# The SLAM Hakka ASR System for Formosa Speech Recognition Challenge 2025

**CHIH-HSI CHEN**
Department of Computer Science and Information Engineering, National Cheng Kung University
chenbrian93@gmail.com

**PEI-JUN LIAO***
Institute of Information Science, Academia Sinica
newsboy3423@iis.sinica.edu.tw

**CHIA-HUA WU**
Institute of Information Science, Academia Sinica
maxwu@iis.sinica.edu.tw

**PANG-CHENG WU**
National Yang Ming Chiao Tung University
andy610125@gmail.com

**HSIN-MIN WANG***
Institute of Information Science, Academia Sinica
whm@iis.sinica.edu.tw

***Corresponding author**

## Abstract

In recent years, large-scale pre-trained speech models such as Whisper have been widely applied to speech recognition. While they achieve strong performance on high-resource languages such as English and Mandarin, dialects and other low-resource languages remain challenging due to limited data availability. The government-led “Formosa Speech in the Wild (FSW) project” is an important cultural preservation initiative for Hakka, a regional dialect, where the development of Hakka ASR systems represents a key technological milestone. Beyond model architecture, data processing and training strategies are also critical. In this paper, we explore data augmentation techniques for Hakka speech, including TTS and MUSAN-based approaches, and analyze different data combinations by fine-tuning the pre-trained Whisper model. We participated in the 2025 Hakka FSR ASR competition (student track) for the Dapu and Zhaoan varieties. In the pilot test, our system achieved 7th place in character recognition (CER: 15.92) and 3rd place in pinyin recognition (SER: 20.49). In the official finals, our system ranked 6 in Hanzi recognition (CER: 15.73) and 4 in Pinyin recognition (SER: 20.68). We believe that such data augmentation strategies can advance research on Hakka ASR and support the long-term preservation of Hakka culture.

Keywords: Hakka, ASR, Low Resource

## 1 Introduction

In recent years, Taiwan has actively invested in the preservation and development of national languages, and has promoted mother-tongue education in primary and secondary schools. In addition to Taiwanese (Southern Min), Indigenous languages, and the languages of Southeast Asian new immigrants, Hakka has also been a major focus. To encourage its daily use, teaching, and revitalization, the “Formosa Speech in the Wild (FSW) project” has launched dialect preservation initiatives, including the organization of the FSR community competition for Hakka automatic speech recognition (ASR). This shared task provides timely benchmarks and resources, with the second edition held in 2025. Hakka belongs to the Sinitic branch and encompasses multiple regional phonological systems. In particular, the Dapu and Zhaoan varieties used in the 2025 competition differ not only in segmental systems but also in prosody, such as tone and duration. Over the long term, the lack of a widely adopted writing system, combined with the declining use of Hakka among

*Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025)*, pages 504–511
November 20-22, 2025 ©2025 Association for Computational Linguistics

younger generations, has restricted the availability of annotated corpora. From a cultural perspective, however, Hakka is central to the transmission of Hakka heritage; from a technological perspective, ASR can support pronunciation assessment and computer-assisted language learning.

We approach Hakka ASR as a data-centric transfer learning challenge, emphasizing the strategic fine-tuning of the powerful general-purpose foundation model Whisper-large-v3 (Radford et al., 2023) to enhance performance on Hakka corpora. We chose Whisper as our backbone model due to its verified multilingual capability, stability in transfer learning, and feasibility on commonly available GPU hardware.

To address the limited training data for the Dapu and Zhaoan dialects, we adopted several strategies:
(i) extending the training set with synthetic speech generated by a Text-to-Speech (TTS) system;
(ii) collecting audio-text pairs from publicly available Hakka learning platforms, following the procedure described by Chen et al. (2023), to construct additional training data for the Dapu and Zhaoan dialects (restricted to Hanzi transcriptions);
(iii) incorporating speech and text from Hakka radio broadcasts in the same dialects. For data augmentation, we first applied MUSAN (Snyder et al., 2015) to inject random noise, and further employed Audiomentations (Ronny, 2020) to introduce dynamic perturbations within each training batch, thereby improving model robustness. In the 2025 FSR Hakka ASR competition (student track), our system achieved 7th place in Hanzi recognition (Character Error Rate, CER: 15.92) and 3rd place in Pinyin recognition (Syllable Error Rate, SER: 20.49) during the pilot test. In the official finals, our system ranked 6th in Hanzi recognition (CER: 15.73) and 4th in Pinyin recognition (Word Error Rate, WER: 20.68).

The following sections describe in detail our strategy for leveraging Whisper, the methods used for data augmentation and corpus expansion, and the full set of experimental results, highlighting the effectiveness and limitations of each approach. Finally, we discuss the broader implications of our findings for speech technology, especially in the context of preserving and revitalizing cultural languages. Through this study, we aim to provide methodological insights and practical tools for the future development of Hakka ASR and other low-resource language technologies.

## 2 Model Architecture

We use the fine-tuned Whisper model as our final submission to the competition. In addition, we perform fine-tuning on LLaMA-Omni for comparison. The details and descriptions of both models are presented below.

### 2.1 Whisper

In this competition, we adopt Whisper as the backbone model, following the approach of Lu et al. (2023). Whisper is an encoder-decoder ASR model pretrained on large-scale speech-text corpora. Our fine-tuning strategy focuses specifically on the decoder for the following reasons:

(i) We aim to fully leverage the pretrained knowledge on the encoder side. We assume that Whisper's encoder, which is responsible for encoding acoustic information, has strong generalization ability across different languages. Therefore, rather than fine-tuning the encoder on a small amount of target data—which might risk degrading this generalization—we retain its pretrained capacity as much as possible.

(ii) We regard the decoder as the component that adapts to the target language. From the perspective of a traditional language model, the decoder primarily handles the mapping from acoustic features to linguistic representations. Since this process must reflect the characteristics of the target language (e.g., Hanzi or Pinyin language models), fine-tuning the decoder is a reasonable and effective choice.

### 2.2 LLaMA-Omni

With the recent rise of large audio–language models (LALMs), such as those proposed by Zhang et al. (2023) and Chu et al. (2024), we conducted additional experiments using the

| Source | Dataset Name | Usage | Duration(hr) | Description |
|---|---|---|---|---|
| Organizer provided | FSR-2025-Train | Train set | 62.0 | Official training corpus released by competition organizers. |
|  | FSR-2025-Record | Train set | 7.2 | Pilot-test subset |
|  |  | Test set | 0.8 |  |
|  | FSR-2025-Media | Train set | 1.6 | Pilot-test subset |
|  |  | Eval set | 0.2 |  |
|  |  | Test set | 0.2 |  |
| Web collected | Hakka Radio | Train set | 11.0 | Transcribed broadcast speech |
|  | Hakka E-Learning | Train set | 16.0 | Educational reading material |
| TTS generated | FSR-Website-TTS | Train set | 335.0 | Synthetic speech from VITS trained on FSR-2025-Train. |
|  | FSR-Media-TTS | Train set | 8.0 | Synthetic speech from VITS trained on FSR-2025-Train |
| Competition test sets | P-test | Test set | 1.0 | 1 hr subset (0.8 Record + 0.2 Media) |
|  | F-test | Test set | 10.0 | Final competition set. |

Table 1 Summary of all speech datasets used in this study

LLaMA-Omni model (Fang et al., 2024). The original architecture employs an 8B large language model; however, in our implementation, we replace it with a smaller 1B-parameter LLaMA (Dubey et al., 2025) variant to better accommodate limited GPU resources. The architecture integrates the Whisper-large-v3 encoder, and a linear adapter is inserted between the encoder and the LLM to align their feature dimensions by projecting the encoder output into the LLM's embedding space.

For fine-tuning, we follow a similar strategy by freezing the Whisper encoder to preserve its pretrained capacity for extracting meaningful speech representations. The adapter and LLM components are then trained jointly, enabling the model to adapt to the downstream task. This setup allows us to leverage the robust acoustic representations from the frozen encoder while focusing computational resources on adapting the modality-bridging adapter and the large language model to the target language domain. This configuration serves as a comparative baseline against our Whisper-only fine-tuning approach.

## 3 Data Sources

We first remove silence segments from all speech data to avoid adverse effects on model training. In addition, all corpora—including both the organizer-provided data and our self-collected resources—are resampled to 16 kHz to ensure consistency with the model requirements. Below, we describe our data augmentation and processing methods, as well as the training mechanism for data utilization. The overall pipeline of our data processing and model fine-tuning framework is illustrated in Figure 1. A detailed summary of all datasets used in this work is provided in Table 1.

### 3.1 FSR Hakka Challenge

As summarized in Table 1, the datasets used in this study can be grouped into three categories:(i) official FSR corpora released by the organizers, (ii) web collected resources, and (iii) TTS-generated synthetic speech. These corpora collectively provide complementary coverage of read and spontaneous Hakka speech, forming the basis for the experiments in Section 4.

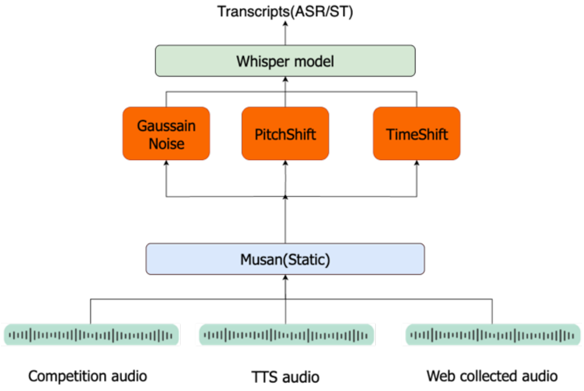

Figure 1: Overall architecture of the AS-SLAM system for Hakka ASR.

### 3.2 Web-Collected

We extracted speech-text pairs for both the Dapu and Zhaoan dialects from two publicly available online resources: Hakka E-learning and Hakka Radio.

The duration and usage of the web-collected datasets are summarized in **Table 1.** Specifically, The Hakka Radio corpus comprises approximately 9 hours of Dapu speech and 2 hours of Zhaoan speech, collected from broadcast programs in which native speakers discuss daily topics in a spontaneous conversational style.

These recordings exhibit diverse acoustic environments, speaker variations, and natural prosodic patterns. Owing to their broadcast nature, we hypothesize that the distribution of Hakka Radio more closely resembles that of the FSR-2025-Media subset. In contrast, the Hakka E-Learning corpus includes 8 hours each for Dapu and Zhaoan, originating from educational materials on the official Hakka E-learning platform. The utterances are primarily short, card-style sentences in which speakers read isolated words or short phrases aloud. Because of its clear articulation and relatively clean recording conditions, their corpus shares greater similarity with the FSR-2025-Record subset. Since the transcripts of both datasets are in Hanzi, they are used exclusively for the Hanzi track of the competition.

### 3.3 TTS-Generated

We adopt the Variational Inference Text-to-Speech (VITS) model (Kim et al., 2021) for speech synthesis, following the approach of Chen et al. (2023). During model training, we use both the official training data provided by the organizers and our self-collected resources. Separate TTS models are trained for the Dapu and Zhaoan dialects. For text prompts, we collect 150,000 example sentences from Hakka dictionary online published by the Hakka Affairs Council[1]. As summarized in Table 1, the official *FSR-2025-Train* corpus includes 123 speakers across both dialects and genders. From this pool, five speakers are randomly selected, and each generates 67 hours of speech, resulting in a total of 335 hours of synthetic data (denoted as FSR-Website-TTS). In addition, to enhance the coverage of media-style speech, we reuse the previously trained TTS models to perform speech synthesis using the transcriptions from the FSR-2025-Media dataset (1.6 hours of text). For each dialect, five speakers are randomly sampled, and each generates 1.6 hours of speech, resulting in a total of 8 hours of synthetic data (denoted as FSR-Media-TTS). Due to time constraints, only Hanzi transcriptions were used for speech synthesis.

## 4 Data Augmentation

We divide our data augmentation into two strategies—static and dynamic—following the two-stage approach proposed by Bhat et al. (2025), which are described as follows.

The overall workflow of both augmentation stages and their integration with the Whisper fine-tuning pipeline is illustrated in Figure 1. As shown in the figure, all audio sources—including competition data, TTS-generated data, and web-collected corpora—first pass through a static augmentation stage (MUSAN), followed by dynamic augmentations applied online during model training.

These two levels of augmentation jointly enhanced the model's robustness to noise, channel, variation, and acoustic mismatch across domains.

[1] https://hakkadict.moe.edu.tw/resource_download/

### 4.1 Static Data Augmentation

We employ the MUSAN (Snyder et al., 2015) toolkit, MetricAug (Wu et al., 2023), and the method proposed by Ko et al. (2023) for data augmentation, adding noise to clean speech before the training stage. The noise level is controlled by a randomly sampled signal-to-noise ratio (SNR) between 0 and 15 dB, following the configuration described in Pligin-SE (Chen et al., 2024). This static stage serves as the offline augmentation block shown in Figure 1 , ensuring that each input waveform exhibits realistic acoustic diversity prior to entering the dynamic augmentation pipeline.

### 4.2 Dynamic Data Augmentation

We apply the Audiomentations [2] toolkit for dynamic data augmentation. Unlike static data augmentation, this method is integrated directly into the training process. Before each sample is fed into the model, the following transformations are independently applied, following the configuration described in Dynamic Mixing (Choi et al., 2022) and Aligned Data Augmentation (Lam et al., 2021): GaussianNoise (minimum amplitude = 0.001, maximum amplitude = 0.015, probability = 0.3), TimeStretch (minimum rate = 0.9, maximum rate = 1.1, probability = 0.3), and PitchShift (minimum semitone = –2, maximum semitone = 2, probability = 0.3). This dynamic augmentation introduces greater variability during training, thereby improving the model's robustness.

## 5 Experimental Setup

### 5.1 FSR Challenge Setting

After the pilot test (stage 1 of the competition), our submitted model showed notably weaker performance on the FSR-2025-Media subset, suggesting that the model was less robust to the media distribution. To address this issue, we extended the training data by adding 1.6 hours of FSR-2025-Media and 7.2 hours of FSR-2025-Record to the original 40 hours of FSR-2025-Train. This new configuration is referred to as FSR-2025-Train-Plus, and served as the baseline for our final experiments. Building on this setup, we designed a series of extended experiments to examine the impact of additional data sources and augmentation strategies. Specifically, we trained three systems before the final submission deadline:

1. **FSR-2025-Train:** The official 40-hour training set only.
2. **FSR-2025-Train-Plus**: FSR-2025-Train+FSR-2025-Record+FSR-2025-Media with the hour combination described above.
3. **FSR**-2025-Train-Final: An extended configuration that further incorporates web-collected corpora and synthetic TTS speech.

All systems were trained with both static and dynamic data augmentation. The remaining experimental variants and comparative results are presented in Section 4.5 (Ablation Study).

In the Pinyin track, we did not incorporate self-collected corpora or TTS-generated data; instead, the system relied solely on the organizer-provided datasets. During the pilot test, training was conducted exclusively on the organizer-provided data with both static and dynamic augmentation strategies applied. In this setting, 20% of the 40-hour dataset was held out as the validation set, resulting in 32 hours of original speech data used for training. In the final stage, due to time constraints, we combined the 40-hour FSR-2025-Train dataset with the full 8 hours of FSR-2025-Record and the complete 2-hour FSR-2025-Media dataset for final model training, without a separate validation set; the model from the last training checkpoint was directly used for prediction. In both tracks, our final submission model was based on the Whisper architecture, fine-tuned at the decoder.

### 5.2 Model Training Details

For all competition submissions, we fine-tuned the Whisper-large-v3 model for 10 epochs, with a learning rate of 1e-5 following Whisper-LM (Zuazo et al., 2025) and an accumulated batch size of 64. For comparison, the LLaMA-Omni model was trained with a learning rate of 1e-4, an accumulated batch size of 12, and for 10 epochs

[2] https://github.com/iver56/audiomentations

in total. All experiments were conducted on Ubuntu using an NVIDIA RTX 3090 GPU.

## 6 Experiment Results

### 6.1 Pilot Test Results

In the pilot test stage, our system achieved a CER of 15.92% on the Hanzi track (Figure 2) and an SER of 20.49% on the Pinyin track (Figure 3) in the official student division results, ranking 7th and 3rd, respectively. The pilot test dataset consists of two subsets: FSR-2025-Record and FSR-2025-Media. To expand the preliminary data for the final competition, we constructed a new test set by combining 0.2 hours from FSR-2025-Media and 0.8 hours from FSR-2025-Record, referred to as P-Test. We then evaluated both the Whisper and LLaMA-Omni models, trained on the final competition training data, using this 1-hour Hanzi test set. The results are presented in Table 2. As shown, under limited training data conditions, Whisper still clearly outperforms LLaMA-Omni. We attribute this to the fact that automatic speech recognition (ASR) is Whisper's original pretraining objective, whereas LLaMA-Omni is designed for more general multimodal purposes. Consequently, Whisper holds a stronger advantage in ASR-specific tasks.

### 6.2 Final Competition Results

In the finals, our system achieved a CER of 15.73% on the Hanzi track and, for the Pinyin track, a WER of 20.68% and a tone-removed WER (WER^) of 13.82% in the official student division results, ranking 6th and 4th, respectively. The rankings and corresponding error rates are summarized in Table 3.

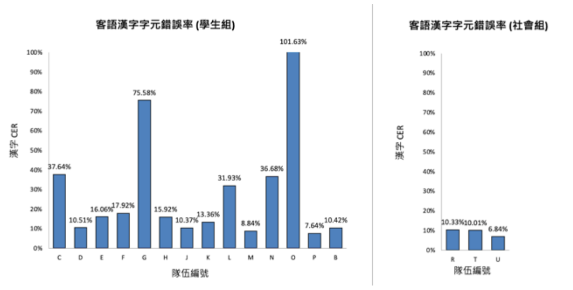

Figure 2: CER results and rankings on the Hanzi track in the pilot test. Our team, labeled as "H," participated in the student division.

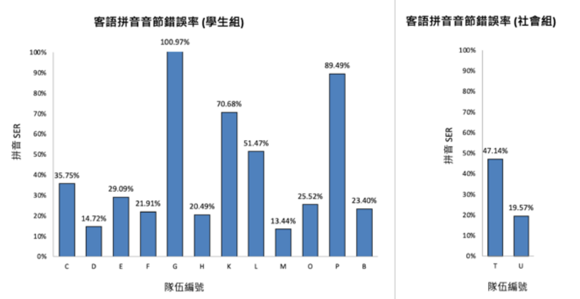

Figure 3: SER results and rankings on the Pinyin track in the pilot test. Our team, labeled as "H," participated in the student division.

| Model | CER |
|---|---|
| LLaMA-Omni | 4.24 |
| Whisper | 2.54 |

Table 2: CER Results on the Hanzi Track of the P-Test for LLaMA-Omni and Whisper. Both models were trained under the same data configuration as FSR-2025-Train-Final

### 6.3 Ablation Study

We conducted a series of ablation experiments to examine the contribution of each data source, as summarized in **Table 4**. We observed that in the P-Test, simply adding the FSR-Website-TTS data led to a performance decline, whereas adding the FSR-Media-TTS data resulted in improved performance on the P-Test but showed the opposite trend on the F-Test. All configurations employed both static (MUSAN) and dynamic (Audiomentations) data augmentation, and were evaluated on the 1-hour **P-Test** and the **F-Test**.

Moreover, combining both types of TTS data did not yield any complementary effect on either test. We speculate that this discrepancy may be attributed to the distributional differences between the two types of TTS data. In contrast, both Hakka Radio and Hakka E-Learning contributed significant improvements on the P-Test and F-Test, with even greater gains when the two were combined.

Notably, on the F-Test, using only these two datasets outperformed all other data combinations. In the Web-collected data experiments, we further observed that adding Hakka Radio yielded better performance than adding Hakka E-Learning. Interestingly, incorporating only Hakka E-Learning caused a performance drop on the P-Test but showed improvement on the F-Test.

When both Hakka E-Learning and Hakka Radio were added together, the performance on the P-Test was slightly worse than using Hakka Radio alone, whereas on the F-Test, the two datasets exhibited complementary effects. We speculate that this is because the data d istribution of Hakka E-Learning differs considerably from that of general media data, while Hakka Radio demonstrates higher generalizability. This effect may also be influenced by the higher proportion of media data in the P-Test compared with the F-Test.

| Hanzi | | Pinyin | | |
|---|---|---|---|---|
| Rank | CER | Rank | WER | WER^ |
| 6 | 15.73% | 4 | 20.68% | 13.82% |

Table 3: In the final results of the Hanzi and Pinyin tracks, the evaluation metric for the Hanzi track is CER, while that for the Pinyin track is WER. WER^ denotes the WER evaluated after tone removal.

| | CER | |
|---|---|---|
| | P-Test | F-Test |
| FSR-2025-Train | 20.10% | 27.79% |
| FSR-2025-Train-Plus | 3.40% | 17.54% |
| + FSR-Website TTS (1) | 3.78% | 17.07% |
| + FSR-Media TTS (2) | 3.29% | 17.33% |
| + (1) + (2) | 3.5% | 17.50% |
| + Hakka Radio (3) | 3.07% | 15.04% |
| + Hakka E-Learning (4) | 3.72% | 16.71% |
| + (3) + (4) | 2.55% | 14.58% |
| + (1) + (2) + (3) + (4) | 2.54% | 15.73% |

Table 4: Comparison of training solely on the original FSR 2025 dataset versus augmenting it with TTS, Hakka E-Learning, and Hakka Radio, evaluated on the Hanzi track of 1-hour Pilot test in the pilot test (P-Test) and Final test set (F-Test) in terms of CER.

| | WER |
|---|---|
| | F-Test |
| FSR-2025-Train | 32.60% |
| FSR-2025-All | 20.68% |

Table 5: Comparison of the results on the Final Pinyin Track using the training data from the preliminary round (FSR-2025-Train) and the training data used for the Final Pinyin Track (FSR-2025-All).

For the Pinyin track, we used all FSR-2025-Media and FSR-2025-Record data in the final stage, while keeping the remaining configurations identical to FSR-2025-Train-Plus. We refer to this training set as FSR-2025-All. The models trained with both the pilot test training data and this final training set were evaluated on the final test set, and the WER results are shown in Table 5. We observed a significant improvement after adding the additional training data, suggesting that future research could further enhance model performance by expanding the amount of Pinyin training data.

## 7 Conclusion

In this competition, we investigated the use of various Hakka datasets and conducted preliminary experiments with existing ASR models such as Whisper and LLaMA-Omni. Our results provide initial evidence that Whisper may outperform LLaMA-Omni for ASR tasks in low-resource languages. In the finals, we achieved 6th place in the Hanzi track and 4th place in the Pinyin track. Moving forward, we plan to explore integrating data across different dialects and experimenting with more model combinations, with the goal of making further progress in low-resource language research.

## Acknowledgments

We would like to express our sincere gratitude to Dr. Hung-Shin Lee from North Co., Ltd., Taiwan, for providing the web-collected data.

## References

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *International Conference on Machine Learning* (pp. 28492-28518). PMLR.

L.-W. Chen, K.-C. Cheng, and H.-S Lee. 2023. The North System for Formosa Speech Recognition Challenge 2023. *In Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023).*

H.-C. Lu, C.-C. Wang, J.-K. Lin, and T.-H. Lo 2023. The NTNU ASR System for Formosa Speech Recognition Challenge 2023. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023).*

D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu. 2023. SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. In *Findings of the Association for Computational Linguistics: ACL 2023*, (pp. 1469-1483).

Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, and others. 2024. Qwen2-Audio Technical Report. *arXiv preprint* arXiv:2407.10759.

Q. Fang, S. Guo, Y. Zhou, Z. Ma, S. Zhang, and Y. Feng. 2024. Llama-Omni: Seamless Speech Interaction with Large Language Models. *In Proceeding of the 13th International Conference on Learning Representations (ICLR 2025).*

A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, and R. Ganapathy 2024. The Llama 3 Herd of Models. *arXiv preprint* arXiv:2407.12345.

J. Kim, J. Kong, and J. Son. 2021. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. In *Proceedings of the 38th International Conference on Machine Learning (ICML).*

P.-K. Chen, B.-J. Huang, C.-T. Chen, H.-M. Wang, and J.-C. Wang. 2023. Enhancing Automatic Speech Recognition Performance Through Multi-Speaker Text-to-Speech. *In Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023).*

C. Bhat, and H. Strik. 2025. Two-Stage Data Augmentation for Improved ASR Performance for Dysarthric Speech. *Computers in Biology and Medicine*, 189, 109954.

D. Snyder, G. Chen, and D. Povey. 2015. MUSAN: A Music, Speech, and Noise Corpus. *arXiv preprint arXiv:1510.08484.*

Y.-T. Wu, and C.-C. Lee. 2023. MetricAug: A Distortion Metric-Lead Augmentation Strategy for Training Noise-Robust Speech Emotion

Recognizer. *In Proceedings of INTERSPEECH 2023. International Speech Communication Association (ISCA).*

T. Ko, V. Peddinti, D. Povey, and S. Khudanpur. 2017. A Study on Data Augmentation of Reverberant Speech for Robust Speech Recognition. *In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Y. Chen, Z. Cui, Y. Gao, J. Feng, C. Deng, and S. Zhang. 2024. Plugin Speech Enhancement: A Universal Speech Enhancement Framework Inspired by Dynamic Neural Network. *arXiv preprint* arXiv:2402.12746.

S. Choi, Y. Lee, J. Park, H. Y. Kim, B.-Y. Kim, Z.-Q. Wang, and S. Watanabe. 2022. An Empirical Study of Training Mixture Generation Strategies on Speech Separation: Dynamic Mixing and Augmentation. I*n Proceedings of the 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC).*

T. K. Lam, M. Ohta, S. Schamoni, and S. Riezler. 2021. On-the-Fly Aligned Data Augmentation for Sequence-to-Sequence ASR. *Proceedings of Interspeech 2021*, (pp. 1299-1303).

X. de Zuazo, E. Navas, I. Saratxaga, and I. Hernáez Rioja. 2025. Whisper-LM: Improving ASR Models with Language Models for Low-Resource Languages. *arXiv preprint* arXiv:2503.23542.