Comparative Analysis of Task-Specific and Task-Agnostic Self-Supervised Pre-Training for Low-Resource ASR on LibriSpeech

Assignee Research

June 13, 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 the accuracy of task-specific self-supervised pre-training for low-resource ASR compare to task-agnostic models when evaluated on the LibriSpeech benchmark with varying amounts of labeled training data?.

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 7.4/10.

3 Results

13 papers retrieved. 12 claims extracted; 9 independently verified. Quality review score: 7.4/10.

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 uses task-specific semi-supervised pre-training to refine self-supervised pre-trained models for the ASR task.	✓	0.30
Wav2vec-S requires only a marginal increment of pre-training time compared to vanilla self-supervised pre-training.	✓	0.23
Wav2vec-S achieves an average relative WER reduction of 24.5% for 1-hour fine-tuning scenarios.	×	0.11
Wav2vec-S achieves an average relative WER reduction of 6.6% for 10-hour fine-tuning scenarios.	×	0.10
Wav2vec-S improves ASR performance on in-domain, cross-domain, and cross-lingual datasets.	✓	0.20
Canonical correlation analysis (CCA) shows that semi-supervised pre-training closes the representation gap between self-	✓	0.24
Character-level supervision yields better performance than phone-level supervision for monolingual semi-supervised pre-t	✓	0.23
Monolingual semi-supervised pre-training exhibits a trade-off where increased training updates improve source language p	✓	0.20
The semi-supervised pre-training step in wav2vec-S costs significantly less time than the self-supervised pre-training s	✓	0.22
Semi-supervised pre-training effectively improves performance when applied to different self-supervised pre-trained mode	✓	0.19
The pre-training source dataset used in the study is the LibriSpeech 100h clean subset.	✓	0.16
Wav2vec-S utilizes unlabeled data through pseudo-labeling during the semi-supervised pre-training phase.	×	0.15

References

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

- http://arxiv.org/abs/2109.14357v1
- http://arxiv.org/abs/2208.05445v1