Neural Discriminative Backend Replacing Generative PLDA for Low-Resource Speaker Verification with Self-Supervised Representations

Assignee Research

June 13, 2026

Abstract

Wav2vec 2.0 is a recently proposed self-supervised framework for speech representation learning. It follows a two-stage training process of pre-training and fine-tuning, and performs well in speech recognition tasks especially ultra-low resource cases. In this work, we attempt to extend self-supervised framework to speaker verification and language identification. First, we use some preliminary experiments to indicate that wav2vec 2.0 can capture the information about the speaker and language. Then we demonstrate the effectiveness of wav2vec 2.0 on the two tasks respectively. For speaker verif

1 Introduction

This paper examines: Exploring wav2vec 2.0 on speaker verification and language identification. Research question: What is the impact of replacing the generative PLDA scoring module with a neural discriminative backend on the equal error rate (EER) for low-resource language speaker verification using self-supervised speech representations?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.8/10.

3 Results

13 papers retrieved. 14 claims extracted; 11 independently verified. Quality review score: 7.8/10.

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
The model achieves an Equal Error Rate (EER) of 3.61% on the VoxCeleb1 dataset for speaker verification.	✓	0.23
The model achieves an Equal Error Rate (EER) of 12.02% on the 1 second condition of the AP17-OLR dataset for language id	✓	0.32
The model achieves an Equal Error Rate (EER) of 3.47% on the full-length condition of the AP17-OLR dataset for language	✓	0.31
The VoxCeleb1 dataset contains over 100,000 utterances from 1,251 celebrities.	✓	0.20
The AP17-OLR dataset consists of 10 different languages.	✓	0.16
The duration of training data for each language in the AP17-OLR dataset is about 10 hours.	✓	0.18
The test set of the AP17-OLR dataset contains three subsets with different durations (1 second, 3 second, and full lengt	✓	0.25
The base model (M-nonetune) is pre-trained on the Librispeech corpus.	✓	0.17
M-sv is fine-tuned on the VoxCeleb1 dataset for speaker verification.	×	0.10
M-lid is fine-tuned on the AP17-OLR dataset for language identification.	×	0.13
M-multi is fine-tuned on the AP17-OLR and VoxCeleb1 datasets simultaneously in a multi-task form.	×	0.14
The wav2vec 2.0 model consists of a CNN-based feature encoder, a Transformer-based context network, and a quantization m	✓	0.26
The CNN encoder in wav2vec 2.0 stacks seven blocks with 512 channels and strides (5, 2, 2, 2, 2, 2, 2) and kernel widths	✓	0.24
The context network in wav2vec 2.0 stacks 12 Transformer blocks with model dimension 768, inner dimension 3,072, and 8 a	✓	0.29

References

- `http://arxiv.org/abs/2012.06185v2`

- http://arxiv.org/abs/2002.03562v2
- http://arxiv.org/abs/2001.07034v2