Cross-Domain Generalization of MWER-Trained ASR Models Evaluated via Intent Classification F1 Scores

Assignee Research

June 11, 2026

Abstract

This study presents a domain adaptation approach for speaker diarization targeting conversational Indonesian audio. We address the challenge of adapting an English-centric diarization pipeline to a low-resource language by employing synthetic data generation using neural Text-to-Speech technology. Experiments were conducted with varying training configurations, a small dataset (171 samples) and a large dataset containing 25 hours of synthetic speech. Results demonstrate that the baseline \texttt{pyannote/segmentation-3.0} model, trained on the AMI Corpus, achieves a Diarization Error Rate (DER

1 Introduction

This paper examines: Domain Adaptation of the Pyannote Diarization Pipeline for Conversational Indonesian Audio. Research question: To what extent do ASR models trained with MWER generalize better across domains (e.g., conversational vs. technical speech) compared to cross-entropy trained models when measured by intent classification F1 scores on domain-specific datasets like TED-LIUM or Switchboard?.

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 6.9/10.

3 Results

13 papers retrieved. 34 claims extracted; 24 independently verified. Quality review score: 6.9/10.

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 study adopts Diarization Error Rate (DER) as the primary evaluation benchmark.	×	0.14
DER is defined as the percentage of input speech time not correctly attributed to the correct speaker, aggregating error	✓	0.26
The mathematical formula for DER is (TFA + TMISS + TCONF) / TTOTAL $\times$ 100%.	×	0.10
TFA (False Alarm) represents speech detected in silence.	×	0.12
TMISS (Missed Detection) represents speech not detected.	×	0.15
TCONF (Confusion) represents speech assigned to the wrong speaker.	✓	0.15
A lower DER indicates better diarization performance.	✓	0.17
The experimental pipeline was implemented using the pyannote.audio open-source toolkit version 3.1.	✓	0.21
The training process leverages pytorch-lightning to manage the optimization loop and GPU acceleration.	✓	0.24
All experiments were conducted on an NVIDIA T4 GPU.	✓	0.18
A custom experimental protocol named DebateIndonesianLarge was defined within the system's database registry configurati	✓	0.24
The protocol partitions the synthetic dataset into Train, Development, and Test sets to prevent data leakage.	×	0.14
The Train Set was used for supervised fine-tuning of model weights.	✓	0.23
The Development Set was used for validation and hyperparameter tuning.	✓	0.16
The Test Set was used solely for final inference and DER calculation.	✓	0.22
The Segmentation task provided by the framework was instantiated for the fine-tuning stage.	✓	0.17
The chunk duration hyperparameter was set to 2.0 seconds.	×	0.08
The batch size hyperparameter was set to 16 audio chunks per batch.	✓	0.15
Experiments were conducted with 1 and 2 training epochs.	×	0.15
The model was trained until the validation loss plateaued.	✓	0.15
The AMI Baseline model achieved a DER of 68.18%.	✓	0.15
The Indo Adapted (2h) model achieved a DER of 53.47%.	✓	0.16
The Indo Adapted (25h) model achieved a DER	×	0.15

References

- http://arxiv.org/abs/2111.03442v2
- http://arxiv.org/abs/1805.04699v4
- http://arxiv.org/abs/2601.03684v1