How does the scalability of M2S-AVSR compare to contrastive audio-visual representation learning methods (e.g.

Assignee Research

June 10, 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 scalability of M2S-AVSR compare to contrastive audio-visual representation learning methods (e.g., AV-Contrast) in terms of throughput and accuracy on the AVSpeech benchmark when trained on varying dataset sizes?.

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 3.8/10.

3 Results

16 papers retrieved. 12 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	×	0.01
wav2vec uses a CNN feature encoder, CNN aggregator, and identifies future features with an output dimension of 512 and 3	×	0.01
wav2vec 2.0 uses a CNN feature encoder, Transformer aggregator, and identifies quantised future features with output dim	×	0.01
wav2vec 2.0 encoder computes latent speech representations from the raw waveform with 7 temporal convolution blocks.	×	0.02
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,	×	0.04
wav2vec 2.0 has base and large architectures with 12 and 24 Transformer blocks in the aggregator, respectively.	×	0.03
Contextual features at the output of the wav2vec 2.0 aggregator are extracted for downstream tasks.	×	0.04
wav2vec 2.0 model can be fine-tuned on a labelled set by adding an extra linear layer on top of the context network and	×	0.02

References

- http://arxiv.org/abs/2506.09781v2
- http://arxiv.org/abs/2109.14357v1

- http://arxiv.org/abs/2106.07732v2