Universal Speech Model vs. Monolingual ASR: WER and Sample Efficiency in Low-Resource Languages

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does the performance of USM's multilingual ASR compare to specialized monolingual models on benchmark datasets like LibriSpeech and Common Voice when evaluated for word error rate (WER) and. 9 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: UML: A Universal Monolingual Output Layer for Multilingual ASR. Research question: How does the performance of USM's multilingual ASR compare to specialized monolingual models on benchmark datasets like LibriSpeech and Common Voice when evaluated for word error rate (WER) and sample efficiency in low-resource languages?.

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 5.5/10.

3 Results

13 papers retrieved. 9 claims extracted; 3 independently verified. Quality review score: 5.5/10.

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
UML is a monolingual output layer shared by all languages.	✓	0.28
In UML, each output node o is mapped to L different monolingual WPMs (W1,o, . . . , WL,o) for L different languages.	×	0.13
In a conventional output layer, each output node is mapped to only one WPM.	✓	0.18
UML uses only one H $\times$ max(V1, . . . , VL)-dimensional output layer to model the sum of Vl WPMs across L languages.	×	0.09
The method using a conventional output layer for all multilingual WPMs requires an H $\times$ (sum of Vl)-dimensional layer.	×	0.14
Methods using L separate monolingual output layers require H $\times$ (sum of Vl) parameters to model the total WPMs.	×	0.10
In UML, each WPM is determined jointly by the Language ID (LID) and the output node index.	×	0.11
UML enables the monolingual ASR decoder structure to be used for multilingual ASR.	✓	0.16
Although UML is introduced for WPMs, it is applicable to other kinds of subword units.	×	0.09

References

- http://arxiv.org/abs/2304.00649v1
- http://arxiv.org/abs/2410.07400v2
- http://arxiv.org/abs/2302.11186v1