Comparison of Self-Supervised Speech Model Inference Efficiency in Flemish Dutch and English for ASR Tasks

Assignee Research

July 23, 2026

Abstract

Recent research in speech processing exhibits a growing interest in unsupervised and self-supervised representation learning from unlabelled data to alleviate the need for large amounts of annotated data. We investigate several popular pre-training methods and apply them to Flemish Dutch. We compare off-the-shelf English pre-trained models to models trained on an increasing amount of Flemish data. We find that the most important factors for positive transfer to downstream speech recognition tasks include a substantial amount of data and a matching pre-training domain. Ideally, we also finetune

1 Introduction

This paper examines: Comparison of Self-Supervised Speech Pre-Training Methods on Flemish Dutch. Research question: How does the inference efficiency of self-supervised speech models pre-trained on Flemish Dutch compare to English pre-trained models when evaluated on downstream automatic speech recognition (ASR) tasks using metrics like Word Error Rate (WER) and real-time factor?.

2 Methodology

Systematic literature search across multiple databases yielded 7 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

7 papers retrieved. 12 claims extracted; 9 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
APC uses a GRU aggregator to reconstruct future frames with an output dimension of 512 and 4.1M parameters.	×	0.14
Mockingjay uses a bidirectional Transformer aggregator to reconstruct masked frames with an output dimension of 768 and	✓	0.16
CPC uses an LSTM aggregator to identify future features with an output dimension of 256 and 1.8M parameters.	✓	0.17
wav2vec uses a CNN aggregator to identify future features with an output dimension of 512 and 32.5M parameters.	✓	0.18
wav2vec 2.0 uses a Transformer aggregator to identify quantised future features with output dimensions of 768 (base) and	✓	0.23
wav2vec 2.0 combines ideas from wav2vec, vq-wav2vec, and MLM.	×	0.14
The wav2vec 2.0 encoder computes latent speech representations from the raw waveform with 7 temporal convolution blocks.	✓	0.17
A quantisation module in wav2vec 2.0 maps the latent feature vectors to discretised versions.	✓	0.19
The final training objective of wav2vec 2.0 is to distinguish the true quantised representation for a masked time step,	✓	0.23
The base architecture of wav2vec 2.0 contains 12 Transformer blocks in the aggregator, while the large architecture cont	×	0.10
The contextual features at the output of the wav2vec 2.0 aggregator are extracted for downstream tasks.	✓	0.16
The wav2vec 2.0 model can be fine-tuned on a labelled set by adding an extra linear layer on top of the context network	✓	0.16

References

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

- http://arxiv.org/abs/2303.06740v1