MetaCORA: A Meta-Learned Curriculum for Adversarial and Contrastive Robustness in Speech Recognition

Yuqian Dai, Chun Fai Chan, Ying Ki Wong, Tsz Ho Pun
Logistics and Supply Chain MultiTech R&D Centre
Hong Kong
{yuqian.dai, cfchan, skwong, thpun}@lscm.hk

Abstract

Pre-trained speech models like Whisper demonstrate impressive performance under ideal conditions but still face robustness challenges in low-resource language scenarios. We introduce Meta Curriculum Optimization for Robust ASR (MetaCORA), a novel meta-curriculum adaptive framework that improves speech recognition for low-resource Hong Kong Cantonese by integrating adversarial training with feature contrastive learning. Our approach dynamically adjusts three critical hyperparameters: adversarial perturbation magnitude, optimization step size, and contrastive learning temperature, allowing the model to adapt to varying training difficulties throughout the learning process. Unlike traditional meta-learning approaches, our framework does not rely on end-to-end differentiability but instead utilizes validation performance as a signal to guide hyperparameter adjustments. Experimental results demonstrate that our approach achieves lower WER than standard Whisper fine-tuning, commercial speech recognition systems, and LLM-based methods. Ablation studies confirm the necessity of each component, as removing any single element leads to a measurable drop in performance. The model also exhibits robustness under noisy conditions, achieving consistently lower WER than baseline systems. Further analysis shows that MetaCORA effectively compresses the distance between adversarial feature representations while maintaining well-separated class boundaries in the embedding space, providing a mechanistic explanation for its improvement.

Keywords: Speech Recognition, Meta Learning, Low-resource Language

1. Introduction

Deep neural architectures have transformed automatic speech recognition (ASR), with transformer-based models achieving impressive accuracy for speech-to-text conversion in controlled settings. However, a critical limitation has become apparent: even minor acoustic perturbations can induce significant transcription errors, particularly in low-resource settings such as Cantonese speech recognition (Cao et al., 2024; An et al., 2024; Cumbal et al., 2024; Yang et al., 2024). This challenge raises fundamental concerns about model robustness and may undermine reliability in high-stakes applications, including healthcare and public safety (Adedeji et al., 2024; Li et al., 2024).

While adversarial training has emerged as a promising defense mechanism, current approaches face two significant limitations. First, they typically implement fixed attack parameters throughout the training process (Jin et al., 2024; Teixeira et al., 2024), creating a fundamental dilemma: overly aggressive attacks in early stages overwhelm the model's developing robustness, while later stages lack sufficient challenge to prepare for novel threats. Second, manually crafted curricula (Kim et al., 2024; Liu et al., 2024) often struggle to accommodate the model's dynamic training. Challenges that are significant during the early stages of training may become inconsequential as learning progresses. Beyond these technical issues lies a deeper problem: adversarial training often struggles to bridge the gap between synthetic perturbations and real-world sound variations. Even if a model is resistant to digital attacks, its robustness often degrades when exposed to unavoidable acoustic disturbances in real-world environments (Jung et al., 2024; Shi and Kawahara, 2024). This challenge becomes particularly acute for low-resource languages, where limited training data substantially exacerbates the difficulty of developing models that generalize effectively across diverse speech conditions.

Drawing from human learning principles, which prioritize understanding foundational concepts before tackling more advanced issues, we propose a new framework that adapts the adversarial and contrastive training strategy of Whisper model (Radford et al., 2023) specifically for low-resource Hong Kong Cantonese ASR. Our approach, Meta Curriculum Optimization for Robust ASR (MetaCORA), introduces three key innovations: (1) a gradient-based adversarial training mechanism that generates challenging perturbations tailored to Cantonese phonetic features, extending Whisper's robustness to address dialect-specific challenges; (2) a noise-invariant representation module that enforces consistency between clean and adversarially perturbed inputs through contrastive learning, bridging the gap between synthetic perturbations and real-world acoustic variations; and (3) an adaptive meta-controller that automatically tunes both

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9999–10008
11–16 May 2026.

adversarial and contrastive parameters, organizing training examples according to difficulty while adjusting adversarial intensity based on the model's evolving performance. The main contributions of this paper are as follows:

- We propose MetaCORA, a meta-learning framework that dynamically adjusts the weighting of adversarial and contrastive training objectives based on validation performance. Experiments on low-resource Hong Kong Cantonese ASR show that models trained with MetaCORA outperform conventional fine-tuning strategies, commercial ASR systems, and LLM-based speech recognizers. The best-performing configuration achieves a word error rate (WER) at least 10% lower than that of existing approaches. Ablation studies confirm the necessity of each component: removing meta-learning, adversarial training, or contrastive learning leads to performance degradation, with the absence of meta-learning resulting in the most significant increase in WER. These results underscore the integral role of adaptive training orchestration in enhancing model robustness and generalization under resource-constrained conditions.
- Experiments under various white noise conditions, with signal-to-noise ratios (SNRs) of -5 dB, 0 dB, +5 dB, and +10 dB, show that the proposed approach consistently outperforms conventionally fine-tuned Whisper models. In most cases, the improvement corresponds to a reduction in WER of at least 2 percentage points across all noise levels, demonstrating robustness to Cantonese speech interference. Furthermore, analysis of the learned feature space reveals that MetaCORA brings clean and adversarially perturbed samples closer together while maintaining well-separated class boundaries. This indicates that the model enhances invariance to noise and preserves discriminative capability, effectively balancing robustness and semantic fidelity in speech recognition.

2. Related Work

Pre-trained models such as Whisper have improved ASR performance, but their robustness in noisy and low-resource settings is still limited. Adversarial training has emerged as an effective strategy for improving model resilience across multiple domains. Carlini and Wagner (2018) demonstrate that ASR models are highly susceptible to carefully crafted adversarial examples, and subsequent work has shown that adversarial training can mitigate such vulnerabilities (Park and Kim, 2024; Fang et al., 2024). However, most current methods rely on fixed hyperparameters for perturbation generation, which may fail to adapt to the evolving dynamics of model learning across training stages. Contrastive learning has also played a central role in advancing self-supervised representation learning in speech. Frameworks such as wav2vec (Baevski et al., 2020) and HuBERT (Hsu et al., 2021) leverage contrastive objectives to learn rich and robust audio representations. Nevertheless, these approaches typically treat adversarial examples as external threats and do not integrate them into the contrastive learning framework (Wang et al., 2024; Chen et al., 2024). Consequently, the combination of contrastive learning and adversarial training for improving robustness in low-resource ASR remains underexplored.

Hyperparameter optimization may emerge as a key challenge in training deep models. Meta-learning frameworks offer a promising solution by automating this process, reducing the need for manual tuning while enhancing model performance. Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017) and Reptile (Nichol et al., 2018) focus on learning optimal initializations that enable fast adaptation to new tasks. In adversarial training, recent work has explored adaptive adjustment of perturbation magnitudes to improve robustness (Zhou et al., 2024; Liu et al., 2025). However, most existing methods rely on end-to-end differentiable frameworks, which require second-order gradients and are computationally intensive, making them impractical for large-scale ASR systems such as Whisper. In contrast, approaches that adjust training hyperparameters based on validation performance, without requiring full differentiability, offer a more scalable alternative. These methods decouple optimization from model architecture, enabling efficient adaptation even in resource-constrained settings. Yet, such strategies remain underexplored in the context of automatic speech recognition, particularly for low-resource languages.

3. Methodology

Our MetaCORA framework enhances the robustness of pre-trained ASR models. Figure 1 illustrates the overall architecture of our approach. First, a multi-step adversarial learning objective (Sec 3.2) generates challenging, feature-level acoustic perturbations. Second, a contrastive learning module (Sec 3.3) enforces consistency between the representations of clean and adversarially perturbed inputs, promoting noise-invariant learning. Finally, to prevent fixed adversarial parameters from overwhelming or under-challenging the model, a meta-curriculum optimizer (Sec 3.4) continuously monitors validation signals to dynamically adjust the perturbation magnitude, step size, and contrastive

temperature.

3.1. Data Preprocessing

Given a set of audio directories and an associated csv file, our data preprocessing pipeline automatically aligns audio files with their corresponding transcriptions. Let x_{raw} denote a raw audio file and t its transcription. First, the audio is resampled to 16 kHz and processed by a pre-trained feature extractor to obtain a feature representation $\mathbf{x} = f(x_{\text{raw}}) \in \mathbb{R}^{T \times D}$, where T is the number of time steps and D is the feature dimension. The corresponding transcription is tokenized into a sequence of identifiers $y = \{y_1, y_2, \ldots, y_L\}$, where L is the length of the sequence.

3.2. Multi-Step Adversarial Learning

Given input features $\mathbf{x} \in \mathbb{R}^{T \times D}$ and ground truth labels y, we formulate the adversarial example generation as a constrained optimization problem:

$$\delta^* = \arg \max_{\|\delta\|_\infty \leq \epsilon} \mathcal{L}(f_\theta(\mathbf{x} + \delta), y) \tag{1}$$

where δ represents the adversarial perturbation, and the constraint $\|\delta\|_\infty \leq \epsilon$ ensures that the perturbation remains within a bounded ball around $\mathbf{x}$. In our dynamic framework, the constant perturbation bound ϵ is subsequently replaced by a time-varying bound ϵ_t (detailed in Sec 3.4).

Our approach introduces three key innovations to improve upon standard projected gradient descent (PGD). Instead of using a fixed step update, we incorporate an iteration-dependent weighting scheme that gradually increases the contribution of the gradient across multiple iterations. Formally, for iteration k (with $k \in \{0, 1, 2\}$), we define the weight w_k as shown in Equation 2, ensuring that later iterations contribute more significantly to the update process.

$$w_k = \frac{k+1}{6} \tag{2}$$

To further enhance the adversarial training process, we dynamically regulate critical hyperparameters during training. In our setting, both the perturbation bound ϵ_t and the step size α_t vary throughout training based on the model's current state:

$$\epsilon_t = f_\epsilon(t, \text{training metrics}) \tag{3}$$

$$\alpha_t = f_\alpha(t, \text{training metrics}) \tag{4}$$

where ϵ_t denotes the allowable perturbation magnitude at time step t, and α_t represents the step size. These parameters are adjusted based on training metrics such as loss values, gradient norms, and accuracy. The specific mechanism for this dynamic adjustment will be detailed in Section 3.4. By dynamically adjusting ϵ_t and α_t, the model can balance exploration (larger perturbations) and exploitation (smaller, more precise perturbations) during training.

At each iteration, the adversarial perturbation is updated according to the following Equation 5:

$$\delta_{k+1} = \text{clip}_{[-\epsilon_t, \epsilon_t]} \left(\delta_k + \alpha_t \cdot \text{sign}(g_k)\right) \tag{5}$$

where g_k is the gradient of the loss with respect to δ computed at the current perturbation δ_k, and $sign$ returns the elementwise sign of its argument, $\text{clip}_{[-\epsilon_t, \epsilon_t]}$ ensures that each element of the perturbation remains within the interval.

3.3. Audio Feature Contrastive Learning

To enhance the model's generalization capability across different audio variants, we introduce a method based on contrastive learning. This method enables the model to learn semantically invariant representations of audio content. Given the original input feature $\mathbf{x}$ and its corresponding adversarial sample $\mathbf{x}^{adv} = \mathbf{x} + \delta$, our objective is to make the model learn similar representations for both inputs while maintaining distinct representations from other samples in the batch. First, we extract feature representations from the encoder:

$$\mathbf{z} = g(\mathbf{x}) = \frac{1}{T} \sum_{t=1}^{T} \mathbf{h}_t \tag{6}$$

where $g(\mathbf{x})$ denotes the function for extracting features from the Whisper encoder, $h_t \in \mathbb{R}^d$ represents the hidden state at time step t, and T is the sequence length. We perform average pooling over the time dimension of the encoder's last hidden states to obtain the global representation $\mathbf{z} \in \mathbb{R}^d$.

Given a mini-batch containing N distinct original audio samples, we generate their corresponding N adversarial samples. We then concatenate the feature representations of these clean and adversarial samples to construct a $2N \times 2N$ cosine similarity matrix:

$$S_{i,j} = \frac{\mathbf{z}_i^T \mathbf{z}_j}{||\mathbf{z}_i|| \cdot ||\mathbf{z}_j||} \cdot \frac{1}{\tau} \tag{7}$$

where τ is the temperature parameter, dynamically adjusted by the meta-optimizer, to control the smoothness of the similarity distribution. A lower temperature value makes the similarity distribution steeper, emphasizing high-similarity pairs, while a higher temperature value makes the distribution smoother, reducing the relative difference in similarities.

We define positive pairs as the original samples and their corresponding adversarial samples of the same audio content, and negative pairs as samples

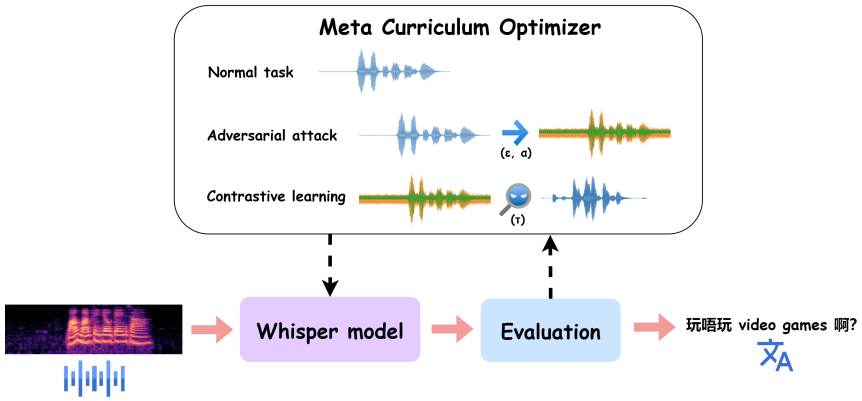

Figure 1: Overview of the proposed MetaCORA framework. The system dynamically adjusts adversarial training parameters (ϵ, α) and contrastive learning temperature (τ) based on model state. Clean examples are used for standard ASR training, while adversarial examples enhance robustness through adversarial training. Additionally, contrastive learning aligns representations between clean and adversarial pairs to further improve model robustness.

from different audio contents. The contrastive loss is computed using the InfoNCE form:

$$\mathcal{L}_{\text{con}} = -\frac{1}{2N}\sum_{i=1}^{2N}\sum_{j=1}^{2N}\frac{M_{i,j}\cdot\log\left(\frac{\exp(S_{i,j})}{\sum_{k=1}^{2N}\exp(S_{i,k})}\right)}{\sum_{j=1}^{2N}M_{i,j}} \tag{8}$$

where $M_{i,j}$ is a binary mask matrix that is 1 when i and j represent different versions of the same audio content (i.e., positive pairs), and 0 otherwise. This loss encourages the model to minimize the distance between clean and adversarial features (positive pairs) while maximizing their distance from unrelated features in the batch (negative pairs). By incorporating this objective, the model learns more robust feature representations that are less sensitive to adversarial perturbations, thereby improving its generalization to noisy or perturbed inputs.

3.4. Meta Curriculum Optimizer

During training, we employ a Meta Curriculum Optimizer to automatically adjust the hyperparameters for adversarial training (perturbation bound ϵ and step size α) as well as the temperature parameter (τ) used in contrastive learning. Specifically, we construct a state feature vector as shown below.

$$\mathbf{s} = \left[\, s_{\text{step}},\ s_{\text{loss}},\ s_{\text{grad}},\ s_{\text{train}},\ s_{\text{val}} \,\right] \in \mathbb{R}^5 \tag{9}$$

where s_{step} represents the normalized training progress (ratio of the current training step to the total number of steps). The terms s_{loss} and s_{val} are normalized training and validation losses, respectively, each scaled within a recent sliding window of loss values. Moreover, s_{grad} denotes the normalized gradient norm, capturing the magnitude of recent gradients, while s_{train} is the average training accuracy computed over several recent iterations. These features serve as proxy signals, indirectly reflecting the effectiveness of the current hyperparameter settings and are then encoded by a two-layer feed-forward network:

$$\mathbf{h} = \text{LayerNorm}(\text{ReLU}(W_1\mathbf{s} + b_1)) \in \mathbb{R}^d \tag{10}$$

The parameter decoder then outputs three values (each normalized to (0,1)):

$$\mathbf{a} = \text{Sigmoid}(W_3(\text{ReLU}(W_2\mathbf{h} + b_2)) + b_3) \in \mathbb{R}^3 \tag{11}$$

Finally, these outputs are linearly scaled to their respective hyperparameter ranges:

$$\epsilon = \epsilon_{\min} + (\epsilon_{\max} - \epsilon_{\min}) \cdot a_1 \tag{12}$$

$$\alpha = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min}) \cdot a_2 \tag{13}$$

$$\tau = \tau_{\min} + (\tau_{\max} - \tau_{\min}) \cdot a_3 \tag{14}$$

where $\epsilon \in [0.03, 0.08]$, $\alpha \in [0.003, 0.01]$, and $\tau \in [0.05, 0.5]$. These specific search boundaries are empirically constrained to maintain training stability and acoustic fidelity. Preliminary evaluations indicated that an unbounded ϵ (e.g., $\epsilon > 0.08$) induces severe semantic distortion in Cantonese tonal features, whereas $\epsilon < 0.03$ yields insufficient adversarial regularization. Provided the predefined ranges encapsulate these functional limits, the policy network demonstrates robustness to minor boundary variations. Furthermore, regarding the architecture of the meta-optimizer, mapping a 5-dimensional state vector $\mathbf{s}$ to a 3-dimensional action space requires minimal representational capacity. We intentionally employ a lightweight two-layer feed-forward network, as early explorations confirmed that deeper architectures provide negligible WER improvements while increasing the susceptibility to overfitting on the validation batch.

At fixed intervals (every 100 training steps), the meta-optimizer adjusts its parameters θ (weights

of the feature encoder and decoder) to minimize the validation loss:

$$\min_{\theta} \mathcal{L}_{\text{meta}} = \mathbb{E}\mathcal{L}_{\text{val}} \tag{15}$$

$$\theta \leftarrow \theta - \eta_{\text{meta}} \nabla_{\theta} \mathcal{L}_{\text{val}} \tag{16}$$

where $\mathbb{E}[\cdot]$ denotes averaging over a sliding window of the 100 most recent validation losses, η_{meta} is the meta learning rate.

Our approach has key differences from traditional meta-learning. While the gradient update form in Equation 17 resembles standard gradient descent, our implementation fundamentally differs from conventional end-to-end differentiable meta-learning frameworks. Specifically:

$$\nabla_{\theta} \mathcal{L}_{\text{meta}} \neq \frac{\partial \mathcal{L}_{\text{meta}}}{\partial w} \cdot \frac{\partial w}{\partial \theta} \tag{17}$$

where w represents the model weights. The gradient path $\frac{\partial w}{\partial \theta}$ is omitted in our implementation, as hyperparameter adjustments influence w indirectly through the training trajectory rather than via direct computational graph connections. This design reflects our fundamental assumption that validation loss dynamics contain statistically significant signals about hyperparameter effectiveness. For instance: Elevated $\mathcal{L}_{\text{val}}$ may indicate excessive ϵ values causing semantic distortion in adversarial examples. By deliberately eschewing end-to-end differentiability, our method leverages the empirical correlation between validation metrics and hyperparameter configurations to establish an implicit curriculum. In this framework, ϵ_t, α_t, and τ_t adjust based on the model's capabilities, increasing the challenge as performance improves and decreasing the difficulty when validation metrics decline. This approach mirrors how an effective teacher tailors instruction according to student progress.

3.5. Loss Integration

Our overall training objective as shown in Equation 18 combines multiple loss functions to balance accuracy and robustness:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{task} + \lambda_2 \mathcal{L}_{adv} + \lambda_3 \mathcal{L}_{con} \tag{18}$$

where $\mathcal{L}_{task}$ is the standard cross-entropy loss for the original audio input, ensuring high accuracy on clean samples. $\mathcal{L}_{adv}$ applies the same cross-entropy loss but on adversarially perturbed inputs $(\mathbf{x}+\delta^*)$, where δ^* is the optimized perturbation from our multi-step adversarial approach. Finally, $\mathcal{L}_{con}$ is the contrastive loss that encourages consistent feature representations between clean and adversarial samples. The weighting parameters $\lambda_1 = 0.8$, $\lambda_2 = 0.1$ and $\lambda_3 = 0.1$ balance the different training objectives. These components address the limitations of using any single objective in isolation, resulting in a model that maintains high accuracy while being substantially more robust to various forms of input distortion encountered in real-world speech recognition scenarios.

4. Experiments

4.1. Datasets and Models

Cantonese remains significantly underrepresented in ASR research, a situation largely caused by the scarcity of standardized speech corpora. Existing resources for its primary variants, Guangzhou and Hong Kong Cantonese, are fragmented across various academic and informal collections. This data scarcity is particularly acute for Hong Kong Cantonese, whose unique lexicon and prevalent code-switching with English are poorly covered by existing datasets.

To address this gap, we construct a Cantonese speech corpus. Our work integrates established open-source datasets, such as Common Voice[1] and MDCC[2], with a newly compiled corpus of Hong Kong Cantonese. This new component is built from diverse, speech sources, mainly including Hong Kong news broadcasts, official speeches and meeting minutes from government agencies, and purpose-recorded speech samples[3]. The final consolidated dataset spans 452 hours and contains approximately 390,000 transcribed sentences, providing coverage of both major Cantonese variants.

To validate our proposed method, we benchmark its performance against the Whisper model family. We systematically evaluate models of varying scales, tiny, base, small, medium, and large-v2. This scaling analysis demonstrates the effectiveness of our approach across a broad spectrum of parameter budgets, ranging from 39 million to 1.55 billion. We reserve 10,279 audio samples (approximately 9 hours) as a test set. The remaining samples serve as training and validation sets for fine-tuning the Whisper models, with 95% used for training and 5% for validation.

All Whisper models are trained with AdamW optimizer (lr=5e-6, weight decay=0.01) and mixed precision training. Our meta-learning framework features a two-layer feed-forward network (hidden dim=64) and utilizes a separate Adam optimizer (lr=1e-4) for curriculum parameter adjustment. The meta-controller continuously monitors training dynamics through a sliding window mechanism to adaptively adjust adversarial training parameters

[1] https://commonvoice.mozilla.org/yue

[2] https://github.com/HLTCHKUST/cantonese-asr

[3] Our dataset can be found at https://huggingface.co/datasets/HKAllen/cantonese-chinese-parallel-audio

and contrastive learning temperature. All experiments are conducted using 7 NVIDIA A100 GPUs and 8 NVIDIA L40S GPUs.

4.2. Model Comparison

We compare models using our approach with those using traditional fine-tuning (without any additional strategies). The original vanilla Whisper models are excluded from the comparison since they consistently failed to output Cantonese transcription. To benchmark our approach against both industry-standard solutions and emerging LLM-based approaches, we incorporate two additional categories of ASR systems: commercial speech recognition engines (Aliyun[4], Google Cloud[5], Azure Speech Studio[6], Cantonese.ai[7]) and LLM-based ASR models (Qwen2-audio-7B[8], GLM-4-Voice-9B[9], GPT-4o Transcribe[10]).

As shown in Table 1, the WER scores of traditionally fine-tuned Whisper models gradually decrease with an increase in the number of parameters, ranging from 33.65 for Whisper-tiny to 16.46 for the best-performing Whisper-large-v2. In contrast, models trained with our proposed MetaCORA framework consistently exhibit superior performance with WER reductions ranging from 28.57 to 14.43 at the same parameter scale. This indicates that the MetaCORA framework has a significant advantage in improving the performance of Cantonese speech recognition. When benchmarked against state-of-the-art commercial ASR systems, MetaCORA-large-v2 demonstrates better performance, surpassing Aliyun, Google Cloud, and Azure Speech Studio. This performance differential of 11.8% relative to the best commercial system underscores the efficacy of our meta-learning contrastive approach for low-resource Cantonese speech recognition.

In addition, our comparison with recent LLM-based ASR models reveals an intriguing trade-off between efficiency and performance. Despite a substantial reduction in parameter count, MetaCORA-large-v2 achieves a 28.4% relative WER reduction compared to Qwen2-Audio and 22.5% compared to GLM-4-Voice, even slightly surpassing GPT-4o-Transcribe. This finding challenges the conventional assumption that a larger number of parameters necessarily leads to superior absolute performance in speech recognition tasks. Instead, it highlights the importance of domain-specific architectural design and targeted training objectives. In specialized tasks such as Cantonese speech recognition, parameter count is not the sole determinant of model performance; the model architecture, training methodology, and the use of domain-appropriate datasets are equally critical.

Model	WER (%)	Params
Traditional Fine-tuned		
Whisper-tiny	33.65	39M
Whisper-base	26.76	74M
Whisper-small	21.96	244M
Whisper-medium	19.42	769M
Whisper-large-v2	**16.46**	1550M
Proposed MetaCORA		
MetaCORA-tiny	28.57	39M
MetaCORA-base	20.21	74M
MetaCORA-small	19.99	244M
MetaCORA-medium	18.05	769M
MetaCORA-large-v2	**14.43**	1550M
Commercial ASR Systems		
Google Cloud	18.16	-
Azure Speech Studio	19.01	-
Cantonese.ai	18.04	-
Aliyun Speech	**16.37**	-
LLM-based Models		
Qwen2-Audio	20.17	7B
GLM-4-Voice	18.64	9B
GPT-4o-Transcribe	**14.52**	-

Table 1: WER comparison between different ASR models on Cantonese test set. Lower values indicate better performance. Best results in each category are bolded, and the overall best result is underlined.

4.3. Ablation Study

To isolate the contribution of each component in our proposed framework, we conduct an ablation study by systematically removing key modules from the full architecture. The removal of a module indicates that the model is trained without that specific functionality. When meta-learning is removed, the parameters for adversarial learning and contrastive learning are fixed based on default starting parameters.

As illustrated in Figure 2, the removal of the meta-learning (w/o Meta) module has the most substantial impact on model performance, resulting in a significant increase in the WER across all model variants. The performance degradation is particularly evident for the small, medium, and large-

[4] https://www.alibabacloud.com/

[5] https://cloud.google.com/speech-to-text

[6] https://azure.microsoft.com/en-us/products/ai-services/ai-speech

[7] https://cantonese.ai/

[8] https://huggingface.co/Qwen/Qwen2-Audio-7B

[9] https://github.com/THUDM/GLM-4-Voice

[10] https://platform.openai.com/docs/models/gpt-4o-transcribe

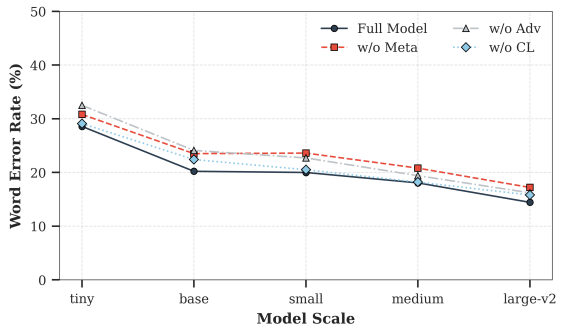

Figure 2: WER across different Whisper model scales under various ablation settings. The "Full Model" includes all components, while "w/o Adv", "w/o Meta", and "w/o CL" indicate removal of adversarial training, meta-learning, and contrastive learning modules, respectively.

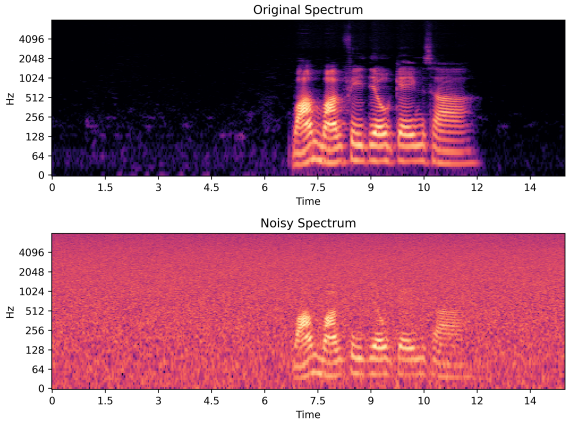

Figure 3: Comparison of mel spectrograms for original and noisy audio signals, with the noisy signal containing additive white noise at SNR = -5 dB.

v2 models, exceeding the effects observed from ablating the adversarial training (w/o Adv) or contrastive learning (w/o CL) modules. This finding underscores the effectiveness of meta-learning in dynamically adjusting model parameters based on learning progress, as compared to traditional manual heuristic parameter tuning, especially in the context of complex speech patterns.

Ablation experiments on adversarial training reveal a greater performance degradation in the tiny and base models. Small-scale models, constrained by their limited parameter capacity, may depend more heavily on the regularization provided by adversarial training to mitigate shifts in the acoustic feature distribution. In contrast, the removal of the contrastive learning module has a relatively lesser impact, indicating that contrastive learning primarily improves performance by optimizing the discriminative representation space of speech features. This optimization appears to be more beneficial for small-scale models, whereas larger models, which inherently possess strong feature disentanglement capabilities, see limited marginal gains from contrastive learning.

4.4. Noise Robustness Analysis

Since real-world noise is highly variable, using a fixed set of real-world recordings may bias evaluation toward specific environmental conditions, limiting result generalizability and complicating the assessment of a model's true robustness. To ensure a controlled and consistent evaluation, we instead add white noise at varying SNRs of –5, 0, 5, and 10 dB (Figure 3). These SNR levels represent different degrees of noise contamination in the audio signals. A lower SNR value indicates a higher level of noise relative to the speech signal, making the recognition task more challenging. Conversely, a higher SNR value indicates a lower level of noise, which is less challenging for the models. Despite this, all SNR levels can be considered noisy environments relative to a clean, noise-free signal.

Figure 4 demonstrates the performance of the MetaCORA approach compared to traditional fine-tuning across different Whisper model sizes in the presence of white noise, simulating a noisy environment. The results indicate that MetaCORA consistently outperforms traditional fine-tuning across all Whisper model scales, achieving at least a 3% absolute reduction in WER even at the highly challenging -5 dB SNR level. This finding supports the hypothesis that our proposed approach enhances the model's ability to maintain speech discriminability in degraded acoustic conditions.

We also observe that the benefits of adaptation diminish as the model capacity increases. Specifically, the WER improvement for the large-v2 model is smaller compared to the base, small, and medium models. We believe it can be attributed to the fact that larger models, due to their increased complexity and capacity, are more robust to noise and thus have less room for improvement through additional adaptation techniques. In contrast, smaller models benefit more from the proposed approach, as they are more susceptible to noise and can leverage the additional training to better handle such conditions. Interestingly, despite the general trend of smaller models benefiting more from adaptation, the tiny model shows the least improvement in most cases. We hypothesize that this is due to the limited capacity of the tiny model, which restricts its ability to capture the complex patterns and features necessary for robust speech recognition, even with the assistance of the proposed approach. Therefore, while MetaCORA provides an improvement to smaller models, the in-

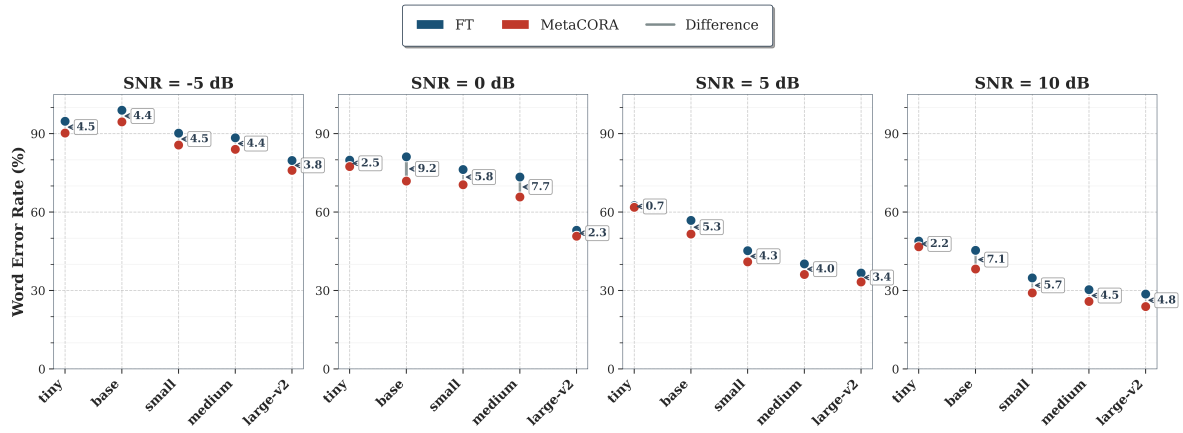

Figure 4: Performance comparison of MetaCORA and traditional fine-tuning across different Whisper model sizes in noisy environments.

herent limitations of the tiny version may constrain its potential gains.

While our evaluation demonstrates consistent robustness improvements under controlled conditions, we acknowledge the limitations of a white-noise-only evaluation strategy. Spectrally flat white noise does not fully capture the complexity of real-world acoustic interference, such as human babble, traffic noise, or room reverberation, which typically exhibit dynamic and frequency-dependent spectral profiles. Consequently, the robustness gains observed in this benchmark may not transfer uniformly to all naturalistic noisy environments.

4.5. Representation Space Analysis

We analyze the representation space of our models to explore how contrastive learning enhances the robustness of speech recognition at the feature level. For audio samples in our test set, we first extract encoder representations from both the baseline models and the proposed models. For each clean sample, we generate a corresponding adversarial sample with perturbation magnitude $\epsilon = 0.05$ and step size $\alpha = 0.005$. We then apply t-SNE to project these high-dimensional features into a two-dimensional space for comparative visualization, as illustrated in Figure 5. To quantify the differences between feature spaces, we compute the average Euclidean distance between each clean sample and its corresponding adversarial sample in the t-SNE embedded space.

Table 2 shows contrastive learning significantly reduces the average distance between clean-adversarial sample pairs in the feature space. For the Tiny model, this distance decreases from 80.11 to 49.70, representing a reduction of approximately 37.96%. Across all model sizes we observe similar improvements, with the average distance consistently reduced by 42-47% for larger models. This

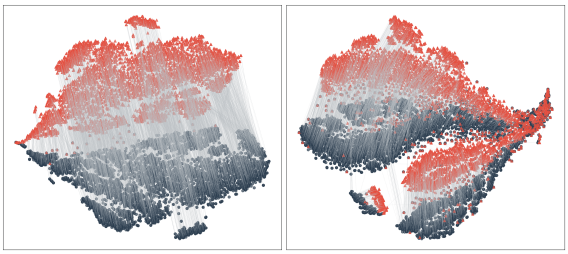

Figure 5: t-SNE visualization of encoder representations in the feature space for the Whisper-medium model. The black points denote clean samples, while the red points represent adversarial samples. The left panel illustrates the feature space distribution after traditional fine-tuning, whereas the right panel demonstrates the improved feature space structure after applying our contrastive learning approach (results from other model sizes followed similar patterns).

indicates that contrastive learning indeed causes the model to map clean samples and adversarial samples of the same content to closer positions in the representation space. This reduction in feature space distance directly corresponds to improved adversarial robustness, as the model learns to treat original inputs and their adversarially perturbed versions as essentially the same content. Another noteworthy observation is that despite the reduced distance between corresponding sample pairs, the overall structure of the feature space maintains good discriminability. It suggests that the model preserves its ability to distinguish different audio content while improving robustness.

5. Conclusion

We present MetaCORA, a method for fine-tuning Whisper on low-resource Hong Kong Cantonese

Model Size	Avg Pair Distance		Diff (%)
	Whisper	MetaCORA	
Tiny	80.11	49.70	37.96
Base	78.22	45.21	42.20
Small	76.58	44.38	42.04
Medium	74.45	41.54	44.20
Large-v2	73.98	39.15	47.08

Table 2: Feature space representation comparison between vanilla Whisper models and our MetaCORA models. The average pair distance measures the Euclidean distance between clean and adversarial sample pairs in t-SNE space. Diff (difference) percentage shows the relative reduction in pair distance achieved by MetaCORA.

that adapts adversarial and contrastive learning during training based on the model's progress. Across all model sizes, MetaCORA yields lower WER scores than standard fine-tuning, commercial ASR systems, and LLM-based alternatives. Ablation studies confirm that both adversarial and contrastive learning are essential, and that dynamically adjusting their relative weights during training is critical to the performance gains. We further observe that MetaCORA improves robustness under noisy conditions and reduces the feature-space distance between clean and adversarial samples while preserving discriminative class boundaries.

While the empirical evidence in this study is limited to Hong Kong Cantonese, the MetaCORA framework is designed to be language-agnostic. The core mechanisms, such as gradient-based adversarial perturbations in the acoustic feature space, contrastive representation learning, and validation-driven meta-optimization, do not rely on Cantonese-specific phonology, lexicons, or linguistic rules. Because the framework operates directly on the latent representations extracted by the pre-trained Whisper encoder, we hypothesize that it possesses broader applicability to other low-resource languages facing similar data scarcity and acoustic variance challenges. In future work, we plan to extend this framework to a wider variety of low-resource languages and evaluate its performance across diverse speech recognition scenarios to fully validate its generalizability.

6. Acknowledgments

We acknowledge LSCM for their crucial support in making this research possible. LSCM's exceptional R&D environment and their dedication to solving real-world challenges in low-resource settings provided the foundational motivation for this work. We thank them for their continuous efforts in cutting-edge AI research and application.

7. Bibliographical References

Ayo Adedeji, Sarita Joshi, and Brendan Doohan. 2024. The sound of healthcare: Improving medical transcription asr accuracy with large language models. *arXiv preprint arXiv:2402.07658*.

Jing An, Yanbing Bai, Jiyi Li, Lifei Wang, Yuyi Jiang, and Yikui Zhang. 2024. Cantonese dialect transcription in diverse sophisticated scenarios via the openai whisper speech recognition model. In *International Conference on Neural Information Processing*, pages 317–328. Springer.

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.

Grace Wenling Cao, Vincent Hughes, Bruce Xiao Wang, and Peggy Mok. 2024. Cross-language forensic voice comparison of hong kong trilingual speakers using filled pauses and an automatic speaker recognition system. In *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 279–283. IEEE.

Nicholas Carlini and David Wagner. 2018. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE security and privacy workshops (SPW)*, pages 1–7. IEEE.

Yaqi Chen, Xukui Yang, Hao Zhang, Wenlin Zhang, Dan Qu, and Cong Chen. 2024. Meta adversarial learning improves low-resource speech recognition. *Computer Speech & Language*, 84:101576.

Ronald Cumbal, Birger Moell, José Lopes, and Olof Engwall. 2024. You don't understand me!: Comparing asr results for l1 and l2 speakers of swedish. *arXiv preprint arXiv:2405.13379*.

Zheng Fang, Tao Wang, Lingchen Zhao, Shenyi Zhang, Bowen Li, Yunjie Ge, Qi Li, Chao Shen, and Qian Wang. 2024. Zero-query adversarial attack on black-box automatic speech recognition systems. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 630–644.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR.

Rupak Raj Ghimire, Prakash Poudyal, and Bal Krishna Bal. 2024. Improving on the limitations of

the ASR model in low-resourced environments using parameter-efficient fine-tuning. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 408–415, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLPAI).

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.

Zhibo Jin, Jiayu Zhang, Zhiyu Zhu, Chenyu Zhang, Jiahao Huang, Jianlong Zhou, and Fang Chen. 2024. Enhancing adversarial attacks via parameter adaptive adversarial attack. *arXiv preprint arXiv:2408.07733*.

Yeonjoon Jung, Jaeseong Lee, Seungtaek Choi, Dohyeon Lee, Minsoo Kim, and Seung-won Hwang. 2024. Interventional speech noise injection for asr generalizable spoken language understanding. *arXiv preprint arXiv:2410.15609*.

Junghun Kim, Ka Hyun Park, and U Kang. 2024. Accurate semi-supervised automatic speech recognition via multi-hypotheses-based curriculum learning. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 40–52. Springer.

Changye Li, Weizhe Xu, Trevor Cohen, and Serguei Pakhomov. 2024. Useful blunders: Can automated speech recognition errors improve downstream dementia classification? *Journal of biomedical informatics*, 150:104598.

Yun Liu, Xuechen Liu, and Junichi Yamagishi. 2024. Improving curriculum learning for target speaker extraction with synthetic speakers. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 364–370. IEEE.

Yunpeng Liu, Xukui Yang, Jiayi Zhang, Yangli Xi, and Dan Qu. 2025. Taml-adapter: Enhancing adapter tuning through task-agnostic meta-learning for low-resource automatic speech recognition. *IEEE Signal Processing Letters*, 32:636–640.

Florian Lux and Ngoc Thang Vu. 2021. Meta-learning for improving rare word recognition in end-to-end asr. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5974–5978.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

Namgyu Park and Jong Kim. 2024. Toward robust asr system against audio adversarial examples using agitated logit. *ACM Transactions on Privacy and Security*, 27(2):1–26.

Andrés Piñeiro-Martín, Carmen García-Mateo, Laura Docío-Fernández, María del Carmen López-Pérez, and Georg Rehm. 2024. Weighted cross-entropy for low-resource languages in multilingual speech recognition. *arXiv preprint arXiv:2409.16954*.

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.

Hao Shi and Tatsuya Kawahara. 2024. Exploration of adapter for noise robust automatic speech recognition. *arXiv preprint arXiv:2402.18275*.

Francisco Teixeira, Karla Pizzi, Raphael Olivier, Alberto Abad, Bhiksha Raj, and Isabel Trancoso. 2024. Improving membership inference in asr model auditing with perturbed loss features. *arXiv preprint arXiv:2405.01207*.

Huimeng Wang, Zengrui Jin, Mengzhe Geng, Shujie Hu, Guinan Li, Tianzi Wang, Haoning Xu, and Xunying Liu. 2024. Enhancing pre-trained asr system fine-tuning for dysarthric speech recognition using adversarial data augmentation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12311–12315. IEEE.

Yufeng Yang, Ashutosh Pandey, and DeLiang Wang. 2024. Towards decoupling frontend enhancement and backend recognition in monaural robust asr. *arXiv preprint arXiv:2403.06387*.

Rui Zhou, Takaki Koshikawa, Akinori Ito, Takashi Nose, and Chia-Ping Chen. 2024. Multilingual meta-transfer learning for low-resource speech recognition. *IEEE Access*, 12:158493–158504.