Adapting Speech Foundation Models for L2 Speakers: Targeted Error Analysis and Adaptation under Proficiency Imbalance

Abstract

Speech foundation models like Whisper have set new benchmarks in various ASR tasks, but they often underperform for second-language (L2) learners due to accent variation, disfluencies, and mispronunciations — speech characteristics that are underrepresented during current model pretraining. In this study, we examine how to adapt large speech foundation models—specifically Whisper—to better serve L2 speakers.

Our framework begins with fine-grained error analysis across speaker proficiency levels, which identifies systematic failure modes such as hesitation insertions and high deletion, insertion & substitution in low-proficiency groups. This motivates adaptation strategies that explicitly account for proficiency-driven variation in L2 speech. Based on these insights, we implement: (1) parameter-efficient multitask learning via LoRA to jointly model transcription and speaker proficiency, and (2) targeted data augmentation simulating disfluency patterns to mitigate recognition bias toward fluent speech.

Preliminary results show that our proficiency-aware multitask model reduces WER across all proficiency levels, with the largest absolute improvement of 4.7% observed in the low proficiency group.

Building on our current framework, we plan to explore several extensions to further enhance adaptation for low-proficiency L2 speech. These include prompt-based decoding with speech-aware LLMs and N-best hypothesis reranking using both phoneme- and word-level representations. We will also investigate dynamic thresholding mechanisms to better handle hesitation phenomena during decoding. These directions aim to expand the adaptability and interpretability of our pipeline, and provide deeper insights into modeling underrepresented L2 speaker populations.

Keywords

Second Language Speech Recognition, Speech Disfluencies, Speech Data Augmentation, Speech-Language Integrated Modeling

Figure

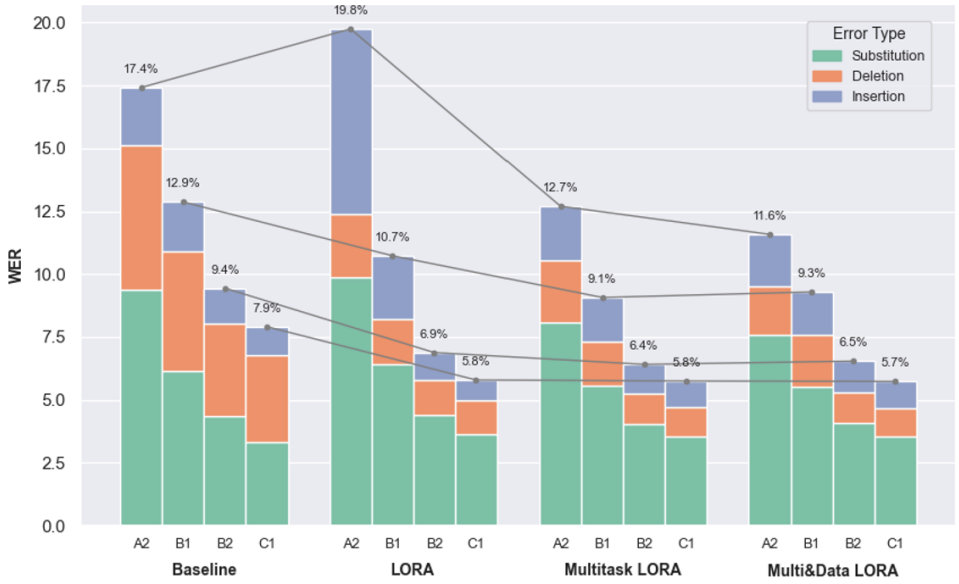

Figure 1: WER and Error Type Composition by Proficiency Level Across ASR Adaptation Models