Multilingual Self-Supervised Pre-Training Accelerates Code-Switched ASR Convergence

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: Does pre-training on multilingual self-supervised audio models improve convergence speed for code-switched ASR tasks compared to monolingual initialization. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Comparison of Self-Supervised Speech Pre-Training Methods on Flemish Dutch. Research question: Does pre-training on multilingual self-supervised audio models improve convergence speed for code-switched ASR tasks compared to monolingual initialization?.

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 3.8/10.

3 Results

14 papers retrieved. 13 claims extracted; 0 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
APC uses a Filterbank feature encoder, GRU aggregator, and reconstructs future frames with an output dimension of 512 an	×	0.01
Mockingjay uses a Filterbank feature encoder, Bidirectional Transformer aggregator, and reconstructs masked frames with	×	0.01
CPC uses a CNN feature encoder, LSTM aggregator, and identifies future features with an output dimension of 256 and 1.8M	×	0.02
wav2vec uses a CNN feature encoder, CNN aggregator, and identifies future features with an output dimension of 512 and 3	×	0.02
wav2vec 2.0 uses a CNN feature encoder, Transformer aggregator, and identifies quantised future features with output dim	×	0.02
wav2vec 2.0's encoder computes latent speech representations from the raw waveform with 7 temporal convolution blocks.	×	0.12
wav2vec 2.0 masks a certain proportion of the latent features before feeding to the aggregator.	×	0.01
wav2vec 2.0 uses a quantisation module to map latent feature vectors to discretised versions.	×	0.01
The final training objective of wav2vec 2.0 is to distinguish the true quantised representation for a masked time step,	×	0.03
wav2vec 2.0 has two architectures: base with 12 Transformer blocks and large with 24 Transformer blocks in the aggregato	×	0.01
The contextual features at the output of the wav2vec 2.0 aggregator are extracted for downstream tasks.	×	0.08
wav2vec 2.0 duplicates contextual features in time to mimic a stride of 10ms instead of 20ms.	×	0.01
wav2vec 2.0 can be fine-tuned on a labelled set by adding an extra linear layer on top of the context network and applyi	×	0.05

References

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

- http://arxiv.org/abs/2007.04134v1
- http://arxiv.org/abs/2401.00273v1