The 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)
Taipei City, Taiwan, October 20-21, 2023. The Association for Computational Linguistics and Chinese Language Processing

Enhancing Automatic Speech Recognition Performance Through Multi-Speaker Text-to-Speech

Po-Kai Chen
Department of CSIE,
National Central University
pokaichen@g.ncu.edu.tw

Bing-Jhih Huang
Department of CSIE,
National Central University
c72599@gmail.com

Chi-Tao Chen
Department of CSIE,
National Central University
apoman123@gmail.com

Hsin-Min Wang
Institute of Information Science,
Academia Sinica
whm@iis.sinica.edu.tw

Jia-Ching Wang
Department of CSIE,
National Central University
jcw@csie.ncu.edu.tw

Abstract

In this study, we present a novel approach to enhancing the performance of our Hakka Automatic Speech Recognition (ASR) model through the strategic use of Text-to-Speech (TTS) amplification techniques. Our investigation explores the integration of diverse speakers to expand our training dataset, leading to a notable reduction of Character Error Rate (CER) approximately 0.2 in the validation set and approximately 3.96 on the test set. These compelling results affirm the effectiveness of multi-speaker TTS strategies in generating ASR data, ultimately bolstering the resilience and precision of our ASR system.

Keywords: Automatic Speech Recognition, Multi-Speaker Text-to-Speech, Data extension

1 Introduction

In the ever-evolving landscape of Automatic Speech Recognition (ASR)(Mishaim Malik, 2020), the quest to unlock the potential of underrepresented languages and dialects remains a paramount challenge. Among these linguistic treasures, Hakka, a prominent language in Taiwan, has long awaited its moment in the spotlight. This paper embarks on a transformative journey, harnessing state-of-the-art ASR models—Whisper and WavLM—as our guiding lights, and employing innovative strategies to bridge the gap between technology and linguistic diversity. The foundation of our research rests on a meticulous curation of training and testing data provided by the competition organizers, augmented by additional data sourced through diligent efforts. To further enhance our dataset, we turn to Text-to-Speech (TTS) systems—a powerful tool that aligns seamlessly with our vision of data augmentation for Hakka ASR. Through these efforts, we amplify our dataset's richness and diversity, fostering the robustness of our ASR system in the face of linguistic variations. This paper unfolds as a testament to the resilience of underrepresented languages and the transformative power of cutting-edge ASR technologies. By integrating Whisper and WavLM, augmenting our data, and embracing TTS as an augmentation tool, we aim to amplify the voices of Hakka speakers,and fostering inclusivity. In the following sections, we delve into the intricacies of our methodology, present empirical findings, and engage in a nuanced discussion, all in the pursuit of advancing Hakka ASR technology and transcending the linguistic frontiers that beckon us forward.

2 Related Works

2.1 Whisper

In the realm of ASR, the Whisper model (Radford et al., 2023) stands as a beacon of innovation and performance, symbolizing a crucial advancement in the pursuit of effective and versatile ASR solutions. As we delve into the intricacies of spoken language understanding, Whisper emerges as a transformative force, exemplifying excellence in multilingual ASR systems. A defining feature of Whisper is its adaptability, rendering it proficient in both major languages and underrepre-

sented dialects. This adaptability is nurtured through multilingual pre-training, followed by fine-tuning on domain-specific datasets. Whisper's resilience in accommodating dialectal nuances positions it as an indispensable tool for ASR researchers, particularly in the context of preserving and promoting linguistic diversity, such as with Hakka. The significance of Whisper reverberates in its integration within the broader ASR landscape. It represents a crucial stepping stone in the democratization of advanced ASR technology, thereby fostering inclusivity and innovation in voice-related applications. Whisper's ease of access and compatibility empower researchers to explore ASR frontiers across diverse linguistic contexts. In the subsequent sections, we delineate the strategic role of Whisper within our ASR framework, elaborating on its contributions to advancing Hakka ASR technology.

2.2 WavLM

In the pursuit of advancing ASR systems, WavLM (Chen et al., 2022)(Team, 2019) emerges as a transformative paradigm shift, championing waveform-based learning and the unification of acoustic signals with linguistic representation. WavLM distinguishes itself through its fundamental departure from conventional ASR methodologies. Rather than relying solely on phonetic or textual transcriptions of speech, WavLM pioneers a waveform-to-waveform approach. This innovative strategy enables it to directly model acoustic waveforms, bridging the gap between raw audio signals and linguistic context—a testament to its ingenuity in tackling underrepresented languages and low-resource ASR scenarios. The core architecture of WavLM is rooted in deep neural networks, wherein it leverages the prowess of neural autoregressive modeling techniques. By directly modeling waveforms, WavLM exhibits an inherent adaptability to diverse speaking styles, regional accents, and varying acoustic conditions WavLM's transformative impact reverberates in its exceptional performance across low-resource and multilingual ASR benchmarks. In these scenarios, it has consistently surpassed traditional ASR systems, underscoring its potential to address the inherent challenges associated with languages such as Hakka, which may face constraints in available training data. In the ensuing sections, we elucidate the strategic integration of WavLM within our ASR framework, detailing the pivotal role it plays in advancing Hakka ASR technology.

2.3 VITS

In the landscape of text-to-speech synthesis, the Variational Inference Text-to-Speech (VITS) model (Kim et al., 2021) emerges as a pioneering advancement with profound implications. As we navigate the realm of natural language understanding and human-computer interaction, VITS has risen to prominence due to its remarkable capability to fuse variational inference with neural network architectures, thereby elevating the quality and expressiveness of synthesized speech. The significance of VITS extends beyond mere proficiency in generating lifelike speech. It empowers users with fine-grained control over latent variables, affording customization of speaking rates, pitch, accents, and emotional inflections—a versatility that aligns well with the exigencies of applications like personalized voice assistants, audiobook narration, and multimodal interfaces. Moreover, VITS thrives in a multilingual and cross-linguistic milieu, learning from diverse corpora and transcending language barriers. This adaptability to linguistic variations underscores its applicability in diverse linguistic contexts In the backdrop of our own research, we harness VITS as a pivotal tool for data augmentation, intricately woven into the fabric of our Hakka ASR endeavors. Through VITS-generated synthetic speech data, we augment our ASR training dataset, bolstering the robustness of our Hakka ASR system. In the ensuing sections, we elucidate the seamless integration of VITS within our ASR pipeline, expounding on the salient advantages it confers upon our research within the context of Hakka ASR.

3 Methodology

In this section, we will describe the whole process including translation, data generation.

3.1 Pinyin

Considering the absence of Pinyin tags in the additional information, particularly in the extra data lacking 客家羅馬數字調 Pinyin an-

notations, we leveraged the **gohakka** platform. This resource facilitated the translation of Hakka Chinese text into 客家羅馬拼音, which was then further processed using a web crawler. The web crawler systematically retrieved and converted the translation results into 客家羅馬數字調拼音. The conversion adhered to the established rules outlined in the "客語拼音方案" published by the Ministry of Education. This meticulous filtering and translation process yielded a substantial dataset comprising 3,619 entries.

3.2 Data Generation

To augment data, we pretrained a VITS which using Hakka character as input. Leveraging the train and eval from FSR-2023 dataset as training data, with all 87 speaker and about 70 hours of audio data in total. Preprocess are stated as follows, For the audio part, first trim off the silence segment with Silero VAD (Team, 2021) then resample to match the VITS sample rate. For the character part, first remove unwilling character which doesn't contribute to audio such as quotation marks and character the speaker doesn't speak then save the string in utf-8 format. The VITS code and setting we use are the same as the official implementation except relax the max input length from 190 to 2000. After training, we use the written text contained within the Extra Data portion 4.1.2 as inference script which contains about 19.2k sentence. For each sentence we uniformly sample 5 out of the 87 speakers to synthesis the audio and build an augmented dataset with about 100k speech in it.

4 Experiment

4.1 Experimental Setup

4.1.1 Data Augmentation

In the realm of data augmentation, we have harnessed the capabilities of the audiomentations library, a third-party Python resource tailor-made for augmenting audio data. Our augmentation strategy encompasses a trio of techniques:

A. **TimeStretch**: To manipulate the temporal dimension, we applied TimeStretch(Waibel, 2020) with parameters specifying a minimum rate of 0.9, a maximum rate of 1.1, and a probability of 0.25. This adjustment introduces controlled time variations into our audio dataset.

B. **PitchShift**: In order to introduce pitch variability, we skillfully employed PitchShift(Kakade, 2018). This technique was configured with a range of semitone shifts, with a minimum of -4 and a maximum of 4, coupled with a probability of 0.25. This augmentation method injects diverse pitch characteristics into our audio samples.

C. **AirAbsorption**: For the simulation of environmental conditions, we judiciously utilized AirAbsorption. With parameters set to a minimum distance of 10.0 units, a maximum distance of 50.0 units, and a probability of 0.5, we recreated the effect of sound absorption in different spatial settings. This enriches our dataset with variations in environmental acoustics.

Furthermore, it's worth noting that we incorporated the SpecAug(Daniel S. Park, 2019) method into the Whisper configuration, initiating it with a probability of 0.1.

4.1.2 Datasets

In this research endeavor, our dataset comprises four distinct components:

1. **HAT-Vol1**: Furnished by the competition organizers, this dataset encompasses a substantial pool of audio resources, consisting of 20,162 training samples and 3,598 verification samples. It serves as a foundational source for our study.
2. **Extra**: In addition to the core dataset, we have benefited from supplementary data generously provided by the Hakka Committee. This supplementary dataset primarily comprises spoken language text, constituting speech-text pairs, as well as written text in plain form.
3. **Self-Sourced**: As part of our data collection efforts, we conducted targeted searches on YouTube to gather Hakka language content. This self-sourced dataset augments the diversity of our corpus, bringing in unique perspectives and spoken language samples.

4. **Generated**: Through the utilization of a TTS model, we have created a dedicated subset of data. This subset is derived from the written text contained within the Extra Data portion. By employing the TTS model, we transformed written content into synthesized audio, expanding the scope of our dataset.

Together, these four datasets form the foundational building blocks of our research, providing a comprehensive and multifaceted corpus for our investigation into the Hakka language. The detailed statistical of all training corpus are shown in Table 1.

Table 1: Statistics of the all traing corpus. These statistics provide an overview of the size and composition of each training corpus after rigorous data cleaning.

Set	**Nbr. of sample**
HAT-Vol1	20,612
Extra	3,619
Self-Sourced	14,031
Generated (1spk/sent)	19,215
Generated (3spk/sent)	57,645

4.1.3 Model Configuration

For model configuration, we prioritize reproducibility and experimental evaluation by utilizing open-source implementations. Specifically, for the Character Track, we use the `openai/whisper-large-v2` pre-trained weights available through Huggingface's platform (Wolf et al., 2020) for the Whisper pre-trained model. Moreover, to enhance the efficiency of our model training, we have implemented LoRA(Edward J. Hu, 2021) adapter training. We initialize LoRA adapters with an init_r as 12, target_r as 4, lora_alpha as 32, lora_dropout as 0.1, and target_modules: {kproj, qproj, vproj, outproj, fc1, fc2}. For the Pinyin Track, we use the baseline model which is provided by the competition organizers.

4.1.4 Training details

In terms of training details, For the Character Track, we employ the AdamW optimizer with a learning rate of 1e-3. The warmup step is set to 500, and we set batch sizes to 64. Additionally, we establish a maximum training epoch limit of 12 to ensure effective model convergence and performance. For the Pinyin Track, our training approach aligns closely with the baseline settings as stipulated by the competition organizers.

4.1.5 Inference details

To accelerate inference, we utilize Faster-Whisper to quantize the model to float16 and employ a beam size of 4 during inference. To mitigate structural limitations in Whisper, we leverage Faster-Whisper's Voice Activity Detection (VAD) with a threshold of 0.5 and a minimum silence time of 1250 milliseconds. These optimizations enhance inference speed and efficiency, especially for real-time applications and long sound files.

Table 2: CER(%) of the results for Character Hakka speech recognition on the FSR-2023 validation set.

Architecture	Validation
Baseline(BSL-3)	6.67
Whisper-LV2+LoRA**(Ours)**	6.79
+Extra Data	5.98
+Self-Sourced Data	5.08
+Generated Data (1spk/sent)	3.51
+Generated Data (3spk/sent)	**3.04**

4.2 Evaluation of Character Track

To assess the influence of data augmentation, we systematically expanded our training dataset while monitoring the model's performance. As presented in Table 4, a discernible trend emerges: an increase in the volume of training data correlates with a consistent decrease in Character Error Rate (CER). Particularly noteworthy is the substantial performance boost observed upon introducing the dataset augmented by TTS during training.

Furthermore, by enriching the dataset with content from multiple speakers (three speakers per sentence), the CER is further reduced by about 0.2. This outcome serves as compelling evidence of the effectiveness of our data augmentation strategy.

4.3 Evaluation of Pinyin Track

In line with the outcomes of section 4.2, we streamlined the overall experimental process

Table 3: SER(%) of the results for Pinyin Hakka speech recognition on the FSR-2023 validation set.

Architecture	Validation
Baseline	7.95
+Extra Data	
+Self-Sourced Data	
+Generated Data (1spk/sent)	**4.54**

by implementing a baseline approach for data augmentation. As presented in Table 3, This pragmatic adjustment yielded an impressive 75% enhancement in Speech Emotion Recognition (SER) performance.

Table 4: CER(%) of the results for Character Hakka speech recognition on the FSR-2023 final test set.

Architecture	Validation
Whisper-LV2+LoRA	20.95
+Extra Data	22.32
+Self-Sourced Data	22.69
+Generated Data (3spk/sent)	**16.99**

4.4 Final Test of Character Track

In the final test of the FSR-2023 Character Track, illustrated in Figure 1, our team's results secured the top position, outperforming all other student's participating teams. Notably, our CER exhibited a remarkable reduction of 26.2% compared to the second-place finisher.

Further analysis, as detailed in Table 4, revealed an interesting observation: the substantial improvement in performance was primarily attributed to the incorporation of data generated by TTS during training. However, when we introduced the Extra Data and Self-Sourced Data into our training set, we observed a marginal decrease in performance on the test set. This phenomenon suggests a potential disparity in data distribution between the Extra Data and Self-Sourced Data compared to the official dataset. It's noteworthy that the TTS model, a key component of our data augmentation strategy, was trained exclusively on the HAT-Vol1 dataset, which may have contributed to these findings.

Our comprehensive analysis underscores the significance of data source compatibility in training ASR models and highlights the effectiveness of TTS-generated data in enhancing ASR system performance.

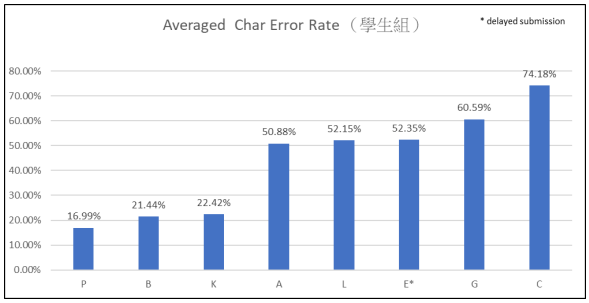

Figure 1: FSR 2023 Final Test Results for all student group.

5 Conclusion

In conclusion, our implementation of the TTS extension strategy has yielded a significant enhancement in the performance of our ASR model. Moreover, the augmentation of diverse speakers to expand our training data has led to notable reductions in the CER of approximately 0.2 on the validation set and approximately 3.96 on the test set. These results strongly emphasize the effectiveness of employing multi-speaker TTS techniques to generate ASR data, ultimately bolstering the resilience and precision of our ASR system.

References

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

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

Phillip Wallis Zeyuan Allen-Zhu Yuanzhi Li Shean Wang Lu Wang Weizhu Chen Edward J. Hu, Yelong Shen. 2021. Lora: Low-rank adaptation of large language models. In *arXiv:2106.09685*.

John Thickstun; Zaid Harchaoui; Dean P. Foster; Sham M. Kakade. 2018. Invariances and data augmentation for supervised music transcription. In *International Conference on Acoustic, Speech and Signal Processing*.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.

Khawar Mehmood Imran Makhdoom Mishaim Malik, Muhammad Kamran Malik. 2020. Automatic speech recognition: a survey.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

ESPNET Team. 2019. Espnet: End-to-end speech processing toolkit. https://github.com/espnet/espnet.

Silero Team. 2021. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad.

Thai-Son Nguyen; Sebastian Stüker; Jan Niehues; Alex Waibel. 2020. Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation. In *International Conference on Acoustics, Speech and Signal Processing*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.