To what extent does the volume of target-language unlabeled data in self-supervised pre-training correlate wit

Assignee Research

June 10, 2026

Abstract

Automatic speech recognition for low-resource languages remains fundamentally constrained by the scarcity of labeled data and computational resources required by state-of-the-art models. We present a systematic investigation into cross-lingual continuous pretraining for low-resource languages, using Perso-Arabic languages (Persian, Arabic, and Urdu) as our primary case study. Our approach demonstrates that strategic utilization of unlabeled speech data can effectively bridge the resource gap without sacrificing recognition accuracy. We construct a 3,000-hour multilingual corpus through a scala

1 Introduction

This paper examines: Efficient ASR for Low-Resource Languages: Leveraging Cross-Lingual Unlabeled Data. Research question: To what extent does the volume of target-language unlabeled data in self-supervised pre-training correlate with downstream ASR accuracy on low-resource dialects?.

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 3.8/10.

3 Results

13 papers retrieved. 19 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
Strategic utilization of cross-lingual unlabeled data can effectively overcome resource constraints without sacrificing	✓	0.23
The pretraining corpus for Persian consists of 850 hours of audio.	×	0.05
The pretraining corpus for Arabic consists of 1350 hours of audio.	×	0.05
The pretraining corpus for Urdu consists of 850 hours of audio.	×	0.05
The pretraining corpus for Urdu consists of 816 hours of audio after VAD-based segmentation.	×	0.04
The pretraining corpus for Persian consists of 878 hours of audio after VAD-based segmentation.	×	0.04
The pretraining corpus for Arabic consists of 1,310 hours of audio after VAD-based segmentation.	×	0.04
The labeled train set for Urdu consists of 60 hours of audio.	×	0.03
The labeled test set for Urdu consists of 10 hours of audio.	×	0.03
The labeled train set for Persian consists of 69 hours of audio.	×	0.03
The labeled test set for Persian consists of 11 hours of audio.	×	0.03
The labeled train set for Arabic consists of 74 hours of audio.	×	0.03
The labeled test set for Arabic consists of 11 hours of audio.	×	0.04
The multilingual corpus consists of 3,000 hours of audio.	×	0.06
Wav2Vec 2.0 Base has 95M parameters.	×	0.05
XLS-R has 300M parameters.	×	0.05
Wav2Vec 2.0 Large has 300M parameters.	×	0.06
XLS-R was pretrained on 436K hours across 128 languages.	×	0.02
Wav2Vec 2.0 Large was pretrained on 65K hours of primarily English speech.	×	0.05

References

- http://arxiv.org/abs/2512.07277v1
- http://arxiv.org/abs/2502.04883v1
- http://arxiv.org/abs/2109.14357v1