Acoustic Diversity Impact on Self-Supervised Speech Model Accuracy

Assignee Research

July 19, 2026

Abstract

Self-supervised learning (SSL) has transformed speech processing, yet its reliance on massive pre-training datasets remains a bottleneck. While robustness is often attributed to scale and diversity, the role of the data distribution is less understood. We systematically examine how curated subsets of pre-training data influence Automatic Speech Recognition (ASR) performance. Surprisingly, optimizing for acoustic, speaker, or linguistic diversity yields no clear improvements over random sampling. Instead, we find that prioritizing the longest utterances achieves superior ASR results while using

1 Introduction

This paper examines: A Study of Data Selection Strategies for Pre-training Self-Supervised Speech Models. Research question: What is the effect of varying the acoustic diversity in pre-training data on the accuracy of self-supervised speech models, as measured by WER on LibriSpeech and Common Voice benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.5/10.

3 Results

12 papers retrieved. 16 claims extracted; 14 independently verified. Quality review score: 8.5/10.

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
Diversity-based sampling methods did not yield significant improvements over the either baseline.	✓	0.20
The speaker-based method on the large split achieved a word error rate of 17.97 on the test set.	✓	0.28
The random baseline achieved a word error rate of 18.54 on the test set.	✓	0.19
The all baseline achieved a word error rate of 18.08 on the test set.	✓	0.18
Utterance length–based methods consistently outperformed both baselines.	✓	0.19
On the medium split, selecting the longest utterances reduced the word error rate to 19.02 on the test set.	✓	0.27
Combining speaker diversity with utterance length further improved performance to 18.97 on the test set.	✓	0.23
On the large split, length-based sampling methods reached 17.77 and 17.42 word error rates, respectively.	✓	0.27
The combined method performed best overall.	×	0.14
Training times for the medium split were similar across all methods.	✓	0.17
On the large split, length-based sampling substantially reduced pre-training time by about 24% compared to the all basel	✓	0.24
The best performing models were pre-trained on subsets consisting of longer utterances.	✓	0.19
The distributions of utterance lengths for the best performing models differ most from the fine-tuning data.	✓	0.18
Enforcing diversity in the pre-training data—whether acoustic, speaker, or linguistic—yields no significant improvement	✓	0.24
Selecting the longest utterances for pre-training achieves superior ASR performance while requiring only half of the ori	✓	0.24
Length-based sampling reduces pre-training time by 24% on a large-scale corpus.	×	0.14

References

- http://arxiv.org/abs/2604.22203v1
- http://arxiv.org/abs/2601.20896v2
- http://arxiv.org/abs/2109.14357v1