Efficient Fine-Tuning of English Speech Models for Flemish Dutch Performance

Assignee Research

July 19, 2026

Abstract

With excellent generalization ability, self-supervised speech models have shown impressive performance on various downstream speech tasks in the pre-training and fine-tuning paradigm. However, as the growing size of pre-trained models, fine-tuning becomes practically unfeasible due to heavy computation and storage overhead, as well as the risk of overfitting. Adapters are lightweight modules inserted into pre-trained models to facilitate parameter-efficient adaptation. In this paper, we propose an effective adapter framework designed for adapting self-supervised speech models to the speaker ve

1 Introduction

This paper examines: Efficient Adapter Tuning of Pre-trained Speech Models for Automatic Speaker Verification. Research question: Can efficient fine-tuning techniques like adapter modules or parameter-efficient tuning improve the downstream task performance of English pre-trained speech models on Flemish Dutch while reducing compute costs, measured by WER and training throughput on LibriSpeech?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.5/10.

3 Results

10 papers retrieved. 14 claims extracted; 10 independently verified. Quality review score: 7.5/10.

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
The VoxCeleb1 dataset contains 148,642 utterances from 1,211 speakers in the development set and 4,874 utterances from 4	✓	0.31
The 1st48-UTD forensic corpus has a total duration of 3.5 hours and more than 50% of utterances are shorter than 2 secon	✓	0.22
The training set of the 1st48-UTD forensic corpus consists of 3,755 utterances from 228 speakers, and the test set conta	✓	0.29
The pre-trained WavLM Base+ model comprises a convolutional feature encoder and 12 Transformer blocks equipped with gate	✓	0.30
The WavLM Base+ model has 94.70 million parameters.	✓	0.16
The speaker verification backend is composed of an average time pooling layer and two FC layers, with an embedding size	✓	0.24
The Inner-layer Adapter consists of two FC layers with a bottleneck dimension of 256, a ReLU activation function between	✓	0.33
The Inter-layer Adapter consists of a FC layer with a hidden dimension of 512, followed by a ReLU activation function an	✓	0.32
The models are trained with the cross-entropy loss using the Adam optimizer with an initial learning rate of 5e-4 for SV	✓	0.28
The learning rates decrease to 2.5e-5 for SV backend and 5e-7 for all other parameters after the first 38k steps.	✓	0.27
The proposed method achieves an EER of 16.75% and a minDCF of 0.734 on the VoxCeleb1 dataset.	×	0.08
Full fine-tuning achieves an EER of 19.13% and a minDCF of 0.842 on the VoxCeleb1 dataset.	×	0.08
Linear probing achieves an EER of 18.52% and a minDCF of 0.829 on the VoxCeleb1 dataset.	×	0.10
Weighted sum method achieves an EER of 18.10% and a minDCF of 0.819 on the VoxCeleb1 dataset.	×	0.12

References

- http://arxiv.org/abs/2109.14357v1
- http://arxiv.org/abs/2606.01947v1
- http://arxiv.org/abs/2403.00293v1