Whisper Finetuning For Hakka Recognition in Low Resource

Min Han Teng[1] **Ci Dao Chen**[1] **You Ting Lin**[1] **Bing Jhih Huang**[1] **Jia-Ching Wang**[1]
[1]Department of Computer Science, National Central University, Taiwan
olivier40103@gmail.com, copeman123@gmail.com, eeyore0624@gmail.com, c725992@gmail.com

Abstract

We study automatic speech recognition (ASR) for Hakka, a low-resource language with substantial dialectal variation. Focusing on Zhaoan and Dapu, we fine-tune Whisper using Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning (AdaLoRA) and apply data augmentation to mitigate data scarcity. Experiments show that AdaLoRA combined with augmentation substantially improves cross-dialect recognition while maintaining parameter efficiency. Our results demonstrate the potential of lightweight adaptation to extend large-scale ASR systems to underrepresented languages, supporting the preservation of Hakka speech and orthography.

1 Introduction

Large-scale pretrained automatic speech recognition (ASR) models, such as Whisper, have achieved strong performance on high-resource languages. However, their applicability to low-resource and dialectally diverse languages remains underexplored. Hakka, a Sinitic language with millions of speakers worldwide, is particularly underrepresented in ASR research despite its cultural and linguistic significance. The lack of standardized resources, limited digital presence, and substantial phonological variation across dialects pose major challenges for building robust ASR systems and heighten the risk of language endangerment.

Among the many Hakka dialects, Zhaoan and Dapu represent two major varieties whose systematic phonological differences further complicate recognition. Developing ASR systems that generalize across these dialects is especially difficult under low-resource conditions.

To address these challenges, we fine-tune Whisper for Hakka using Low-Rank Adaptation (AdaLoRA), a parameter-efficient method well-suited for low-resource adaptation of large-scale pretrained models. To further mitigate data scarcity and improve cross-dialect robustness, we incorporate data augmentation techniques that expand the training set and enhance generalization.

This work not only provides a practical solution for advancing Hakka ASR, but also offers a replicable framework for extending large-scale ASR models to other underrepresented languages, contributing to both language preservation and applied speech technologies.

Our main contributions are as follows:

- We propose training strategies and data augmentation techniques that improve model performance and robustness in low-resource, multi-dialect settings.
- We empirically demonstrate consistent gains in cross-dialect recognition, measured by pinyin WER and character CER, validating the effectiveness of lightweight adaptation for underrepresented languages.

2 Related Works

Automatic Speech Recognition (ASR) Large-scale pretrained ASR models have greatly advanced multilingual speech recognition. Whisper (Radford et al., 2022),achieves state-of-the-art results in high-resource languages but struggles in low-resource settings due to limited data and dialectal variation. Prior work has explored transfer learning (Wang et al., 2021), and parameter-efficient fine-tuning (Zhang et al., 2023; Liu et al., 2022). Low-Rank Adaptation (AdaLoRA) (Zhang et al., 2023) has proven particularly effective for adapting large ASR models, making it suitable for extending Whisper large-v2 to Hakka.

Text-to-Speech (TTS) TTS has progressed with multilingual and multispeaker pretrained models. Neural approaches such as Tacotron 2 (Shen et al., 2018) and FastSpeech (Ren et al., 2019) achieved natural synthesis in high-resource languages, while

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 450–453
November 20-22, 2025 ©2025 Association for Computational Linguistics

YourTTS (Casanova et al., 2022), based on VITS (Kim et al., 2021), introduced zero-shot multilingual, multispeaker capabilities. Trained on VCTK (Yamagishi et al., 2019), YourTTS enables speaker adaptation with under one minute of speech. Beyond synthesis, TTS has been used to generate synthetic data for ASR in low-resource settings (Rosenberg et al., 2019; Li et al., 2020). Here, we adopt YourTTS to augment scarce Hakka corpora for more robust ASR training.

Data Augmentation for Speech Data augmentation is widely used to improve ASR robustness in low-resource contexts. Common methods include noise injection, Utterance concatenation, Time stretching, Pitch shifting, Air absorption, and Environmental impulse response (Ko et al., 2017; Zahid and Qazi, 2025; van der Meer, 2022; Kates and Brandewie, 2020; Bryan, 2019). More recent work explores cross-lingual transfer and TTS-based generation (Hsu et al., 2020; Rosenberg et al., 2019). SpecAugment (Park et al., 2019) is now standard in ASR, while TTS-based augmentation (Li et al., 2020; Jia et al., 2019) creates labeled data for low-resource languages.

3 Method

3.1 ASR Model Fine-tuning

We adopt the Whisper large-v2 model (Radford et al., 2022) as the backbone for Automatic Speech Recognition (ASR). To efficiently adapt the large-scale model to underrepresented Hakka dialects, we employ adaptive Low-Rank Adaptation (AdaLoRA) (Zhang et al., 2023), which enables parameter-efficient fine-tuning without retraining the full model. This setup allows the model to retain general multilingual knowledge while specializing in Hakka recognition.

3.2 Synthetic Data Generation with TTS

To address the scarcity of annotated Hakka speech, we leverage YourTTS (Casanova et al., 2022), a multilingual and multispeaker zero-shot TTS system based on the VITS architecture (Kim et al., 2021). We use YourTTS to synthesize Hakka utterances across different accents, thereby enriching the training corpus and improving model generalization. In addition, we specifically generate synthetic data for coarticulated syllables using self-collected and organizer-provided texts, ensuring better phonological coverage. Our training corpus combines both human-recorded and synthetic speech. The human data consist of Hakka Dapu and Zhaoan recordings provided by the competition organizers. To complement this limited corpus, we generated synthetic speech using YourTTS across multiple accents. In addition, we produced accent-specific utterances focusing on coarticulated syllables to improve phonological coverage and enhance recognition robustness.

3.3 Data Preprocessing and Augmentation

To further improve robustness under low-resource and cross-dialect conditions, we applied a set of augmentation techniques during training, including noise injection, utterance concatenation, time stretching, pitch shifting, air absorption, and environmental impulse response. These methods enrich the acoustic diversity of the training data, enabling the model to better recognize Hakka speech across varied speaking styles and noisy environments.

4 Experiments

4.1 Experimental Setup

Dataset

The dataset comprises both human-recorded and synthetic Hakka speech. The human portion includes approximately 70 hours of Dapu and Zhaoan recordings provided by the competition organizers. To expand this limited resource, we used YourTTS to generate around 310,000 utterances for each dialect, from which 25,000 per dialect were sampled to form a balanced synthetic training set. In addition, about 28,000 utterances targeting coarticulated syllables were synthesized to enhance phonological coverage, yielding roughly 100,000 synthetic training samples in total.

Metrics

We evaluated performance using two complementary metrics CER for character track and WER for pinyin track. These metrics jointly capture both graphemic and phonemic aspects of recognition, providing a comprehensive assessment of Hakka ASR performance.

Finetune Details

We fine-tuned the Whisper large-v2 model (Radford et al., 2022) for Hakka ASR using AdaLoRA. The adaptation targeted key modules including `k_proj`, `q_proj`, `v_proj`, `out_proj`, `fc1`, and `fc2`. Training was conducted for 25 epochs with the AdamW optimizer, a learning rate of 1×10^{-4}, and mixed-precision optimization (fp16).

To enhance robustness, several data augmentation strategies were applied with explicit probabilities and parameter ranges. Time stretching was applied with a probability of 0.25, adjusting the speaking rate within a range of 0.9 to 1.1 (Ko et al., 2015; McFee et al., 2015). Pitch shifting was also used with a probability of 0.25, modifying the fundamental frequency by up to ± 4 semitones (Ko et al., 2015; McFee et al., 2015). To simulate far-field effects, air absorption was introduced with a probability of 0.5, modeling distances between 10 and 50 meters (Habets, 2006). Environmental reverberation was added using impulse responses from RIR at a probability of 0.25 (Habets, 2006). Short environmental noises from the ESC-50 (Piczak, 2015) dataset were injected with a probability of 0.75, using signal-to-noise ratios (SNR) between 3 and 30 dB and durations of 2–8 seconds. Finally, Gaussian noise was added with a probability of 0.25, with SNR ranging from 5 to 40 dB (Ko et al., 2015).

Baselines

For the baseline system, we fine-tuned the Whisper large-v2 model directly on the original Hakka corpus provided by the organizers, which consists of Zhaoan and Dapu recordings. No synthetic speech data or augmentation techniques were applied. The model was trained under the same setup as our proposed method, using AdaLoRA for parameter-efficient fine-tuning with 25 training epochs, the AdamW optimizer, a learning rate of 1×10^{-4}, and mixed-precision optimization (fp16). This baseline serves as a reference point to evaluate the effectiveness of our data augmentation and synthetic data generation strategies.

4.2 Main Results

Table 1 summarizes the effectiveness of our proposed strategies. In low-resource and cross-dialect settings, ASR models often lack sufficient acoustic variability to generalize to real-world scenarios. To address this limitation, we deliberately designed and introduced a series of data augmentation techniques, including noise injection, utterance concatenation, time stretching, pitch shifting, air absorption, and environmental reverberation. These augmentations explicitly enriched the acoustic variability of the training corpus, thereby equipping the model with the ability to cope with noisy conditions, diverse speaking rates, phonological variation, and far-field speech scenarios. In addition, we incorporated TTS-generated synthetic data to compensate for limited phonological coverage and to expand the overall training corpus. As a result of combining enhanced acoustic diversity with more comprehensive phonological coverage, the model achieved substantial performance gains in cross-dialect recognition. Compared with the baseline trained solely on the original corpus (CER = 31.9%, WER = 51.5%), the final system reduced the character CER to **5.6%** and the pinyin WER to **16.7%**. These reductions—over fivefold in CER and more than threefold in WER—clearly demonstrate that synthetic data generation, together with carefully designed augmentation, provides the critical capabilities necessary for improving cross-dialect ASR robustness under low-resource conditions.

System	CER (%)	WER (%)
Baseline	31.9	51.5
Proposed Method	**5.6**	**16.7**

Table 1: Recognition performance on Hakka ASR under different training setups. The incorporation of TTS-generated synthetic data and augmentation methods yields substantial improvements over the baseline system

4.3 Error Analysis and Re-sampling

We performed detailed error analysis to identify persistent recognition errors. To address these, additional utterances were re-sampled from the 310,000 TTS-generated dataset for targeted fine-tuning. However, this error-driven re-sampling did not yield further performance gains, suggesting that improvements depend more critically on data quality and phonological coverage than on the sheer quantity of synthetic speech.

5 Conclusion

This work tackled the challenge of Hakka ASR under low-resource and dialectal variation by fine-tuning Whisper with LoRA and augmenting data with synthetic speech. Our approach reduced CER from 31.9% to 5.6% and WER from 51.5% to 16.7%, demonstrating that parameter-efficient adaptation with augmentation can yield substantial cross-dialect gains. Beyond Hakka, the framework offers a replicable path for extending large-scale ASR to other low-resource languages. Future work will target broader dialect coverage, more natural synthetic speech, and cross-lingual transfer, further supporting the preservation of endangered languages in the era of large-scale AI.

References

Nicholas J. Bryan. 2019. Impulse response data augmentation and simulation as alternatives to rir collections for speech recognition. In *Proceedings of WASPAA*, pages 229–233.

Edresson Casanova, Juliano Weber, Christopher Shulby, Antonio Junior, Rodolfo da Silva, Moacir Antonelli Ponti, Sandra Aluisio, Junichi Yamagishi, Yossi Adi, Mohamed Haidar, and 1 others. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning (ICML)*.

Emanuël AP Habets. 2006. Room impulse response generator. Technical report, Technische Universiteit Eindhoven. Technical Report.

Wei-Ning Hsu, Benjamin Bolte, Yung-Sung Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2020. Meta learning for low-resource speech recognition. In *ICASSP*.

Ye Jia, Heiga Zen, Ron J Weiss, and 1 others. 2019. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *ICASSP*.

James M. Kates and Eugene J. Brandewie. 2020. Adding air absorption to simulated room acoustic models. *The Journal of the Acoustical Society of America*, 148(5):EL408–EL413.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *ICML*.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Interspeech*.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2017. A study on data augmentation of reverberant speech for robust speech recognition. In *Proceedings of ICASSP*, pages 5220–5224.

Bo Li, Abdelrahman Mohamed, and Geoffrey Zweig. 2020. Training data augmentation for end-to-end speech recognition using text-to-speech synthesis. In *ICASSP*.

X Liu, Jonas Pfeiffer, Sebastian Ruder, and 1 others. 2022. Adapterfusion: Non-destructive task composition for transfer learning. In *EACL*.

Brian McFee, Colin Raffel, Daniel Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, pages 18–25.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech*.

Karol J Piczak. 2015. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018. ACM.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. In *NeurIPS*.

Andrew Rosenberg, Bhuvana Ramabhadran, Abhinav Sethy, and 1 others. 2019. Speech synthesis for data augmentation in noisy speech recognition. In *ICASSP*.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, and 1 others. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *ICASSP*.

Jelle van der Meer. 2022. Evaluating the use of pitch shifting to improve automatic speech recognition. Master's thesis, Delft University of Technology.

Changhan Wang, Wei-Ning Hsu, and 1 others. 2021. Improving low-resource speech recognition with cross-lingual self-supervised learning. In *ICASSP*.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*.

Muhammad Zahid and Imran Qazi. 2025. Pitch-speed feature space data augmentation for automatic speech recognition improvement in low-resource scenario. *International Journal of Speech Technology*.

Qianxi Zhang, Zhenheng Yang, Tianlong Chen, and Zhangyang Wang. 2023. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations (ICLR)*.