Adaptation of Speech Foundation Model for Under Resource Task

**Chapter 1-5 [109 page]**

**Chapter 1: Introduction [3 page]**

1.1 **Background and Motivation [1 page]**

* What is speech foundation model (SFM)? Why is it important?
* What is the limitation of SFMs? Discuss the limitations of SFMs (transcribing rare words, limited number of supported languages)
* How to solve it? -> low-resource adaptation
* What are the 2 low-resource adaptation challenges that this thesis focusses on?
* Why cross-lanugage adaptation is a low-resource adaptation problem?
* Why rare word adaptation is a low-resource adaptation problem?

1.3 **Thesis Objectives and Contributions [1.5 page]**

* What is the goal of the thesis: developing effective adaptation strategies for rare word adaptation and cross-language adaptation
* **Cross-language adaptation** (gradient surgery, embedding layer surgery, two-stage LoRA) [0.5 page]
* **Rare word adaptation** (TCPGen with rare-word-aware loss, Trie-based postprocessing with K-step prediction) [0.5 page]

1.4 **Thesis Organization [0.5 page]**

* What is the structure of the thesis, guiding the reader through the chapters and how each builds on the previous.

**2. Related Work [24 page]**

2.1 Speech Foundation Model [4 page]

* What is speech foundation models (SFMs): large-scale pretrained neural architectures designed to understand, process, and generate human speech.
* What are the Core Components: Acoustic models, language models, and end-to-end architectures (e.g., Transformer-based models like Whisper).
* What are the key Techniques, which is better in what scenario, why are they better: Deep learning, recurrent neural networks (RNNs), and especially Transformer architectures with self-attention mechanisms that enable strong generalization across languages and domains.
* Why is SFM important: Role of SFMs in enabling robust ASR, multilingual processing, speech synthesis, and enhancement.
* What are the recent Advances: Brief mention of large-scale pretraining on diverse datasets and challenges in adapting SFMs to new domains or languages efficiently.
* 图示

  AI 生成的内容可能不正确。 https://github.com/ga642381/speech-trident
  1. Low-resource Adaptation [10.5 page]
* Why do we need low-resource adaptation: Need for efficient adaptation of large SFMs to target domains, rare words, or new languages under low-resource conditions. [2.5 page]
* What Categories of Adaptation are there:
  + Full Fine-Tuning: Updating all model parameters, often costly and prone to overfitting or catastrophic forgetting.
  + Parameter-Efficient Fine-tuning: Techniques like adapters, LoRA (Low-Rank Adaptation), and modular gating that update small subsets of parameters or add lightweight modules. [4 page]
  + Continual Learning: Emphasise on preventing catastrophic forgetting without full retraining on all datasets [4 page]
* Trade-offs: Balancing adaptation effectiveness, computational cost, and retention of original model generalization.
  1. Multilingual ASR [2.5 page]
* What is MASR, What are the history and development of MASR
* What continual learning (CL) methods are there for multilingual ASR:focusing on techniques like LoRA and adapters.
* Which method is better
* Why is it better
* Examine the challenges of mitigating catastrophic forgetting when expanding ASR models to new languages and the role of language identification (LID) in improving multilingual models.

2.4 Rare Word Recognition in ASR [3 page]

* What is Rare Word Recognition, what is the history and development of rare word adaptation
* What are the existing methods for rare word recognition: including trie-based contextual biasing approaches.
* Which method is better
* Why is it better
* Discuss deep biasing techniques like TCPGen and postprocessing methods, comparing their strengths and weaknesses in terms of accuracy, efficiency, and training requirements.

2.6 Research Gap [4 page]

* Identify key gaps in the current literature, organized by the two main areas of focus in this thesis:

2.6.1 Rare Word adaptation [2 page]

* + Discuss the challenges of rare word recognition, particularly for domain-specific words in low-resource environments.
  + Examine the computational inefficiencies of traditional methods like trie-based biasing and the need for more efficient solutions to handle rare words effectively.

2.6.2 Cross language adaptation [2 page]

* + Discuss the limitations in expanding ASR models to support multiple languages while avoiding catastrophic forgetting.
  + Emphasize the need for parameter-efficient and low-resource adaptation techniques for unseen languages.

**3. Cross-language adaptation [43 page]**

Supporting paper:

Chin Yuen Kwok, Jia Qi Yip and Eng Siong Chng. "Continual Learning Optimizations for Auto-regressive Decoder of Multilingual ASR systems." INTERSPEECH. 2024.

Chin Yuen Kwok, Jia Qi Yip and Eng Siong Chng. "CONTINUAL LEARNING WITH EMBEDDING LAYER SURGERY AND TASK-WISE BEAM SEARCH USING WHISPER." 2024 IEEE Spoken Language Technology Workshop. 2024

Chin Yuen Kwok, He Xin Liu, Jia Qi Yip, Sheng Li and Eng Siong Chng. "A Two-Stage LoRA Strategy for Expanding Language Capabilities in Multilingual ASR Models" Submitted to TASLP

3.1 **Problem Definition and Importance [1 page]**

* Define the problem of unseen language adaptation and discuss its significance in the context of ASR.
* Explain the challenges of catastrophic forgetting and language ID errors when adapting to new languages under low-resource conditions.

3.3 **Gradient Surgery and Learning Stabilization [8 page]**

* Introduction [1 page]
* Related Work [2 page]
* Experience replay and A-GEM [1 page]
* Gradient surgery technique [1 page]
* embedding freezing [1.5 page]
* output suppression [0.5 page]
* learning rate re-scaling [0.5 page]
* Explain how these methods help avoid catastrophic forgetting during decoder adaptation. [0.5 page]

3.4 **Embedding Layer Surgery [7 page]**

* Introduction [1 page]
* Related Work [2 page]
* Introduce **Embedding Layer Surgery** as a solution to the problem of token embedding overwriting during continual learning. [2 page]
* Discuss **Task-Wise Beam Search** and how it enhances language identification accuracy, mitigating errors in the multilingual decoding process. [2 page]

3.5 **Two-Stage LoRA Approach with Task-Wise Beam Search [8 page]**

* Introduction [1 page]
* Related Work [2 page]
* Describe LoRA [1 page]
* Describe the two-stage LoRA approach for separating language identification (LID) and ASR adaptation using modular LoRAs. [2 page]
* Explain how is language-wise beam search implemented more efficiently by early stopping. [2 page]

3.6 **Experimental Results and Discussion [18 page]**

* Outline [0.5 page]
* Datasets [1.5 page]
* Experiment Setup [0.5 page]
* Evaluation Metric [1.5 page]
* Language-aware decoding results [4 pages]
* Language-agnostic decoding results [3 pages]
* Sequential adaptation to ten languages [2 pages]
* Language identification [1.5 pages]
* Task-wise beam search [1.5 pages]
* The Encoder’s Role for Language Adaptation [1 page]
* Computation and Memory Overhead of Two-stage LoRA [1 page]

3.7 **Summary [0.5 page]**

* Summarize the findings from the continual learning experiments, highlighting the effectiveness of the proposed methods in unseen language adaptation.

**4. Rare word adaptation [30 page]**

Supporting paper:

Chin Yuen Kwok, Jia Qi Yip, and Eng Siong Chng. "Improving Synthetic Data Training for Contextual Biasing Models with a Keyword-Aware Cost Function" Submitted to Interspeech25

Chin Yuen Kwok, Jia Qi Yip, and Eng Siong Chng. " Efficient Trie-based Biasing using K-step Prediction for Out-of-vocabulary Word Recognition" Submitted to Interspeech25

4.1 Problem Definition and Motivation [2.5 page]

* Define rare word recognition as a key low-resource ASR challenge.
* Explain why handling rare words efficiently is crucial for practical ASR systems, particularly in domains with evolving vocabulary.

4.2 TCPGen-Based Contextual Biasing [7.5 page]

* Introduction [0.5 page]
* Related Work [1 page]
* Introduce TCPGen [2.5 page]
* Explain the limitations of TCPGen [1 page]
* Keyword-Aware Loss [2.5 page]

4.3 Hypothesis Pruning using K-Step Prediction [8.5 page]

* Introduction [1 page]
* Related Work [1 page]
* Beam search [1.5 page]
* Contextual biasing problem formulation [1 page]
* Trie-based Postprocessing [1.5 page]
* Reward revocation [1.5 page]
* K-step prediction [1 page]

4.4 Experimental Results and Discussion [10.5 page]

* Experiment Setup (Evaluation Metric) [1.5 page]
  + Synthetic Data generation pipeline for Rare word dataset [2.5 page]
* TCPGen with keyword aware loss [3.5 page]
* Trie-based postprocessing with K-step prediction [3 page]

4.5 Summary [1 page]

* Summarize the insights gained from the experiments on rare word recognition, emphasizing the advantages of trie-based postprocessing over deep biasing.

**5. Conclusion and Future Work [3.5 page]**

5.1 **Summary of Contributions [2 page]**

* Recap the main contributions of the thesis: effective low-resource ASR adaptation strategies for both rare word recognition and unseen language expansion.

5.2 **Practical Implications [0.5 page]**

* Discuss the real-world applications of the proposed methods, including healthcare, multilingual services, and edge devices where low-resource adaptation is critical.

5.3 **Limitations and Open Challenges [0.5 page]**

* Identify ongoing challenges, such as cross-lingual rare word transfer, few-shot learning, and developing unified adaptation frameworks.

5.4 **Directions for Future Research [0.5 page]**

* Suggest future research avenues: improved LID gating, joint adaptation of biasing and language-specific models, and ethical considerations in using synthetic data for training ASR models.

Appendices

- Implementation details

- Additional experimental results

- Dataset descriptions

- Training configurations

- Model architectures

- Evaluation metrics

- Other publications

1. Chin Yuen Kwok, Jia Qi Yip, and Eng Siong Chng. “Robust Audio Deepfake Detection using Ensemble Confidence Calibration” 2025 International Conference on Acoustics, Speech, and Signal Processing. IEEE, 2025
2. Chin Yuen Kwok, Jia Qi Yip, and Eng Siong Chng. "Low Resource Language Adaptation using Two-stage Regularization for Multilingual ASR." 2025 International Journal on Asian Language Processing (IJALP). IEEE, 2025.
3. Chin Yuen Kwok, Jia Qi Yip and Eng Siong Chng. " Low Resource Language Adaptation using Two-stage Regularization for Multilingual ASR." The 28th International Conference on Asian Language Processing. 2024 (Best Paper Award)
4. Chin Yuen Kwok, Jia Qi Yip and Eng Siong Chng. "Improved Alignment for Score Combination of RNN-T and CTC Decoder for Online Decoding." 27th International Conference on Text, Speech and Dialogue. 2024.
5. Chin Yuen Kwok, Haoyang Li, and Eng Siong Chng. "ASR Model Adaptation for Rare Words Using Synthetic Data Generated by Multiple Text-To-Speech Systems." 2023 APSIPA ASC. IEEE, 2023.
6. Chin Yuen Kwok, Jia Qi Yip and Eng Siong Chng. "Bona fide Cross Testing Reveals Weak Spot in Audio Deepfake Detection Systems." Submitted to Interspeech25