Adaptation of Speech Foundation Model for Under Resource Task

Abstract:

Speech foundation models (SFMs) form the backbone of modern automatic speech recognition (ASR) systems, offering strong generalization across domains, speakers, and languages. However, their effectiveness diminishes in low-resource scenarios—particularly when adapting to new languages beyond the training distribution or recognizing infrequent, domain-specific vocabulary. These challenges are common in real-world deployments, where large-scale retraining is infeasible and annotated data is scarce.

This thesis addresses these limitations by proposing a unified framework for *low-resource adaptation* of SFMs. The focus lies on two key challenges: (1) **cross-language adaptation**, where extending models to new languages typically requires retraining on large multilingual datasets, leading to prohibitive computational costs; and (2) **rare word adaptation**, where the long-tail distribution of vocabulary prevents effective learning of low-frequency terms due to limited data availability.

To enable efficient cross-language adaptation, we propose three Transformer decoder–focused techniques: (i) **decoder training stabilization**, which combines gradient surgery, embedding freezing, and learning rate re-scaling; (ii) **embedding layer surgery**, which selectively updates embeddings to support new language tokens; and (iii) a **two-stage LoRA-based adaptation** framework enhanced with language-wise beam search. Applied to ten new languages in the Common Voice corpus, each with less than or equal to 10 hours of training data, these methods achieve up to 13.8% and 10.4% relative WER reductions on existing and new languages respectively, while maintaining low computational overhead.

For rare word adaptation, we introduce strategies that require little modification to the pretrained backbone. These include: (i) a **contextual biasing framework (TCPGen)** that attaches an external decoder to bias outputs toward rare words using synthetic audio; (ii) a **rare-word-aware loss function** combining masked cross-entropy with a word detection term to improve decoding precision; and (iii) a **training-free trie-based post-processing method** enhanced with K-step prediction and hypothesis pruning. Evaluated on DSTC2 and NSC Part 2, these methods reduce WER from 29.7% to 11.8% and from 26.5% to 7.0%, respectively, using only 10 hours of training data.

Together, these contributions provide scalable, data- and compute-efficient solutions for adapting SFMs in low-resource settings. By addressing both language and vocabulary gaps, this work advances the practical deployment of ASR systems across diverse, resource-constrained applications.