CUPE Universal Phoneme Encoder Substitution for Low-Resource Word Error Rate Reduction in Common Voice

Assignee Research

June 11, 2026

Abstract

Word-piece models (WPMs) are commonly used subword units in state-of-the-art end-to-end automatic speech recognition (ASR) systems. For multilingual ASR, due to the differences in written scripts across languages, multilingual WPMs bring the challenges of having overly large output layers and scaling to more languages. In this work, we propose a universal monolingual output layer (UML) to address such problems. Instead of one output node for only one WPM, UML re-associates each output node with multiple WPMs, one for each language, and results in a smaller monolingual output layer shared acros

1 Introduction

This paper examines: UML: A Universal Monolingual Output Layer for Multilingual ASR. Research question: What is the impact of replacing language-specific output layers with the CUPE universal phoneme encoder on word error rate metrics for low-resource languages in the Common Voice dataset?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.0/10.

3 Results

14 papers retrieved. 10 claims extracted; 9 independently verified. Quality review score: 8.0/10.

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.	✓	0.21
In UML, each output node o is mapped to L different monolingual WPMs (W1,o, . . . , WL,o) for L different languages.	✓	0.26
In a conventional output layer, each output node is mapped to only one WPM.	✓	0.19
UML uses only one H $\times$ max(V1, . . . , VL)-dimensional output layer to model the sum of Vl WPMs across L languages.	✓	0.26
Methods using a conventional output layer for all multilingual WPMs require an H $\times$ (sum of Vl)-dimensional layer.	✓	0.19
Methods using L separate monolingual output layers require H $\times$ (sum of Vl) parameters to model the total WPMs.	×	0.15
In UML, each WPM is determined jointly by the Language ID (LID) and the output node index.	✓	0.19
LIDs need to be taken into account during inference in UML.	✓	0.17
UML enables the monolingual ASR decoder structure to be used for multilingual ASR.	✓	0.21
Although UML is introduced for WPMs, it is applicable to other kinds of subword units.	✓	0.19

References

- http://arxiv.org/abs/2508.15316v1
- http://arxiv.org/abs/2302.11186v1
- http://arxiv.org/abs/2304.00649v1