A Channel-Aware Anomaly-Guided Data Augmentation Framework for the FSR-2025 Hakka Speech Recognition Challenge

Siang-Ting Lin, Arthur Hao, Chiun-Yu Hua, Kuan-Tang Huang, and Berlin Chen
National Taiwan Normal University, Taiwan
{61347114s, 41247050s, 614k0009c, 61347002s, berlin}@ntnu.edu.tw

Abstract

The Formosa Speech Recognition Challenge 2025 (FSR-2025) focuses on Taiwanese Hakka, a low-resource language with limited data diversity and channel coverage. To address this challenge, we propose a channel-aware, data-centric framework that leverages multilingual foundation models to mitigate mismatches between field recordings and training data. Our method integrates unsupervised anomaly detection and channel-conditioned augmentation to enhance data representativeness before ASR fine-tuning, aiming to explore the potential for improving robustness in low-resource Hakka speech recognition.

Keywords: Hakka Speech Recognition, Low-Resource Language, Domain Adaptation, Anomaly Detection, Data Augmentation

1 Introduction

Hakka remains a low-resource language for Automatic Speech Recognition (ASR). The challenge goes beyond limited overall data availability: it is particularly difficult to collect speech that adequately covers diverse real-world conditions, such as background noise, channel variability, and device or room effects. As a result, existing systems trained on insufficiently diverse data often lack robustness to these factors, which severely undermines practical deployment (Lu et al., 2023; Yang et al., 2023; Chen et al., 2023).

To address this gap, we adopt a data-centric pipeline that leverages multilingual resources while explicitly targeting the mismatch between field recordings and training data. Concretely, we first perform channel-aware data preprocessing and augmentation, and fine-tune Whisper (Radford et al., 2022) on the curated data.

An overview of the proposed data-centric pipeline is illustrated in Figure 1. The framework comprises three main modules corresponding to the system workflow: (1) **Target Data Selection**, where an anomaly detector based on Deep SVDD (Ruff et al., 2018) scores the test set to identify anomalous samples; (2) **Simulation Data Generation**, which employs CADA-GAN (Wang et al., 2025) to synthesize channel-aware augmented data; and (3) **ASR Fine-Tuning**, where the augmented and original training sets are jointly used to fine-tune the Whisper-based model. This three-stage pipeline unifies anomaly detection, simulation, and fine-tuning in a data-centric manner to address the channel mismatch problem in low-resource Hakka ASR.

Our design is pragmatic for the Hakka-in-the-wild setting: distribution shifts are often dominated by channel and environmental factors, e.g. device, room, reverberation, intermittent noise, which are only weakly captured by content- or speaker-centric supervision. We therefore separate two roles. First, a *channel-aware anomaly detector* operates on utterance embeddings to surface target-domain risks *without labels* and to prioritize channel conditions that the original training set undercovers. Second, a *channel-aware augmentation* stage consumes these rankings/statistics to expose the model to those under-represented conditions before fine-tuning.

Methodologically, our detector reuses an MFA-Conformer (Zhang et al., 2022) backbone to produce utterance-level embeddings, with channel supervision following prior channel-aware work. Per channel group, we adopt a lightweight two-layer Multi-Layer Per-

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 499–503

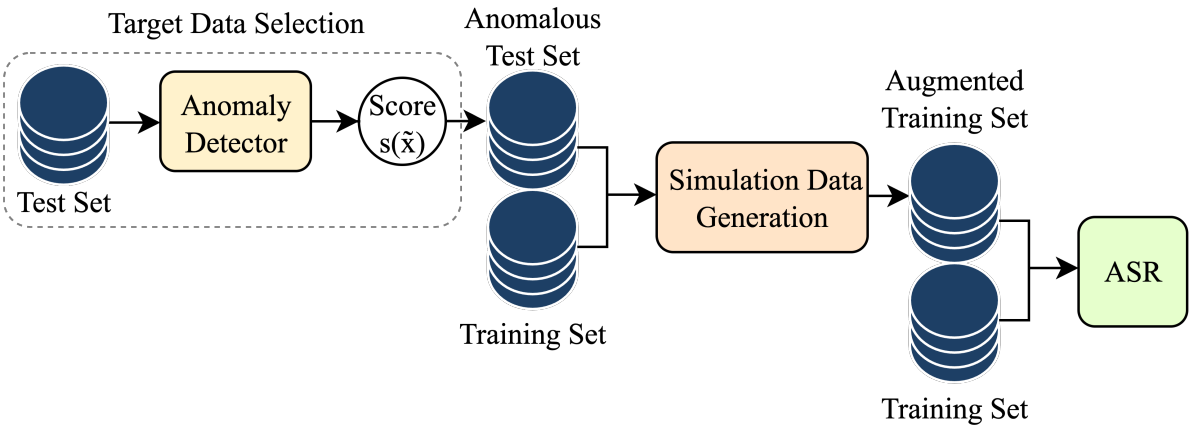

Figure 1: Overview of our data-centric pipeline. After training the anomaly detector with the training set using Deep SVDD, the detector scores the test set to filter anomalous samples (see Sec. 3.1 Target Data Selection). The selected data are then used to drive channel-aware simulation and data augmentation (see Sec. 3.2 Simulation Data Generation). Finally, the augmented and original training data are combined to fine-tune the Whisper-based ASR model (see Sec. 4.2 Model Configuration).

ceptron (MLP) with a **Soft-boundary deep SVDD** objective to score outliers; thresholds are derived from training-score quantiles and lightly calibrated on the target domain without retraining. This keeps the pipeline simple, label-free on the target side, and aligned with downstream augmentation and Whisper-based fine-tuning.

Contributions. (1) A channel-aware, data-centric pipeline for low-resource Hakka ASR that couples unsupervised detection with targeted augmentation prior to Whisper-Large fine-tuning. (2) A per-channel anomaly detector (MFA-Conformer embeddings + **Soft-boundary deep SVDD**) with bounded, no-retraining calibration to steer flag rates toward a target band. (3) An end-to-end recipe that prioritizes under-covered channel conditions and demonstrates improved robustness, evaluated with character error rate (CER) in realistic Hakka settings.

2 Background

2.1 Anomaly Detection

Anomaly detection identifies samples that deviate from the prevailing data distribution and is widely used in fraud, security, and industrial monitoring (Chandola et al., 2009; Schölkopf et al., 1999). In our low-resource Hakka automatic speech recognition (ASR) task, its role is to *surface target-domain risks without labels* and *prioritize channel conditions* that the original training set under-covers. Concretely, we use it to: (i) group utterances by channel, (ii) score and flag outliers per channel, and (iii) hand off ranked items/statistics to the downstream channel-aware augmentation stage (Sec. 3.1).

Why channel supervision (and how it relates to this task). Utterances are encoded by an MFA-Conformer backbone. In line with channel-aware work such as CADA-GAN (Wang et al., 2025), we train the encoder with *channel supervision* and at deployment reuse the learned channel discriminator to assign each utterance to a channel group. We adopt channel supervision as a pragmatic match to anticipated sources of shift in field Hakka recordings: prior channel-aware studies indicate it can emphasize channel factors and partially *disentangle* them from speaker or linguistic content. We do not claim general superiority over speaker-, phonetic-, or noise-type supervision; rather, this choice aligns with the channel-conditional analysis and augmentation used in our pipeline.

2.1.1 Deep SVDD

Deep SVDD (Ruff et al., 2018) learns an end-to-end hypersphere in a task-specific feature space so that *normal* data lie inside while violations indicate anomalies. We adopt the soft-bound variant with a lightweight two-layer MLP.

Score. Let $\mathbf{x}$ be the encoder embedding for a sample assigned to group g. For the detector we standardize $\mathbf{x}$ and apply PCA to 128 dimensions: $\tilde{\mathbf{x}} = \mathrm{PCA}_{128}(\mathrm{Standardize}(\mathbf{x}))$. With a two-layer MLP f_θ (hidden 128, output 64) and group center $\mathbf{c}_g$, the anomaly score is

$$s(\tilde{\mathbf{x}}) = \left\| f_\theta(\tilde{\mathbf{x}}) - \mathbf{c}_g \right\|_2^2. \tag{1}$$

Soft-boundary Deep SVDD loss. Within group g we optimize

$$\mathcal{L}_g(\theta, R_g) = R_g^2 + \frac{1}{\nu_g} E\big[\max\big(0,\, s(\tilde{\mathbf{x}}) - R_g^2\big)\big] + \frac{\lambda}{2} \sum_l \|W^l\|_F^2, \tag{2}$$

where R_g is the radius, $\nu_g \in (0, 1)$ trades tightness vs. violations, and $\{W^l\}$ are layer weights (implemented via AdamW weight decay). After each epoch, R_g^2 is set to the $(1-\nu_g)$ quantile of *training* scores; the decision threshold is $\tau_g = R_g^2$. A test utterance is anomalous in group g iff $s(\tilde{\mathbf{x}}) > \tau_g$. For cross-group prioritization we use a rarity indicator computed against each group's training-score distribution (no label usage), which feeds the channel-aware augmentation in Sec. 3.1.

3 Method

3.1 Target Data Selection

Scope. The unlabeled target domain (final test audio) is used solely for *unsupervised* scoring, per-channel thresholding, and ranking; no labels are accessed and no model parameters are updated with target data.

Pipeline. (1) *Channel grouping.* Reuse the channel-supervised encoder (Wang et al., 2025) to assign each utterance to a group g (grouping uses original encoder embeddings). (2) *Detector features.* For Deep SVDD we standardize embeddings and apply PCA to 128 dimensions (PCA=128). (3) *Detector model.* Within each group, train a two-layer MLP f_θ (hidden 128, output 64). Let $\mathbf{x}$ be the encoder embedding and $\tilde{\mathbf{x}} = \mathrm{PCA}_{128}(\mathrm{Standardize}(\mathbf{x}))$. Define $\mathbf{z} = f_\theta(\tilde{\mathbf{x}})$ and the group center $\mathbf{c}_g$ (mean of training $\mathbf{z}$).

Thresholding and calibration (no retraining). After each epoch we set R_g^2 to the $(1-\nu_g)$ quantile of *training* scores in group g; the decision threshold is $\tau_g = R_g^2$. At test time we keep f_θ and $\mathbf{c}_g$ fixed and adjust only ν_g (hence τ_g) within bounds (e.g., $[0.01, 0.10]$) to steer the group's flag rate toward a target band ($\sim$5%). This *auto-calibration* accommodates train—test mismatch without updating model parameters.

Decision and ranking. A test utterance in group g is anomalous iff $s(\tilde{\mathbf{x}}) > \tau_g$. For cross-group prioritization we use a stable ordering: (1) anomalous first $\Rightarrow$ (2) smaller rarity indicator (tail probability, computed against the group's training scores) $\Rightarrow$ (3) larger s. The resulting per-channel flag rates and thresholds $\{(\nu_g, \tau_g)\}$ guide channel-aware augmentation to expose the ASR model to characteristics under-covered by the original training set.

3.2 Simulation data generation

We adopt CADA-GAN, a Channel-Aware Domain-Adaptive Generative Adversarial Network proposed by Wang et al. (Wang et al., 2025). The model is specifically designed to address channel mismatch in ASR by generating augmented speech data conditioned on channel characteristics. In our framework, CADA-GAN is used to synthesize additional training utterances, enriching the channel diversity of the training set.

Channel encoder: The data identified by the Deep SVDD method are used as the target source and processed by the MFA conformer to extract channel-aware representations. These representations are subsequently employed in the generator via Feature-wise Linear Modulation (FiLM)(Perez et al., 2018), where they are transformed into weights and biases to modulate the data generation process.

Generator and Discriminator: During this process, the generator integrates the encoded source data with FiLM to synthesize simulated data, while the discriminator enforces consistency between the generated data and both the intrinsic characteristics of the original source data and the embeddings of the target data.

4 Experimental Setup

4.1 Dataset

	Sentences	Hours
Train	21,879	52
Eval	5,470	8
Test(warm-up)	4,404	10
Total	31,753	70

Table 1: Dataset statistics of the FSR-2025-Hakka corpus.

We use the FSR-2025-Hakka corpus as our primary dataset. The train set contains a total of 60 hours of speech, evenly divided between two dialects: Dapu and Zhao'an (30 hours each). From this corpus, 20% of the data is randomly selected as the Eval set, while the remaining 80% is used as the Train set. The Test set consists of 10 hours of speech released for the warm-up phase, which is employed to evaluate inference performance after fine-tuning. The dataset composition is summarized in Table 1.

4.2 Model Configuration

We employed OpenAI's Whisper-Large model as our base architecture. The model configuration consisted of the following components:

Pre-trained Model We utilized the "openai/whisper-large" pre-trained model, which provides robust multilingual speech recognition capabilities. To optimize training efficiency and prevent catastrophic forgetting of learned features, we applied encoder freezing strategy, allowing only the decoder parameters to be updated during fine-tuning.

Training Strategy Our training approach employed the Seq2SeqTrainer framework. We set the batch size to 8 and accumulated gradients over 8 steps, yielding an effective batch size of 64. The model was optimized with a learning rate of 1×10^{-4}, scheduled linearly with 1,000 warmup steps. Training proceeded for 20 epochs with early stopping based on validation performance. To mitigate overfitting, we applied a weight decay of 0.01, while gradient clipping was enforced with a maximum norm of 1.0. For efficiency, we enabled mixed-precision training (FP16) and activated gradient checkpointing to reduce memory consumption.

The training dataset comprised Hakka speech data from Dapu and Zhao'an dialect variants with total 60hr data

4.3 Evaluation Metrics

Following established practices in automatic speech recognition evaluation , we employed Character Error Rate (CER) as our primary evaluation metric, which is particularly suitable for Chinese languages including Hakka.

CER The CER measures recognition accuracy at the character level and is computed as:

$$CER = \frac{S + D + I}{N} \times 100\%, \tag{3}$$

where S represents character substitutions, D represents deletions, I represents insertions, and N is the total number of characters in the reference transcript.

5 Results

Table 2 shows the CER of different training settings. Without preprocessing, the baseline system achieved a CER of 16.07%. Through systematic experimentation with different augmentation ratios, we identified 13% augmented data as the optimal configuration, yielding a CER of 15.13% when the augmented samples were generated to simulate the test set channel characteristics.

These results demonstrate that our proposed augmentation method achieves substantial performance improvement, with the optimal 13% augmentation ratio providing a 0.94 percentage point reduction in CER compared to the baseline, confirming the effectiveness of our channel simulation approach.

Method	CER
w/o Preprocessing	16.07%
Add 13% augmented data	15.13%

Table 2: CER Compare Table.

6 Conclusion and Future Work

This work presents a channel-aware, data-centric pipeline that combines unsupervised anomaly detection with targeted augmentation to address channel mismatch in low-resource Hakka ASR. By incorporating 13%

channel-simulated data, our approach reduces CER to 15.13%, achieving a 0.94-point improvement over the baseline. Our results demonstrate enhanced model robustness in realistic, noisy environments, validating the effectiveness of channel-focused augmentation.

For future work, we plan to extend the preprocessing pipeline to include semantic- and noise-specific analysis, enabling more fine-grained supervision of both linguistic and acoustic variations. In particular, long-duration noise segments, which may currently be misclassified as channel shifts, will be addressed through targeted refinement. Moreover, we will further investigate the role of FiLM modulation, as excessive influence from the generator may overpower the modulation process and reduce the contribution of source data, potentially limiting the effectiveness of synthetic augmentation.

7 Limitation

Data scale and coverage. Hakka remains low-resource; the amount and channel diversity of transcribed training audio constrain fine-tuning effectiveness. Coverage gaps (devices/rooms/reverberation patterns) limit how well Whisper-Large can adapt, even with targeted augmentation.

References

Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58.

Li-Wei Chen, Kai-Chen Cheng, and Hung-Shin 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)*, pages 386–389, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Hao-Chien Lu, Chung-Chun Wang, Jhen-Ke Lin, and Tien-Hong 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)*, pages 397–402, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

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

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. 2018. Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4393–4402. PMLR.

Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. 1999. Support vector method for novelty detection. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press.

Chien-Chun Wang, Li-Wei Chen, Cheng-Kang Chou, Hung-Shin Lee, Berlin Chen, and Hsin-Min Wang. 2025. Channel-aware domain-adaptive generative adversarial network for robust speech recognition.

Tzu-Ting Yang, Hsin-Wei Wang, Meng-Ting Tsai, and Berlin Chen. 2023. The NTNU super monster team (SPMT) system for the Formosa speech recognition challenge 2023 - Hakka ASR. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 414–422, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Yang Zhang, Zhiqiang Lv, Haibin Wu, Shanshan Zhang, Pengfei Hu, Zhiyong Wu, Hung yi Lee, and Helen Meng. 2022. Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification.