the amount of data used to train a linear	✓	0.18
The research explores the effects of leveraging a projector pretrained on English data (Librispeech 100 and 100 hours of	✓	0.29

References

- http://arxiv.org/abs/2501.05976v1
- http://arxiv.org/abs/2508.05149v1
- http://arxiv.org/abs/2603.27981v1