Cross-Lingual Transferability of Wav2Vec 2.0 vs. Traditional Speaker Verification in Low-Resource Languages

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 7 peer-reviewed papers addressing the following research question: How does the cross-lingual transferability of wav2vec 2.0 compare to traditional speaker verification methods when evaluated on low-resource languages using metrics such as EER and minDCF on datasets. 8 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: NPLDA: A Deep Neural PLDA Model for Speaker Verification. Research question: How does the cross-lingual transferability of wav2vec 2.0 compare to traditional speaker verification methods when evaluated on low-resource languages using metrics such as EER and minDCF on datasets like Common Voice or VoxLingua107?.

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 4.8/10.

3 Results

7 papers retrieved. 8 claims extracted; 2 independently verified. Quality review score: 4.8/10.

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 proposed neural net architecture is compared with discriminative PLDA, pairwise Gaussian backend, and a baseline sys	×	0.14
For all the pairwise generative/discriminative models, the backend is trained using randomly sampled target and non-targ	×	0.05
Experiments are performed by sampling trials from the clean VoxCeleb segments and the augmented set, with about 6.6 mill	×	0.02
The Kaldi implementation of the PLDA backend models the average embedding x-vector of each training speaker, with x-vect	×	0.05
The Gaussian backend is trained on the same pairs of target and non-target x-vector trials, after centering, LDA, and le	×	0.04
The discriminative PLDA (DPLDA) model is trained on processed x-vectors, with a portion of the Voxceleb training trials	×	0.04
The proposed neural PLDA (NPLDA) model operates on pairs of x-vector embeddings and outputs a score to decide target ver	✓	0.20
Experiments on the SITW dataset and the VOiCES development and evaluation datasets show that the proposed NPLDA approach	✓	0.20

References

- http://arxiv.org/abs/2002.03562v2
- http://arxiv.org/abs/2412.10008v1
- http://arxiv.org/abs/2111.02735v3