Multi-Representation Distribution Alignment in Domain-Adaptive Speech Recognition on XTREME-S

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the alignment of multi-representation distributions in domain adaptation affect the accuracy of speech recognition models on the XTREME-S benchmark when scaling across diverse linguistic. 10 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: TTA: Transcribe, Translate and Alignment for Cross-lingual Speech Representation. Research question: How does the alignment of multi-representation distributions in domain adaptation affect the accuracy of speech recognition models on the XTREME-S benchmark when scaling across diverse linguistic domains?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

14 papers retrieved. 10 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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 TTA model significantly outperforms Whisper models on widely used Chinese and English test sets, including Aishell,	×	0.04
The TTA model demonstrates a considerable WER on the CommonVoice dataset, even compared with Whisper Large-v3 (6.76% vs.	×	0.04
The TTA model shows clear advantages on MLS and VoxPopuli datasets.	×	0.04
For zero-shot evaluation on Fleurs, the TTA model fails to surpass Whisper Large models, while still behaves better than	×	0.06
The TTA model exhibits better performance than Whisper Medium on CoVoSTv2 for speech translation performance measured wi	×	0.08
The TTA model achieves 100% accuracy for all 10 training languages on the LID task evaluated on Fleurs.	×	0.05
Whisper Large-v3 performs generally the same as the TTA model but worse in Indonesian with 81% accuracy on the LID task	×	0.04
Models with Alignment components perform consistently better in the ablation study on the probing task of ST.	×	0.04
The TTA model is trained on 358k hours of speech data on multilingual speech recognition (MASR), speech translation (ST)	✓	0.28
The Zipformer encoder is a fast and memory-efficient variant of the Conformer architecture.	×	0.01

References

- http://arxiv.org/abs/2511.14410v2

- http://arxiv.org/abs/2203.10752v3
- http://arxiv.org/abs/2201.01002v1