Cross-lingual vs. monolingual model performance in accented speech recognition

Assignee Research

June 28, 2026

Abstract

Utilizing Self-Supervised Learning (SSL) models for Speech Emotion Recognition (SER) has proven effective, yet limited research has explored cross-lingual scenarios. This study presents a comparative analysis between human performance and SSL models, beginning with a layer-wise analysis and an exploration of parameter-efficient fine-tuning strategies in monolingual, cross-lingual, and transfer learning contexts. We further compare the SER ability of models and humans at both utterance- and segment-levels. Additionally, we investigate the impact of dialect on cross-lingual SER through human eva

1 Introduction

This paper examines: Cross-Lingual Speech Emotion Recognition: Humans vs. Self-Supervised Models. Research question: What is the comparative performance of cross-lingual transfer from English self-supervised models versus monolingual Flemish models on accented speech recognition tasks?.

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 9.0/10.

3 Results

7 papers retrieved. 14 claims extracted; 14 independently verified. Quality review score: 9.0/10.

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 same amount of data is used for CN, DE, and EN to reduce the effect of varying training data sizes.	✓	0.17
An equal number of samples for each emotion is used to ensure a balanced emotion distribution.	✓	0.18
For ESD, PAVOQUE, and IEMOCAP, 5-fold cross-validation is applied for model training with 400 utterances per emotion cat	✓	0.28
200 utterances are randomly selected for validation and test sets, respectively.	✓	0.24
12 sentences per emotion category, totaling 144 utterances for all languages (12 sentences $\times$ 4 emotions $\times$ 3 datasets) ar	✓	0.29
For SED, 8 utterances per emotion are randomly selected from ZED, totaling 24 utterances, for comparison with human eval	✓	0.29
A classification head projecting from dimension 768 to 4 for SER is used with a learning rate of 1e-4, epsilon of 1e-8,	✓	0.39
Cross-entropy is used as the loss criterion and training stops if the validation loss does not decrease for 10 consecuti	✓	0.31
For PEFT strategies, the same classification head configuration as in the layer-wise analysis is used.	✓	0.18
For the LoRA module, the attention head is set to 8, alpha for scaling is 16, with a dropout rate of 0.1.	✓	0.27
For the BA module, the reduction factor is 16.	✓	0.17
Models are trained for 100 epochs with a batch size of 16 for PEFT strategies.	✓	0.26
In the monolingual setting, both the CN and DE models demonstrate strong performance on their respective source language	✓	0.25
In the cross-lingual setting, both models show a significant drop in accuracy.	✓	0.22

References

- http://arxiv.org/abs/2409.16920v2

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