A Whisper-Based System with Multi-Faceted Data Augmentation for Low-Resource Language

Pin-Cheng Chen*[1] Yu-Chi Chen*[1] Chia-Chun Liang*[1]
Cheng-Yu Lin*[1] Ping-Juei Tsai*[1] Wei-Yun Ma†[1]

[1]Institute of Information Science, Academia Sinica, Taipei, Taiwan
b10102102@ntu.edu.tw, chi.mg10@nycu.edu.tw, d14948004@ntu.edu.tw
gary2004930518@gmail.com, allcare.c@nycu.edu.tw, ma@iis.sinica.edu.tw

Abstract

This paper presents a comprehensive approach for the Formosa Speech Recognition Challenge 2025 (FSR-2025), targeting automatic speech recognition (ASR) for the under-resourced Dapu and Zhao'an dialects of Taiwanese Hakka. Our method integrates data augmentation and robustness techniques, including SpecAugment, dialect-aware special tokens, text-to-speech (TTS) augmentation, noise/reverberation mixing, and speed perturbation, to mitigate data scarcity and domain mismatch. Experiments on the official FSR-2025 datasets show consistent improvements in both character error rate (CER) and word error rate (WER). Extensive ablation studies further confirm that each component contributes positively. These results offer a practical path toward robust ASR for under-resourced Hakka dialects and suggest broader applicability to other low-resource languages.

Keywords: Automatic Speech Recognition, Data Augmentation, Low-resource, Taiwanese Hakka

1 Introduction

Automatic Speech Recognition (ASR) has made remarkable progress in recent years, driven by large-scale speech corpora and powerful deep learning models. End-to-end pipelines have become standard, where Connectionist Temporal Classification (CTC)-based models provide efficient monotonic alignment (Graves et al., 2006), and attention/Transformer frameworks enhance long-range modeling (Chan et al., 2016; Dong et al., 2018; Barrault et al., 2023). For scenarios with limited labeled data, self-supervised pretraining (e.g., wav2vec, HuBERT) yields substantial gains by learning robust acoustic representations from unlabeled audio (Schneider et al., 2019; Baevski et al., 2020; Hsu et al., 2021). Building on this foundation, large models like Whisper leverage multilingual, multitask training to generalize across diverse languages and domains (Radford et al., 2022). More recently, Multimodal Large Language Models (MLLMs) have extended this paradigm by processing speech and text within a unified framework, facilitating cross-modal reasoning and transfer (Rubenstein et al., 2023; Zhang et al., 2023).

However, these advances are unevenly distributed: most data and models target high-resource languages (e.g., English, Mandarin), whereas minority and dialectal languages remain underserved. Taiwanese Hakka is a low-resource Sinitic language with multiple dialects; among them, Dapu and Zhao'an are particularly under-resourced. These challenges make it difficult for standard ASR models to achieve satisfactory performance, creating a significant technological gap for their speakers. The Formosa Speech Recognition Challenge 2025 (FSR-2025) directly addresses this issue by providing a benchmark dataset to foster research in this area.

To address these challenges, our methodology centers on a multifaceted data augmentation strategy, combining SpecAugment, text-to-speech (TTS) synthesis, noise and reverberation mixing, and speed perturbation. We also introduce dialect-aware special tokens to guide the model in distinguishing between the Dapu and Zhao'an dialects. The effectiveness of this approach and the contribution of each component are systematically evaluated through a se-

*Equal contribution. Authors are listed in alphabetical order by last name, except the advisor.

†Advisor and corresponding author.

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 489–498
November 20-22, 2025 ©2025 Association for Computational Linguistics

ries of experiments and ablation studies on the official dataset, as detailed in the subsequent sections.

2 Dataset and Task Definition

Our study is based on the HAT-Vol-2 corpus, provided by the organizers. The corpus contains roughly **70** hours of audio from approximately **100** speakers across Taiwan and is divided into **three** official splits: *train*, *evaluation*, and *final-release* (test set).

The FSR-2025 challenge structure the task into two parallel tracks, each corresponding to a different orthography. This dual-track system defines the output targets for our models and the metrics for evaluation.

2.1 Orthography and Evaluation Tracks

- **Track 1: Recommended Hakka Characters.** This track uses a set of Han characters promoted by the Taiwanese Ministry of Education for writing Hakka. While leveraging semantic context familiar to readers of Sinitic languages, these characters often lack a one-to-one phonetic correspondence. For instance, the character '行' can have multiple pronunciations depending on the context. Performance on this track is measured by **Character Error Rate (CER)**.

- **Track 2: Hakka Pinyin System.** This track employs a phonemic transcription system that precisely represents initials, finals, and tones. It clearly distinguishes dialectal variations (e.g., the word "person" (人) is transcribed as `ngin113` in Dapu vs. `ngin53` in Zhao'an). However, this system is unfamiliar to most native speakers. Performance on this track is measured by **Word Error Rate (WER)**, where each Pinyin syllable is treated as a word.

In addition to the official data, we employ VoxHakka, a multi-accent, multi-speaker text-to-speech (TTS) system (Chen et al., 2024)[1], to synthesize additional Hakka speech. This mitigates data scarcity and expands both lexical and speaker coverage; our generation policy and settings are detailed in Section 4.3.

3 Model

3.1 Whisper

The Whisper model, developed by OpenAI (Radford et al., 2022), is an end-to-end ASR system with strong multilingual performance. Our study builds upon the **`whisper-large-v3-taiwanese-hakka`** checkpoint (hakka-whisper) (FormoSpeech, 2025), already fine-tuned on six Hakka dialects, and we further fine-tuned it on Dapu and Zhao'an data for FSR-2025. It's quite notable that the further evaluation revealed a divergence between character-based (Track 1) and pinyin-based (Track 2) transcription: the adapted checkpoint improved character recognition, but the original Whisper model performed better on pinyin, likely due to its broader phonetic generalization. This also highlights a trade-off between dialect adaptation and phonetic robustness.

3.2 MLLM-based model

In addition, we evaluate LLM-based approaches for speech transcription. Specifically, we use Kimi-Audio, which is based on the Qwen architecture, as the backbone of the language model (KimiTeam et al., 2025). Kimi-Audio employs a 12.5 Hz audio tokenizer and has been trained on large-scale Chinese text and audio data; it shows strong performance on Mandarin ASR benchmarks—indicating robust capability for Sinitic phonetic and orthographic patterns. This setup allows us to probe how well a large Chinese-trained LLM can transfer its knowledge to low-resource dialects such as Dapu and Zhao'an Hakka, and whether the model can effectively leverage its linguistic knowledge to compensate for the scarcity of labeled speech data.

4 Metholodgy

Due to the different effects on each track, we applied different methods to each of them. The utilized results are summarized in Table 1.

[1] `https://voxhakka.github.io/`

Table 1: Methods for Track 1 & 2

Track	SpecAugment	Special token	TTS
Track 1	✓	✓	✓
Track 2		✓	✓

4.1 SpecAugment

SpecAugment is a simple yet effective method that operates directly on the log-mel spectrogram (Park et al., 2019). Instead of relying on additional data, it improves model robustness toward noise by applying several types of transformations: time warping, frequency masking, and time masking. Time warping distorts the spectrogram along the temporal axis, while frequency masking and time masking randomly remove consecutive frequency channels or time steps, respectively. In our setting , we adopted frequency masking and time masking with a progressive enhancement strategy during training (Li et al., 2022; Lu and Li, 2024), which is also applied in images with good performance (Jarca et al., 2024).

4.2 Special token

In Whisper, special tokens can be utilized to control specific attributes of the speech recognition process such as task type, target language, and timestamping behavior (Radford et al., 2022). In practice, the token serves as a high-level cue for the model, guiding the model's acoustic and lexical predictions. Recent studies demonstrate that leveraging special tokens, which is often termed *prompt-based control*, can significantly improve Whisper's performance, particularly in low-resource or unseen language scenarios (Peng and Yan, 2023; Yang et al., 2024; Huang et al., 2025). For instance, studies have shown that introducing explicit prompts, such as language-family tags or even learnable soft prompts, helps guide the model toward more accurate transcriptions for underrepresented languages (Yang et al., 2025). Inspired by these findings, our work investigates a similar approach by introducing dialect-aware special tokens. We hypothesize that providing an explicit cue to distinguish between the closely related Dapu and Zhao'an dialects will enable the model to better activate dialect-specific acoustic and linguistic knowledge, thereby improving recognition accuracy for both.

4.3 Text-to-Speech

We synthesize additional Hakka speech with VoxHakka, a YourTTS-based, multi-speaker, multi-dialect TTS system for Taiwanese Hakka (Chen et al., 2024). We adopt a twofold generation policy with external and internal sources. On top of that, each transcript is generated in **three** voices, sampled randomly from VoxHakka's multi-speaker bank.

External sources *External* denotes text not included in the official data transcripts. We collect sentences from the Ministry of Education Hakka Dictionary[2] and the online teaching materials released by the Hakka Affairs Council (HAC)[3]. Given Han-character inputs, VoxHakka synthesizes the corresponding waveforms and generates pinyin labels, which are not provided by these sources. The synthesized utterances enrich the training set with terms and sentences that are rarely observed in spontaneous speech.

Internal sources *Internal* denotes text derived from official data transcriptions. We employ two strategies:

1. **Tokenized rare-term augmentation** We observed that official evaluation set often contains proper nouns and other low-frequency words that are scarce in the training set, making them a common source of recognition errors. To mitigate this, we first identify these rare lexical items from the training transcripts using a GPT-4o model (Hurst et al., 2024) guided by a carefully designed few-shot prompt. After de-duplication, each unique term is synthesized into an audio clip using the VoxHakka TTS system. This provides the ASR model with explicit acoustic examples of rare and potentially out-of-vocabulary (OOV) terms.

2. **Voice conversion** The released training set contains many repeated prompts recorded by multiple speakers–some sentences are read by up to 14 speakers–while

[2] https://hakkadict.moe.edu.tw

[3] https://elearning.hakka.gov.tw

other sentences occur only once or twice. Motivated by prior findings that Voice conversion (VC) based speaker augmentation improves ASR in low-resource settings (Baas and Kamper, 2021), we apply VC to under-covered sentences to increase speaker diversity: each such sentence is uttered by at least three distinct speakers.

5 Experiments

For evaluation, we adopted the official scoring mechanism provided by the competition[4]. Specifically:

- **Track 1:** Character Error Rate (CER) was used as the primary metric.
- **Track 2:** Word Error Rate (WER) was used as the primary metric.

5.1 Models

Our initial experiments focused on the Kimi-Audio model. It was fine-tuned on the *FSR-2025-train* set and evaluated on the *FSR-2025-evaluation* set. The system obtained a CER of 51.87% on Track 1 and a WER of 89.49% on Track 2. Notably, the outputs contained several abnormal generation artifacts[5]; manually correcting for these reduced the CER to 33.47%.

We then evaluated Kimi-Audio-Instruct, a variant trained predominantly on Mandarin data (KimiTeam et al., 2025), under the same configuration. This model yielded a CER of 45.70% on Track 1, which improved to 28.27% after correcting for the same abnormal outputs. For comparison, a hakka-whisper baseline trained with an identical setup achieved a markedly lower CER of 7.64%. Thus, while Kimi-Audio-Instruct outperformed the original Kimi-Audio, both models remained significantly behind the specialized hakka-whisper system.

Furthermore, our error analysis of both Kimi-Audio models revealed a particular weakness in processing longer utterances and tokens rare in the training data (e.g., proper nouns and transliterated names). These conditions not only yielded substantially higher error rates but also occasionally triggered the generative artifacts noted above, severely degrading overall performance.

To investigate the effect of data distribution, we conducted a controlled experiment. We created a new data split by merging the *FSR-2025-train* and *FSR-2025-evaluation* sets. From this combined pool, we held out 20% as a new development set and randomly sampled 5,000 utterances for a test set. Under this controlled setting, Kimi-audio achieved a greatly improved performance of **6.13% CER** (Track 1) and **7.56% WER** (Track 2). This result, summarized in Table 2, indicates that the model's performance improves dramatically when the evaluation data distribution is well-represented in its training data—especially concerning rare words and proper nouns.

Despite this promising result, we prioritized the Whisper-based system for the final challenge submission. This decision was based on the observed instability (i.e., the generation of abnormal outputs) and the higher computational cost of the Kimi-based models, which posed practical risks when facing an unknown final test set. Nevertheless, our findings suggest that MLLM-based approaches like Kimi-Audio hold considerable promise for future work, provided sufficient data coverage and improved model stability.

Table 2: Performance of Kimi-audio under a controlled split (train+evaluation merged; 20% held out; 5,000-item test sample).

Model	**CER (Track 1)**	**WER (Track 2)**
Kimi-Audio	6.13%	7.56%

5.2 Methods

5.2.1 Evaluation Setup

After confirming the model, to evaluate the usability of the proposed methods for ASR tasks, we adopted the following data split for testing on a Whisper-like model:

- **Training data:** 90% of (*FSR-2025-train* + 80% of *FSR-2025-evaluation*).
- **Validation data:** 10% of (*FSR-2025-train* + 80% of *FSR-2025-evaluation*).

[4] https://github.com/yfliao/FSR-2023-Hakka-ASR-Scoring

[5] For examples, see https://github.com/MoonshotAI/Kimi-Audio/issues/101

- **Testing data:** 20% of *FSR-2025-evaluation.*

We trained a *hakka-whisper* model on the training data, which serves as the **baseline** for comparison against the results of the different methods.

5.2.2 SpecAugment

In our setting, we set the following corrected progressive strategy: in the early stage, SpecAugment was applied with a probability of 30%, which means each batch would have a 30% chance of being perturbed using a time mask of 40 frames and a frequency mask of 14 bins. In the middle stage, the probability was increased to 50% with masking parameters set to 60 frames for time masking and 20 bins for frequency masking. Finally, in the late stage, the probability was set to 70% , with stronger augmentation using 80 frames for time masking and 27 bins for frequency masking.

The 30% initial probability (rather than 10%) provides sufficient augmentation from the start to prevent early overfitting, while the 70% final probability (rather than 80%) avoids over-augmentation that could harm model convergence. This balanced progression aligns with curriculum learning principles where moderate difficulty increases lead to better generalization (Jarca et al., 2024).

The progressive augmentation probability at step t is defined as:

$$p(t) = \begin{cases} 0.3 & \text{if } t/t_{max} < 0.3 \\ 0.5 & \text{if } 0.3 \leq t/t_{max} < 0.7 \\ 0.7 & \text{if } t/t_{max} \geq 0.7 \end{cases}$$

where t_{max} represents the total training steps.

At first, we compared the baseline to the one with noise mask. We conducted this test on Track 1. Surprisingly, the baseline achieved a CER of 4.47%, while applying the proposed method reduced the CER to 3.77% at 5000-step training. This corresponds to a relative reduction of 15.66% in CER, indicating a substantial improvement.

Secondly, we evaluated the effect of the progressive and the stationary enhancement. This time we conduct the experiment on Track 2. As shown in Fig. 1, we observed that under the stationary setup, the error rate plateaued after approximately 4,000 steps. By 13,000 steps, the validation WER was still around 32%. In contrast, the progressive setup continued to improve throughout training. After 20,000 steps, the validation WER reached 5.92% (final value: 6.35%).

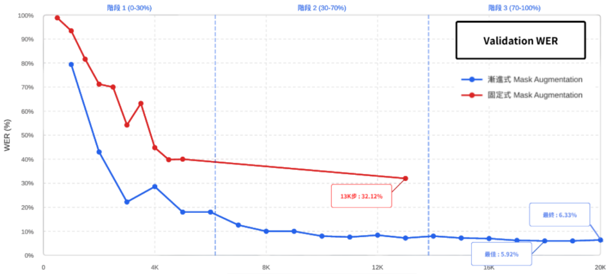

Figure 1: Comparison of validation WER between stationary (red-line) and progressive (blue-line) enhancement setups on Track 2. We utilized *Mask* Augmentation. Under the stationary setup, the error rate plateaued after around 4,000 steps and remained at about 32% by 13,000 steps. In contrast, the progressive setup continued to improve throughout training, reaching a validation WER of 5.92% at 20,000 steps (final value: 6.35%).

On the test set, the stationary setup yielded a WER of 46.53%, while the progressive setup achieved a significantly lower WER of 21.59%. This result validates our 30%-50%-70% progressive schedule, where the moderate initial augmentation (30% vs. 10%) allows faster convergence while the conservative final stage (70% vs. 80%) maintains stability.

5.2.3 Source-Aware Speed Perturbation

Unlike traditional uniform speed augmentation, we implement a source-aware speed distribution strategy that adapts to the inherent characteristics of different audio sources in the FSR-2025 dataset—which we identified based on the accompanying metadata—as shown in Table 3. Our analysis reveals significant heterogeneity: media sources (broadcasts, podcasts, etc.) exhibit fast speaking rates (1.1-1.3x relative to conversational speech) with high audio quality, while recorded conversational data show slower rates (0.8-1.0x) but suffers from reverberation and environmental noise.

Design rationale for asymmetric speed distributions:

- **Media sources** are heavily biased to-

Table 3: Source-aware speed factor distribution across different types of sources. The table summarizes the relative proportions of playback speed factors observed for *Media Source*, *Recorded Source*, and *General Source*. Overall, Media Source tends to concentrate in the slower range (0.70–1.00x) with a peak at 0.75x, while Recorded Source shifts toward faster speeds (0.90–1.15x) peaking at 1.00x. General Source covers a wider range (0.85–1.20x) and peaks at 1.05x.

Speed Factor	**Media Source**	**Recorded Source**	**General Source**
0.70x	15%	-	-
0.75x	25%	-	-
0.80x	20%	-	-
0.85x	15%	-	5%
0.90x	10%	10%	10%
0.95x	10%	15%	15%
1.00x	5%	20%	20%
1.05x	-	25%	20%
1.10x	-	20%	15%
1.15x	-	10%	10%
1.20x	-	-	5%
Range	0.70-1.00	0.90-1.15	0.85-1.20
Peak	0.75x (25%)	1.00x (20%)	1.05x (20%)

wards slowdown factors (75% probability in 0.70-0.85x range) to compensate for their inherently fast speaking rate. This prevents the model from overfitting to rapid speech patterns that are rare in target applications.

- **Recorded sources** receive balanced bidirectional augmentation with a peak at 1.0x (20%) and symmetric distribution (1.00-1.15x speedup, 0.90-0.95x slowdown). This addresses the slower baseline rate while maintaining natural variation.
- **General sources** adopt the widest range (0.85-1.20x) with a slight speedup bias (peak at 1.05x, 20%), maximizing robustness to diverse speaking rates in unknown data.

The progressive speed augmentation schedule follows three distinct phases (see Table 4), synchronized with SpecAugment and noise augmentation to achieve curriculum learning effects.

Table 4: Progressive learning schedule across different training phases. The early stage (0–30% epochs) adopts conservative settings for foundation learning, the middle stage (30–70%) uses maximum augmentation for robustness building, and the late stage (70–100%) reduces augmentation to stabilize convergence.

Parameter	**Early Stage**	**Middle Stage**	**Late Stage**
Objective	Warm-up	Intensive	Stabilization
Epoch Range	0–30%	30–70%	70–100%
SpecAug Prob	0.3	0.6	0.4
Speed Prob	0.3	0.6	0.4
Noise Prob	0.2	0.5	0.3
Mask Intensity	0.7x	1.0x	1.2x

Rationale for progressive probability schedule: The middle stage employs the highest augmentation probability (0.6) as the model has developed sufficient acoustic modeling capacity to benefit from aggressive data perturbation while avoiding early-stage confusion. The late stage deliberately reduces augmentation intensity (0.4) to prevent over-regularization that could harm fine-grained learning of tonal patterns, critical for Hakka's complex tone system. Noise augmentation follows a particularly conservative schedule (0.2 → 0.5 → 0.3) because excessive noise can disrupt the *fundamental frequency* (F0) contour information essential for tone discrimination in Hakka.

This coordinated multi-type augmentation strategy, validated through 30% relative CER reduction compared to uniform augmentation (from 4.47% to 3.13%), demonstrates the effectiveness of curriculum-based training for low-resource ASR.

Next, as shown in Table 5 with Track 1, we want to know the effects from the three masks: spectrum deformation (Spec), noise addition (Noise), and speed variation (Speed).

Overall, adding Speed augmentation leads

Table 5: CER results for different augmentation settings across training steps.

Setting	Stage	Step	CER (%)
Baseline	-	20k	4.33
Spec+Noise	Early (30%)	3k	4.02
Spec+Noise	Early (30%)	4k	3.91
Spec+Noise	Mid (50%)	5k	3.47
Speed+Spec+Noise	Early (30%)	3k	3.71
Speed+Spec+Noise	Mid (50%)	6k	3.40
Speed+Spec+Noise	Late (70%)	12k	**3.13**

to a consistent decrease in CER as the training steps increase. The best result, achieved at 12000 steps with Speed+Spec+Noise, shows an improvement of approximately 27.7% over the baseline (from 4.33% down to 3.13%).

5.3 Special tokens

We design an enhanced dialect conditioning mechanism by injecting dialect-specific tokens into the decoding process:

Dialect Token Insertion & Detection. We define a set of dialect tokens: <| 大埔腔 |>, <| 詔安腔 |>, and <| 未知腔 |>. Since the training datasets are labeled with the corresponding dialect, our system detects the dialect of each audio file based on rule-based metadata during preprocessing. And the system would prepends the appropriate token to the transcription text.

Balanced Sampling. To prevent majority dialects from dominating training, we employ a balanced sampling strategy. Depending on the configuration, batches can be drawn either equally from each dialect (equal strategy), or weighted to favor minority dialects.

We integrated both dialect tokens and balanced training and the results are presented in Table 6 and Table 7.

Table 6: CER results on Track 1 for with and without the Special token setting.

Setting	Step	CER (%)
Baseline	20000	4.33
Special token	30000	**3.48**

Table 7: WER results on Track 2 for with and without the Special token setting.

Setting	Step	WER
Baseline (BS) (hakka-whisper)	20000	9.31%
Special token (hakka-whisper)	30000	13.09%
Special token (openai whisper)	30000	12.82%

Table 8: Summary of synthesized TTS datasets

Data source	Nb. of Entries
External	64,950
Internal	28,046
Total	**92,996**

5.4 Text-to-Speech

To mitigate data scarcity, we significantly expanded our training set with synthesized audio from various TTS sources, following the strategies detailed in Section 4.3. A summary of the augmented data is provided in Table 8.

Ablation Study on Short-Utterance Mismatch. A key concern with synthetic data is the potential for distributional mismatch with the official dataset. We identified a significant difference in utterance length: our externally sourced, dictionary-based TTS data consists of very short clips (mean duration of 1.32 s), whereas utterances in the official training data are much longer (mean duration of 8.4 s).

To assess whether injecting a large volume of short clips would degrade model performance, we conducted a targeted ablation study. We created a data subset named **Half-Dict**, comprising approximately 38k dictionary-based utterances (totaling around 7 hours), carefully balanced between the Dapu and Zhao'an dialects.

The results, presented in Table 9, show that the inclusion of short TTS clips does not degrade performance; in fact, it provides a slight improvement in CER. We then supposed that these additional dictionary-derived utterances can expand lexical coverage, allowing the model to encounter more of the vocabulary likely to appear in the final test set.

6 Results

Based on our experiments and ablation studies, we submitted two distinct systems to the FSR-2025 challenge. The final configurations on the final dataset are summarized in Ta-

Table 9: CER result for the short-utterance (dictionary TTS) ablation.

Setting	CER (%)
Baseline	4.33
+ Half Dict.	**4.18**

Table 10: The final result on the Final dataset for the FSR-2025-challenge.

Track	Baseline	Final Result
Track 1 (CER)	10.45 %	**8.99%**
Track 2 (WER)	20.02%	**19.22%**
Track 2 (WER) (no tone value)	-	**12.36%**

ble 10.

6.1 Track 1: Recommended Characters (CER)

For Track 1, our system, which is built upon `hakka-whisper` and enhanced with our full suite of data augmentation and dialect conditioning techniques, achieved a final CER of 8.99% on the official test set. This performance secured the second-place rank among all participating teams (Fig. 2) and represents a significant 19.8% relative error reduction compared to the third-place team.

6.2 Track 2: Hakka Pinyin (WER)

For Track 2, our final system used the general-purpose `whisper-large-v3`, which outperformed the Hakka-fine-tuned variant in development. On the official test set, it achieved a WER of 19.22%, ranking third among all teams (Fig. 3). Under the competition' s tone-ignored metric, the error rate further decreased to 12.36%.

7 Conclusion

This work presented a comprehensive approach for the ASR task of the under-resourced Dapu and Zhao'an dialects of Taiwanese Hakka for the FSR-2025 challenge. By integrating multiple data augmentation and robustness techniques including SpecAugment, dialect-aware special tokens, TTS augmentation, noise/reverberation mixing, and speed perturbation, our systems effectively mitigated the challenges posed by limited training data and domain mismatch.

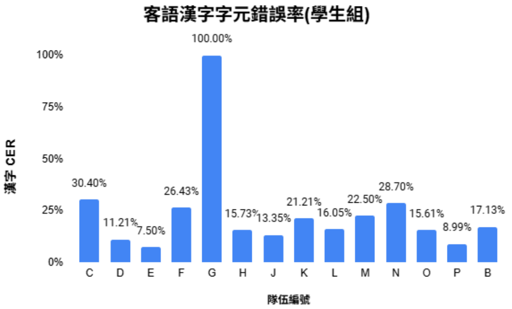

Figure 2: Official CER results for the FSR-2025 student group. Our team ranked second.

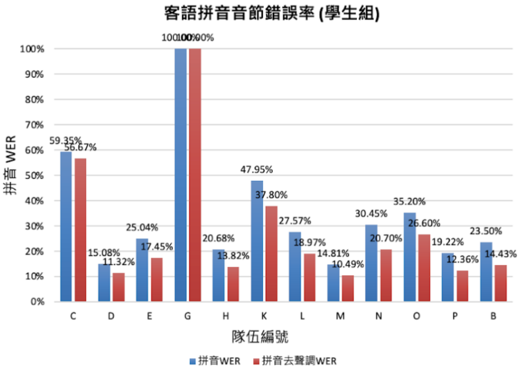

Figure 3: Official WER results for the FSR-2025 student group. Our team ranked third.

This report also consisted of experimental results that demonstrated substantial improvements in both CER and WER, with our Track 1 system achieving 8.99% CER (2^{nd} place in academic groups) and our Track 2 system achieving 19.22% WER (3^{rd} place in academic groups), further reduced to 12.36% without considering tone value. Ablation studies confirmed that each component contributed positively to overall performance.

These results highlight the effectiveness of a combined augmentation and robustness strategy for low-resource ASR, providing a practical path toward robust recognition for Hakka dialects and offering insights applicable to other under-resourced languages.

Acknowledgments

We thank the National Center for High-performance Computing (NCHC), Taiwan, for providing essential computational and storage resources.

References

Matthew Baas and Herman Kamper. 2021. Voice conversion can improve asr in very low-resource settings. *arXiv preprint arXiv:2111.02674*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023. Seamlessm4t: massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4960–4964. IEEE.

Li-Wei Chen, Hung-Shin Lee, and Chen-Chi Chang. 2024. Voxhakka: A dialectally diverse multi-speaker text-to-speech system for taiwanese hakka. In *2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6.

Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5884–5888. IEEE.

FormoSpeech. 2025. whisper-large-v3-taiwanese-hakka. `https://huggingface.co/formospeech/whisper-large-v3-taiwanese-hakka`. Accessed: 2025-09-10.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Shao-Syuan Huang, Kuan-Po Huang, Andy T Liu, and Hung-Yi Lee. 2025. Enhancing multilingual asr for unseen languages via language embedding modeling. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Andrei Jarca, Florinel-Alin Croitoru, and Radu Tudor Ionescu. 2024. Cbm: Curriculum by masking. *arXiv preprint arXiv:2407.05193*.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. 2025. Kimi-audio technical report.

Rui Li, Guodong Ma, Dexin Zhao, Ranran Zeng, Xiaoyu Li, and Hao Huang. 2022. A policy-based approach to the specaugment method for low resource e2e asr. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 630–635. IEEE.

Hongxuan Lu and Biao Li. 2024. Sample adaptive data augmentation with progressive scheduling. *arXiv preprint arXiv:2412.00415*.

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. *arXiv preprint arXiv:1904.08779*.

Puyuan Peng and Brian Yan. 2023. Prompting the hidden talent of web-scale speech models for zero-shot task generalization.

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

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Chih-Kai Yang, Kuan-Po Huang, and Hung-yi Lee. 2024. Do prompts really prompt? exploring the prompt understanding capability of whisper. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 1–8. IEEE.

Hongli Yang, Yizhou Peng, Hao Huang, and Sheng Li. 2025. Adapting whisper for parameter-efficient code-switching speech recognition via soft prompt tuning. *arXiv preprint arXiv:2506.21576*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

8 Appendix

8.1 Strategy Selection and Cross-Track Evaluation

We initially observed consistent improvements in the Character Error Rate (CER), which suggested that joint training would not cause conflicts across different objectives. Motivated by this trend, we applied the same strategy to Track 2 and evaluated Word Error Rate (WER). In practice, however, performance on the pinyin-based task degraded, indicating that *speech-to-Chinese-character* and *speech-to-pinyin* are inherently different and should not necessarily share identical optimization recipes.

During early screening, SpecAugment was found to hurt WER and was therefore excluded from the final Track 2 configuration. As summarized in Table 11, both dialect special tokens and TTS (Half-Dict.) improved or maintained character-level recognition but further reduced WER relative to the baseline. Given competition timelines and limited compute, we could not conduct deeper; we leave these analyses to future work.

Table 11: Overview of strategies in our FSR-2025 implementation. "–" indicates the setting was excluded from Track 2 after early negative results.

Strategy	CER	WER
Baseline	4.33%	9.31%
Baseline+SpecAug	3.13%	–
Baseline+Special token	3.48%	12.82%
Baseline+TTS (Half Dict.)	4.18%	13.76%

8.2 Special tokens with additional tones

This additional part is for we have rich Hailu (304.123 hrs) and Sixian (312.369 hrs) dialect data compared to Dapu (34 hrs) and Zhaoan (34 hrs), we conducted an additional experiment with special tokens. We implemented two configurations: a baseline without dialect information and a hard-prompt approach that prepends dialect-specific tokens (e.g., `<|dialect_sixian|>`, `<|dialect_hailu|>`) to the input sequence. The intention was to leverage these high-resource dialects to improve the model's generalization to low-resource dialects and reduce overfitting. Under the CER (track 1) setting, the baseline achieved 5.00% while the hard-prompt system obtained 5.54%. Although explicit dialect prompts did not improve performance at this stage, these results provide insights for future approaches such as weighted dialect embeddings or automatic dialect inference.