How does cross-lingual self-supervised pre-training impact convergence speed and word error rate on code-switc

Assignee Research

June 10, 2026

Abstract

Self-supervised pre-training could effectively improve the performance of low-resource automatic speech recognition (ASR). However, existing self-supervised pre-training are task-agnostic, i.e., could be applied to various downstream tasks. Although it enlarges the scope of its application, the capacity of the pre-trained model is not fully utilized for the ASR task, and the learned representations may not be optimal for ASR. In this work, in order to build a better pre-trained model for low-resource ASR, we propose a pre-training approach called wav2vec-S, where we use task-specific semi-supe

1 Introduction

This paper examines: Wav2vec-S: Semi-Supervised Pre-Training for Low-Resource ASR. Research question: How does cross-lingual self-supervised pre-training impact convergence speed and word error rate on code-switched ASR benchmarks compared to monolingual initialization?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.7/10.

3 Results

16 papers retrieved. 11 claims extracted; 6 independently verified. Quality review score: 6.7/10.

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
wav2vec-S achieves an average relative WER reduction of 24.5% for 1h fine-tuning.	✓	0.15
wav2vec-S achieves an average relative WER reduction of 6.6% for 10h fine-tuning.	✓	0.15
Semi-supervised pre-training can improve the performance and generalization of the self-supervised pre-trained model, i.	✓	0.31
Character-level supervision is better than phone-level for monolingual semi-supervised pre-training even on a cross-ling	×	0.10
Monolingual semi-supervised pre-training has a trade-off between performance of the source language and other languages.	×	0.09
The semi-supervised pre-training step costs much less time than self-supervised pre-training.	×	0.14
Semi-supervised pre-training effectively improves different self-supervised pre-trained models, e.g., wav2vec 2.0 [1], d	✓	0.19
Semi-supervised pre-training closes the representation gap between the pre-trained and fine-tuned models.	✓	0.19
The pre-training (source) dataset is LibriSpeech [28], where the 100h clean subset is used.	×	0.04
The semi-supervised pre-training is also learning task-specific representations.	✓	0.23
The two steps in wav2vec-S, i.e., self-supervised and semi-supervised pre-training, are loosely coupled.	×	0.12

References

- http://arxiv.org/abs/2110.04484v2

- http://arxiv.org/abs/2601.20896v2
- http://arxiv.org/abs/2509.12101v1