Beyond Acoustic Sparsity and Linguistic Bias: A Prompt-Free Paradigm for Mispronunciation Detection and Diagnosis

Anonymous submission to Interspeech 2026

Abstract

Mispronunciation Detection and Diagnosis (MDD) requires modeling fine-grained acoustic deviations. However, current ASR-derived MDD systems often face inherent limitations. In particular, CTC-based models favor sequence-level alignments that neglect transient mispronunciation cues, while explicit canonical priors bias predictions toward intended targets. To address these bottlenecks, we propose a prompt-free framework decoupling acoustic fidelity from canonical guidance. First, we introduce CROTTC, an acoustic model enforcing monotonic, frame-level alignment to accurately capture pronunciation deviations. Second, we implicitly inject mispronunciation information via the IF strategy under the knowledge transfer principle. Experiments show CROTTC-IF achieves a 71.77% F1-score on L2-ARCTIC and 71.70% F1 on the Iqra'Eval2 challenge. With empirical analysis, we demonstrate that decoupling acoustics from explicit priors provides highly robust MDD[1].

Index Terms: Mispronunciation detection and diagnosis Optimal transport, Consistency Regularization, Knowledge Transfer, Large Language Model.

1. Introduction

Mispronunciation Detection and Diagnosis (MDD) plays an indispensable role across a spectrum of applications, ranging from general Computer-Aided Pronunciation Training (CAPT) for L2 learners to religious domains such as Qur'anic recitation, as highlighted by the Iqra'Eval Challenges [1, 2]. The core challenge of MDD lies in acoustic fidelity: transcribing speech exactly as realized to pinpoint deviations from canonical norms. Historically, MDD was formulated as a downstream application of Automatic Speech Recognition (ASR). Conventional systems typically used ASR features to generate Goodness-of-Pronunciation (GOP) scores [3] or relied on Extended Recognition Networks (ERNs) to catch predefined error patterns [4, 5]. In this classical paradigm, MDD was essentially an ASR by-product.

With the rapid advancement of end-to-end (E2E) architectures and self-supervised learning (SSL) speech representations, the paradigm has shifted. MDD is now a standalone task, directly optimized for fine-grained phoneme diagnosis. However, despite this structural independence, MDD research inevitably remains under the heavy methodological shadow of ASR. While general ASR focuses on deducing semantic intent of a speaker despite acoustic imperfections, MDD demands objective, fine-grained acoustic fidelity. By uncritically adopting ASR's acoustic and linguistic optimization strategies, we believe that current MDD studies frequently fall into two fatal traps.

The first is the *Acoustic Trap*, driven by the direct adoption of Connectionist Temporal Classification (CTC) [6]. As a de facto standard for E2E ASR, CTC usually serves as an acoustic model (AM) by maximizing the marginal probability of all valid alignment paths, forcing the model to prioritize global sequence correctness. However, when ported to the MDD task, this global optimization actively smooths over the subtle, transient acoustic variations (e.g., co-articulation, brief onset substitutions). In doing so, CTC might erase the exact phonetic evidence that is critical for mispronunciation diagnosis.

The second is the *Linguistic Trap*, driven by the reliance on explicit canonical prompts or strong language model (LM) post-processing. In typical ASR systems such as CTC/AED [7] or RNN-T [8], LMs are powerful components to correct acoustic errors. By adjusting the AM/LM weight, ASR models can shift their preference to favor either acoustic or linguistic features. In MDD, however, when models are guided by canonical priors, they naturally exhibit an over-correction tendency. Instead of objectively diagnosing the actual phonetic deviations from the acoustic signal, the system defaults to the linguistically most likely sequence, which is almost always the canonical text.

To break free from ASR's legacy and overcome these bottlenecks, we propose **CROTTC-IF**, a unified, canonical prompt-free paradigm, designed exclusively for the standalone MDD task. Our main contributions are summarized as follows:

- **Frame-wise Acoustic Modeling (CROTTC):** We introduce **C**onsistency-**R**egularized **O**ptimal **T**emporal **T**ransport **C**lassification. Unlike standard CTC, this acoustic model enforces strict monotonic, frame-wise alignments, effectively mitigating inherent sparsity and delayed emission issues to preserve fine-grained, transient mispronunciation cues.
- **Knowledge Transfer via Indirect Fusion (IF):** Grounded in the Learning Using Privileged Information (LUPI) paradigm, we introduce **I**ndirect **F**usion stragety to implicitly integrate linguistic priors. By treating canonical texts and mispronunciation patterns as privileged data during training, IF provides soft linguistic guidance to the language model, preventing the override of faithful acoustic evidence during inference.
- **Comprehensive Analysis of Canonical Bias:** With delicately designed Multi-modality Large Language Models (LLMs) and various prompt templates, we quantitatively investigate the impact of explicit canonical prior in MDD. Our findings demonstrated that over-reliance on canonical priors can degrade detection sensitivity, highlighting the necessity of balancing linguistic context with acoustic fidelity.
- **State-of-the-Art Performance:** Operating entirely without auxiliary data or explicit canonical prompts, the CROTTC-IF framework achieves highly competitive results. It demon-

[1] Anonymized for double-blind review. Source code, recipes, and checkpoints will be publicly available upon acceptance.

strates strong generalization across diverse benchmarks, ranging from general L2 English corpora (L2-ARCTIC, ERJ, speechocean762) to the specialized Arabic Qur'anic recitation task (Iqra'Eval2).

2. Related Works

Currently, modern MDD approaches broadly fall into two categories: dictation-style and text-prompting style.

2.1. Dictation-Style MDD and the Acoustic Trap

Dictation-style methods aim to recognize the uttered phoneme sequence exclusively from acoustic-related features. For example, Leung et al. introduced a CTC framework and demonstrated the feasibility of capturing L2 phoneme information from acoustic features alone [9]. More recent studies have leveraged self-supervised learning (SSL) speech representations, such as wav2vec 2.0 [10], HuBERT [11, 12] and WavLM [13], which provide robust phonemic context features. For instance, [14] employed a fine-tuned wav2vec 2.0 model with a CTC decoder, achieving promising diagnostic accuracy. To strengthen free phoneme recognition, researchers have also incorporated semi-supervised pseudo-labeling [15], dual-path SSL for predicting manner-of-articulation [16], and acoustic-to-articulatory inversion leveraging electromagnetic articulography (EMA) data [17].

This dictation-style MDD also dominated the recent Iqra'Eval 2025 challenge [1]. For instance, *Hafs2Vec* [18] utilized extensive professional reciter data, while *Metapseud* [19] applied curriculum learning and beam-search decoding. Despite these feature and data enhancements, the fundamental reliance on CTC decoders introduces a potential vulnerability: the *Acoustic Trap*[2].

Conversely, frame-wise alignment methods have demonstrated greater potential for MDD. For example, Feng et al. pioneered the use of attention-based sequence labeling to map individual acoustic frames directly to phone labels [20]. Similarly, Lin et al. modeled MDD as a sequence labeling task via a joint-alignment mechanism [21]. More recently, Tu et al. introduced a retrieval-based framework for frame-wise classification, achieving competitive performance without task-specific training [22]. These advancements indicate that leveraging **dense, frame-wise representations** provides superior performance, outperforming sparse features for robust MDD.

2.2. Text-Prompting MDD and the Linguistic Trap

In reading-aloud scenarios, canonical phoneme sequences are inherently available. Consequently, integrating this prior knowledge into MDD models has become another standard practice. For instance, Ye et al. proposed an approach to leverage acoustic, phonetic and linguistic embeddings for MDD [23]. Peng et al. employed a gating mechanism and contrastive loss to dynamically regulate canonical embeddings [24]. Zheng et al. utilized coupled-cross attention for explicit acoustic-canonical fusion [25]. Similarly, Yan et al. injected implicit linguistic priors using phoneme lookup tables and graph networks to model frequent mispronunciation patterns [26, 27].

Entering the era of Large Language Models (LLMs), recent studies have shifted toward leveraging massive linguistic priors for posterior correction in MDD. Current methods range from adapting LLM-based ASR architectures [28] to fine-tuning native SpeechLM [29]. For instance, Wu et al. adapted LLMs for MDD by coupling a pretrained acoustic front-end with a trainable projector and LLM backbone [30]. Other researchers adapted multi-modality LLM to embed canonical sequences and predefined error patterns directly into the latent space [31, 32]. Beyond simple transcription, LLM reasoning is also deployed to generate educational feedback, offering nuanced diagnostics alongside canonical norms [33, 34].

Despite reported successes, the practical superiority of text-prompted and LLM-assisted methods remains debatable. Many systems are evaluated on private datasets, precluding reproducible comparisons. And on open benchmarks, they frequently underperform compared to conventional models that rely solely on acoustic modeling.

Beyond evaluation discrepancies, we believe that some of these canonical-dependent methods inherently fall into the *Linguistic Trap* due to a shared flaw: **canonical information leakage**. Whether by explicitly including canonical information as part of the input, or by using a strong language model as the model's prior knowledge, these approaches can all be categorized as suffering from canonical information leakage. Although many models carefully design their canonical information usage, canonical information can easily override the subtle acoustic information. Consequently, the model becomes overly tolerant of errors, masking critical acoustic deviations and sacrificing diagnostic objectivity. Furthermore, requiring explicit canonical texts during inference renders these methods entirely impractical for spontaneous speech or shadowing scenarios [35–37], restricting the real-world applicability of MDD.

3. Consistency Regularization on Optimal Temporal Transport Classification

In this section, we analyze the inherent limitations of CTC and introduce the architecture of our proposed frame-wise dense acoustic model: **CROTTC**. Our approach, as illustrated in the left panel of Figure 1, is built upon two core pillars: Consistency Regularization **(CR)** and Optimal Temporal Transport Classification **(OTTC)**.

3.1. Limitations of Connectionist Temporal Classification

Connectionist Temporal Classification (CTC) [6] has established itself as a fundamental criterion for sequence-to-sequence modeling, particularly in the ASR domain. By introducing a blank token and optimizing the marginal probability over all valid alignments, CTC enables end-to-end training without requiring explicit frame-level annotations for target units (e.g., phonemes or subwords). Consequently, the model tends to produce a highly sparse posterior distribution, concentrating probability mass on a few highly discriminative frames [38]. Moreover, these non-blank emission peaks are frequently delayed, shifting towards the tail end of their corresponding acoustic segments—a phenomenon widely recognized as CTC's delayed behavior [39, 40].

This sparsity and delayed behavior pose a significant challenge for precise segment detection tasks. In practice, a mispronounced phoneme in MDD does not necessarily manifest as a complete substitution starting from the very first frame; rather, it often involves co-articulation errors or a partial blend of canonical and abnormal acoustic traits. To illustrate, consider a speaker attempting to produce the diphthong /aɪ/ (as in "buy") but erroneously substituting the initial vowel element to produce /ɔɪ/ (as in "boy"). Acoustically, both diphthongs share

[2]Further detailed in Sec. 3.1

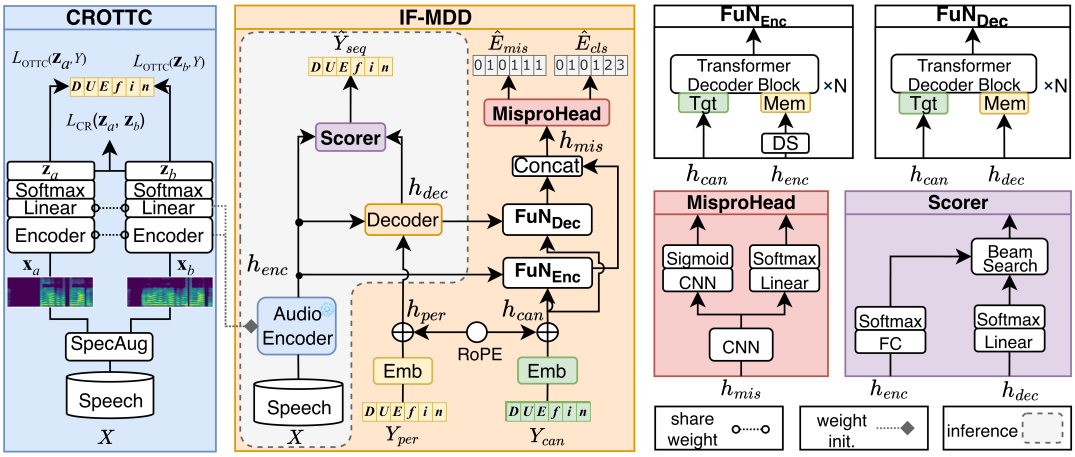

Figure 1: *Overview of the CROTTC-IF architecture. From left to right, it comprises the CROTTC AM (blue), the lightweight IF-MDD LM (orange), and the detailed IF components.*

an identical trailing off-glide (/ɪ/). Driven by CTC's inherent delayed emission tendency, the model will typically align the probability peak for the target label (/aɪ/) exclusively with this shared off-glide. Consequently, the preceding frames containing the crucial evidence of the mispronunciation (/ɔ/) can be masked by blank tokens and entirely ignored. While this highly discriminative frame selection is beneficial for standard ASR, it discards the fine-grained acoustic cues that are essential for sensitive diagnostic tasks like MDD.

3.2. Optimal Temporal Transport Classification

To address the inherent shortcomings of CTC, we leverage Optimal Temporal Transport Classification (OTTC) [41], a differentiable framework rooted in one-dimensional optimal transport that enables dense, frame-level alignments.

3.2.1. Sequence Alignment via Optimal Transport

Let $\mathbf{X} = \{x_i\}_{i=1}^{n}$ denote the acoustic frame sequence and $\mathbf{Y} = \{y_j\}_{j=1}^{m}$ the perceived phoneme sequence. We define that the alignment between $\mathbf{X}$ and $\mathbf{Y}$ can be solved with an optimal transport plan $\gamma \in \mathbb{R}_{+}^{n\times m}$, which serves as a monotonic mapping governing the frame-to-label correspondence. In optimal transport theory, γ is parameterized by two discrete distributions: the frame weights $\alpha \in \Delta^n$ derived from $\mathbf{X}$, and the label weights $\beta \in \Delta^m$ associated with $\mathbf{Y}$, where $\Delta^k = \{v \in \mathbb{R}_{+}^{k} \mid \sum_{i=1}^{k} v_i = 1\}$ is the probability simplex. Here, the frame weights α can be predicted from the input via a neural network W:

$$\alpha[\mathbf{X}, W] = \mathrm{softmax}(W(x_1), \ldots, W(x_n))^{\top}, \qquad (1)$$

and β is typically fixed as a uniform distribution to evenly assign each target token's contribution. Subsequently, the alignment plan $\gamma(\alpha, \beta)$ is computed as the solution to the 1D optimal transport problem:

$$\gamma(\alpha, \beta) = \arg\min_{\gamma \in \Gamma_{\alpha,\beta}} \sum_{i=1}^{n}\sum_{j=1}^{m} \gamma_{i,j} \, \|i - j\|_2^2, \qquad (2)$$

where $\Gamma_{\alpha,\beta} = \left\{\gamma \in \mathbb{R}_{+}^{n\times m} \mid \gamma \mathbf{1}_m = \alpha, \, \gamma^{\top}\mathbf{1}_n = \beta\right\}$ denotes the set of valid couplings preserving the marginal distributions. Based on the *Sequence Optimal Transport Distance (SOTD)* framework [41], which defines a pseudo-metric over sequence space, we derive the OTTC loss as the expected transport cost under the cross-entropy metric:

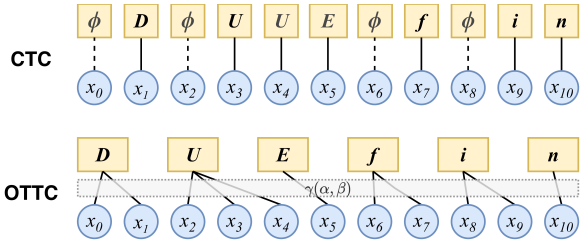

Figure 2: *Comparison between CTC and OTTC, where ϕ is the blank token and $\gamma(\alpha, \beta)$ is the optimal transport plan.*

$$\mathcal{L}_{\mathrm{OTTC}} = -\sum_{i=1}^{n}\sum_{j=1}^{m} \gamma_{i,j}(\alpha, \beta) \cdot \log p(y_j \mid x_i), \qquad (3)$$

where $p(y_j \mid x_i)$ is the posterior probability. Since α is parameterized via a neural network W (Eq. 1) and β is fixed, the model parameters can be optimized in an end-to-end manner by minimizing $\mathcal{L}_{\mathrm{OTTC}}$.

Comparison to CTC. Fig. 2 illustrates the fundamental distinction in alignment strategies. Specifically, CTC maximizes the marginal probability over *all* valid monotonic paths:

$$\mathcal{L}_{\mathrm{CTC}} = -\log \sum_{\pi \in \mathcal{B}^{-1}(\mathbf{Y})} \prod_{i=1}^{n} p(\pi_i \mid x_i), \qquad (4)$$

where $\mathcal{B}^{-1}(\mathbf{Y})$ denotes the set of all paths collapsed to $\mathbf{Y}$. In contrast, OTTC maximizes the probability of a *single* optimal monotonic path $\gamma(\alpha, \beta)$ via Eq. (2) and (3). This formulation yields dense, frame-level alignments without blank-token dominance, as γ explicitly models the frame-to-label correspondence rather than marginalizing over sparse paths.

3.3. Consistency Regularization

Since OTTC takes all frames into consideration, acoustic models can be over-sensitive to local variations, leading to spurious insertions. To mitigate this, we also introduce *Consistency Regularization (CR)* into our training pipeline [42]. Specifically, as shown in Fig 1, for each input utterance $\mathbf{X}$, we generate two augmented views $\mathbf{X}_a$ and $\mathbf{X}_b$ using stochastic perturbations[3]. Let $\mathbf{Z}_a$ and $\mathbf{Z}_b$ denote the posterior probability distributions output by the model for each view, where $\mathbf{Z}^{(i)} \in \Delta^K$ represents the probability vector at frame i (equivalent to $p(y \mid x_i)$

[3]See Sec. 6.1 for spectrogram augmentation details.

in Eq. 3). The consistency regularization loss is defined as the symmetric Kullback-Leibler (KL) divergence between these distributions:

$$\mathcal{L}_{\text{CR}} = \frac{1}{2n}\sum_{i=1}^{n}\left(\text{KL}\left(\mathbf{Z}_a^{(i)}||\mathbf{Z}_b^{(i)}\right) + \text{KL}\left(\mathbf{Z}_b^{(i)}||\mathbf{Z}_a^{(i)}\right)\right), \quad (5)$$

where n is the number of frames. By minimizing the distribution distance and maximizing the mutual information between the two branches, CR enables each branch to receive supervision from the other, which encourages the shared encoder to perform self-distillation on frame-level posterior distributions. Additionally, relatively long-term time masking implicitly encourages *masked language modeling*-like behavior: the model learns to reconstruct occluded acoustic frames by leveraging contextual information from unmasked regions. Together, these mechanisms stabilize frame-level predictions and reduce sensitivity to local acoustic noise. Consequently, the training criteria for CROTTC is defined as:

$$\mathcal{L}_{\text{AM}} = \mathcal{L}_{\text{CR}} + \eta\left(\mathcal{L}_{\text{OTTC}}(\mathbf{Z}_a, \mathbf{Y}) + \mathcal{L}_{\text{OTTC}}(\mathbf{Z}_b, \mathbf{Y})\right), \quad (6)$$

where hyperparameter η is set to 1.0.

4. Indirect Fusion of Mispronunciation Information via Knowledge Transfer

In this section, we detail our Indirect Fusion (IF) strategy for incorporating mispronunciation information into LM, named **IF-MDD**. As illustrated in the middle and right panels of Fig. 1, IF-MDD treats canonical phonemes and mispronunciation cues as privileged information available exclusively during the training phase. These cues are leveraged to guide the model's latent representations via backpropagation.

4.1. Learning using Privilege Information

Learning using Privileged Information (LUPI) is a training paradigm that fundamentally operates as a knowledge transfer mechanism, leveraging additional information available exclusively during the training phase to implicitly guide the model [43, 44]. While this paradigm has already been successfully applied to various speech-related tasks [45–47], its most compelling advantage lies in its effectiveness when training data is limited or labels are scarce [48–50]. Given the inherent data scarcity of expert-annotated MDD corpora, adopting the LUPI framework is a natural and highly principled choice.

4.2. Indirect Fusion of Mispronunciation Information via Knowledge Transfer

As illustrated in Fig. 1, IF-MDD consists of two main components: 1) an encoder–decoder backbone; 2) an auxiliary mispronunciation-aware teacher network.

4.2.1. Encoder-decoder backbone

We employ an encoder–decoder architecture serving as the primary backbone of IF-MDD. Let the input waveform be $\mathbf{X} = \{x_i\}_{i=1}^{n}$ and the perceived phoneme sequence be $\mathbf{Y}_{\text{per}} = \{y_j^{\text{per}}\}_{j=1}^{m}$, where n and m denote the lengths of the speech signal and phoneme sequence, respectively. The backbone is formulated as:

$$\mathbf{h}_{\text{enc}} = \text{Enc}(\text{RoPE}(\mathbf{X})), \quad (7)$$

$$\mathbf{h}_{\text{per}} = \text{RoPE}(\mathbf{Y}_{\text{per}}), \quad (8)$$

$$\mathbf{h}_{\text{dec}} = \text{Dec}(\mathbf{h}_{\text{per}}, \mathbf{h}_{\text{enc}}), \quad (9)$$

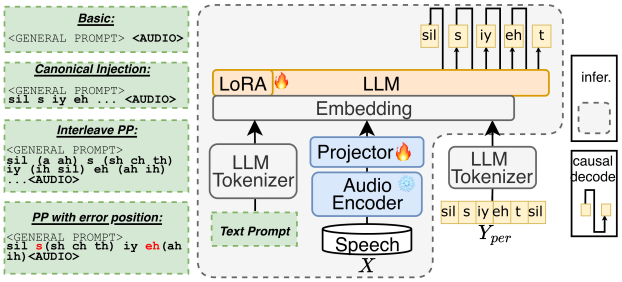

Figure 3: *Illustration of LLM-based MDD, with alternative prompts.*

where $\text{Enc}(\cdot)$ and $\text{Dec}(\cdot)$ represent the encoder (AM) and decoder (LM), respectively. Rotary position embeddings $\text{RoPE}(\cdot)$ are applied to both $\mathbf{h}_{\text{enc}} \in \mathbb{R}^{n\times D}$ and $\mathbf{h}_{\text{dec}} \in \mathbb{R}^{m\times D}$ to enhance positional modeling [51].

4.2.2. Auxiliary mispronunciation-detection teacher network

To explicitly model the mispronunciation information indicated by the misalignment between acoustic and canonical features, we introduce an auxiliary teacher network. This module leverages the canonical phoneme sequence $\mathbf{Y}_{\text{can}} = \{y_j^{\text{can}}\}_{j=1}^{m}$ and the corresponding phoneme-level ground-truth error sequence $\mathbf{E} = \{e_j\}_{j=1}^{m}$ as privileged information. It consists of two sequential stages: feature fusion and comprehensive error detection.

First, two parallel fusion networks (FuN) are employed to model multi-aspect feature interactions:

$$\mathbf{h}_{\text{can}} = \text{RoPE}(\mathbf{Y}_{\text{can}}), \quad (10)$$

$$\mathbf{h}_{\text{mis}}^{\text{enc}} = \text{FuN}_{\text{enc}}(\mathbf{h}_{\text{can}}, \text{DS}(\mathbf{h}_{\text{enc}})), \quad (11)$$

$$\mathbf{h}_{\text{mis}}^{\text{dec}} = \text{FuN}_{\text{dec}}(\mathbf{h}_{\text{can}}, \mathbf{h}_{\text{dec}}), \quad (12)$$

$$\mathbf{h}_{\text{mis}} = \text{Concat}(\mathbf{h}_{\text{mis}}^{\text{enc}}, \mathbf{h}_{\text{mis}}^{\text{dec}}), \quad (13)$$

where $\text{FuN}(\cdot)$ is composed of a transformer decoder structure and $\mathbf{h}_{\text{mis}} \in \mathbb{R}^{m\times D}$ denotes the fused representation. In both branches, $\text{FuN}(\cdot)$ uses $\mathbf{h}_{\text{can}}$ as the query and $\mathbf{h}_{\text{enc}}$ or $\mathbf{h}_{\text{dec}}$ as the memory. Notably, a downsampler $\text{DS}(\cdot)$ is applied within $\text{FuN}_{\text{enc}}(\cdot)$ to improve alignment.

Subsequently, our teacher network employs a dual-path detection sub-network to provide a comprehensive diagnosis. Specifically, we apply a shared CNN trunk on the fused representation $\mathbf{h}_{\text{mis}}$, followed by two task-specific projection heads:

$$\mathbf{u} = \text{CNN}(\mathbf{h}_{\text{mis}}), \quad (14)$$

$$\hat{E}_{\text{mis}} = \sigma(\text{CNN}_{\text{bin}}(\mathbf{u})), \quad (15)$$

$$\hat{E}_{\text{cls}} = \text{Softmax}(\text{Linear}_{\text{cls}}(\mathbf{u})), \quad (16)$$

where $\hat{E}_{\text{mis}}$ and $\hat{E}_{\text{cls}}$ are the predicted counterparts for the position and classification facets of $\mathbf{E}$, respectively. Here, $\hat{E}_{\text{mis}} \in [0,1]^m$ indicates the probability of mispronunciation, while $\hat{E}_{\text{cls}}$ denotes the distribution over $\{$*correct*, *substitution*, *deletion*, *insertion*$\}$.

Crucially, this architectural design naturally facilitates knowledge transfer. Since the auxiliary mispronunciation heads are relatively lightweight and receive direct supervision from $\mathbf{E}$, they converge significantly faster than the primary sequence-to-sequence backbone. During training, this fast-converging teacher provides a strong, discriminative gradient signal that backpropagates through the fusion networks into both $\mathbf{h}_{\text{enc}}$ and $\mathbf{h}_{\text{dec}}$. Consequently, the diagnostic features from the teacher

network act as a catalyst, helping the deeper backbone to converge effectively and internalize the mispronunciation-aware representations required for prompt-free inference.

4.3. Inference and Training Criteria

During inference, the auxiliary teacher is discarded. A scorer searches for the optimal hypothesis $\hat{\mathbf{Y}}$ by maximizing the interpolated log-probabilities between the AM and LM:

$$\hat{\mathbf{Y}} = \arg\max_{\mathbf{Y}} \big\{ \lambda \log P_{\text{AM}}(\mathbf{Y}|\mathbf{h}_{\text{enc}}) + (1-\lambda) \log P_{\text{LM}}(\mathbf{Y}|\mathbf{h}_{\text{enc}}) \big\}, \tag{17}$$

where $\lambda \in [0,1]$ is a AM decoding weight. Here, P_{AM} represents the acoustic probability predicted by the CROTTC branch, and P_{LM} is the contextual probability evaluated autoregressively by $\text{Dec}(\cdot)$. The IF-MDD model is optimized via a multi-task objective:

$$\mathcal{L}_{\text{total}} = \omega_1 \mathcal{L}_{\text{AM}} + (1-\omega_1)\mathcal{L}_{\text{LM}} + \omega_2(\mathcal{L}_{\text{pos}} + \mathcal{L}_{\text{type}}) + \omega_3 \mathcal{L}_{\text{ga}}, \tag{18}$$

where $\mathcal{L}_{\text{ga}}$ is the guided-attention loss [52] ensuring monotonic fusion in FuN. Hyperparameters $\omega_1, \omega_2, \omega_3$ are set to 0.3, 1.0, and 10.0, respectively.

5. Leveraging LLMs to Investigate Canonical Information Effect on MDD

To empirically investigate the impact of explicit canonical information and strong linguistic priors, this section introduces an LLM-based architecture, denoted as **LLM-MDD**. As illustrated in Fig. 3, we replace the conventional lightweight LM with an open-source LLM, which serves as a high-capacity linguistic processor for the acoustic embeddings. We adopt an LLM for this canonical information analysis for two primary reasons. First, LLMs inherently possess massive language modeling capabilities; therefore, leveraging an LLM to replace a lightweight LM itself represents a significant amplification of linguistic priors. Second, multimodal LLMs natively support the direct integration of text (canonical prompts) and speech (acoustic embeddings). This allows us to isolate and strictly verify the exact impact of various canonical injection strategies, while entirely bypassing complex and task-specific fusion mechanisms.

5.1. LLM-MDD Framework

Inspired by the SLAM-LLM architecture [28], our LLM-MDD system formulates the MDD task as mapping acoustic features and canonical cues to a perceived phoneme sequence. The architecture comprises three primary components: a pretrained AM, a trainable projector, and a backbone LLM.

During training, the AM remains frozen. We optimize the LLM via Low-Rank Adaptation (LoRA) [53] and fully fine-tune the projector to align acoustic features to the LLM's embedding space. The system is optimized using standard cross-entropy loss against ground-truth perceived phoneme sequences. By injecting these massive linguistic priors, we aim to evaluate whether the emergent capabilities of LLMs can genuinely outperform traditional lightweight decoders in diagnosing subtle mispronunciations.

5.2. Canonical Information Injection via Prompting

To explore how explicit canonical information influences the model's diagnostic accuracy, we design four distinct prompting strategies for LLM-MDD:

Basic: A standard SLAM-style sequence where the speech features are followed by the perceived phonemes: `[Prompt]<speech><BOS><perc_phns><EOS>`.

Canonical Injection: Following the implementation in [30], the canonical phoneme sequence is appended to a general prompt: `[Prompt]<cano_phns><speech><BOS><perc_phns><EOS>`.

Interleaved Potential Pronunciation: Following the implementation in [31, 32], we interleave canonical phonemes with potential pronunciation (PP) alternatives: `[Prompt]<cano_phns_with_PP><speech><BOS><perc_phns><EOS>`. In both the training and inference stages, these PP candidates are uniformly assigned to each canonical phoneme. This design ensures that the model receives no prior clues regarding the specific *location* or *type* of error.

Potential Pronunciation with Error Position: Unlike the uniform assignment, we only provide PP candidates for phonemes explicitly annotated as mispronounced: `[Prompt]<cano_phns_with_PP_err_pos><speech><BOS><perc_phns><EOS>`. Since the mispronunciation position is pinpointed, the task reduces to a forced-choice mispronunciation diagnosis. Although this non-causal prompting style is impractical for real-world applications, this configuration is designed to test the performance upper bound of LLMs—specifically, whether they prioritize the canonical or the acoustic features in MDD.

6. Experimental Evaluation

This section comprehensively evaluates our CROTTC-IF framework against conventional dictation- and text-prompting-style baselines, alongside an empirical analysis of canonical priors using LLM-MDD.

6.1. Datasets

Table 1 summarizes the L2 English datasets utilized in our experiments. The most widely recognized among these is L2-ARCTIC, which standardized test set with six speakers (NJS, TLV, TNI, TXHC, YKWK, and ZHAA). Besides, we incorporated two supplementary datasets: Speechocean762 (SO762) and ERJ. SO762 was originally designed for pronunciation assessment, assigning each phoneme a score ranging from 0.0 to 2.0. Phonemes scoring below 0.5 are identified as mispronunciations, and their corresponding expert-annotated phonetic realizations are provided as ground truth. ERJ is a Japanese-accented L2-English corpus that serves as an out-of-domain (OOD) dataset to evaluate generalization ability. For vocabulary preparation, we adopted a 44-unit ARPAbet-style inventory (39 phonemes plus 5 special tokens `<bos>`, `<eos>`, `<sil>`, `<error>`, and `<blank>`).

6.2. Implementation Details

Acoustic Model (AM). For the AM backbone described in Sec. 3, we utilize WavLM Large [13] followed by a 2-layer Conformer [54] with a kernel and stride size of 3. The hidden dimension is set to 384. For the Consistency Regularization, we apply time warping with a factor of 80 alongside time and frequency masking following [42]. Specifically, we use a maximum of 3 mask blocks with a masking ratio $\rho \in (0.1, 0.3]$. In the time domain, the minimum masking length is enforced to ensure the CR mechanism leverages neighboring context rather than acting as simple data augmentation.

Language Model (LM). As introduced in Sec. 4, our LM comprises a 2-layer Transformer decoder with a hidden size

Table 1: *Overview of the L2-ARCTIC, ERJ, and speechocean762 datasets.*

Dataset	L1	Volume	Train Utt.	Val Utt.	Test Utt.	Test Spk.
L2-ARCTIC [57]	Mixed[4]	3.66h	2429	268	900	6
SO762 [58]	Mandarin	5.58h	2250	250	2500	125
ERJ [59]	Japanese[5]	0.99h	674	74	144	8

of 384. We employ a two-stage training strategy: (1) initial joint CTC/Transformer training, followed by (2) substituting the CTC AM with the pretrained CROTTC AM and fine-tuning the LM until convergence. This approach resolves convergence issues arising from the absence of blank tokens in CROTTC, which otherwise provide crucial implicit segmental boundaries. The AM/LM training weight ω_1 is set to 0.5, and the effect of the decoding weight is analyzed in Sec. 6.4.4.

Auxiliary Teacher Network. The teacher network in Sec. 4.2.2 comprises two parallel branches: an encoder-side $\text{FuN}_{\text{enc}}(\cdot)$ and a decoder-side $\text{FuN}_{\text{dec}}(\cdot)$, both implemented with 2-layer Transformer decoders with a hidden size of 384. Notably, the 1D CNN down-sampler (factor=4) is integrated into $\text{FuN}_{\text{enc}}(\cdot)$ for alignment improvement. The shared CNN trunk from the mispronunciation head has a dimension of 128, followed by a 64-dim CNN branch for error-position detection and a linear branch for error type detection, respectively.

LLM-MDD. For the modules mentioned in Sec. 5, we adapted AMs fine-tuned with both CTC and CROTTC. The projector is a 1D-CNN with a down-sampling factor of 4. We employed LLaMA-3.2-1B-Instruct [55] and Qwen3-4B-Thinking-2507 [56] as backbone LLMs, with rank=16 for LoRA fine-tuning. For the potential pronunciation prompting, we conducted a corpus-level statistical analysis, where the top 5 alternatives were selected as the PP candidates.

Miscellaneous. The AMs were trained for 300 epochs with a batch size of 32; learning rates were set to 10^{-5} for the SSL backbone and 3×10^{-4} for the following modules. The LM and Teacher Network were trained for 200 epochs (batch size 32, learning rate 2×10^{-4}). LLM training was limited to 20 epochs (batch size 4, learning rate 2×10^{-4}). During inference, both the LM and LLMs utilized beam search decoding with a beam size of 10 and a temperature of 1.1. All experiments were conducted on a single NVIDIA GH200 GPU. Model selection was determined by the optimal $F1$-score on the validation set.

6.3. Evaluation metrics

We followed the hierarchical evaluation protocol of [61]. Detection was measured by true acceptance (TA), true rejection (TR), false acceptance (FA), and false rejection (FR), while diagnosis was evaluated with correct diagnosis (CD) and error diagnosis (ED) as illustrated in Table 2. The main performance is then summarized with the F1 score, treating TR as the positive class:

$$\text{P} = \frac{TR}{TR+FR}, \quad \text{R} = \frac{TR}{TR+FA}, \quad \text{F1} = \frac{2\text{PR}}{\text{P}+\text{R}}.$$

We also report phoneme error rate (PER) and Correction Rate (COR) as a measure of recognition accuracy:

$$\text{PER} = \frac{S+D+I}{N}, \qquad \text{COR} = 1 - \frac{S+D}{N},$$

where S, D, and I denote the numbers of substitutions, deletions, and insertions, and N is the total number of **perceived phonemes**.

[4]Hindi, Korean, Mandarin, Spanish, Arabic, and Vietnamese.

[5]Only expert-annotated speech were utilized as described in [60].

Table 2: *An illustration of MDD metrics.*

Canonical phoneme	s	p	iy	k	t
Perceived phoneme	s	b	iy	g	d
Predicted phoneme	s	p	ih	g	th
Evaluation result	TA	FA	FR	TR CD	TR ED

Table 3: *Comparison of the proposed methods against various baseline models on L2-ARCTIC.* ***Bold*** *and* underlined *values indicate the best and second-best results among fair-comparison models, respectively.* ↑/↓*: higher/lower is better.*

Model	Detection			Diagnosis			Recognition	
	F1 ↑	P ↑	R↑	FRR ↓	FAR ↓	EDR ↓	PER ↓	COR ↑
Dictation-Style Baselines								
MPL-MDD [15]	55.42	60.39	51.20	5.60	48.80	22.71	14.36	-
RNN-T [62]	59.10	63.40	55.30	5.30	44.70	-	15.47	-
MV-w2v2 [16]	60.31	59.23	61.43	-	-	-	14.13	-
w2v2-CTC [14]	60.44	62.86	58.57	5.70	41.80	29.28	16.20	-
Meta-Learn [63]	61.45	**91.60**	46.24	29.75	**8.40**	-	42.25	-
Text-prompting-Style Baselines								
Qwen2 [34]	50.60	71.62	39.12	-	-	-	-	-
Qwen2-sub [34]	55.00	63.27	48.64	-	-	-	-	-
AEL w/o Pos. [64]	56.33	58.36	55.00	6.55	45.00	25.72	14.81	-
MDDGCN [26]	56.49	51.90	61.97	9.18	38.03	25.27	-	-
Peppanet [27]	56.81	64.53	51.38	9.61	36.47	25.88	-	-
PG-MDD [17]	60.10	60.06	60.15	6.65	39.94	23.13	13.92	-
TG+Contrast. [24]	61.75	62.12	61.38	6.19	38.62	28.92	-	-
Frame-wise MDD Baselines								
Joint-align [21]	63.04	77.12	53.31	-	-	-	-	-
PER-MDD [22]	69.60	71.78	**67.56**	4.43	32.44	37.77	104.08	90.42
Proposed Methods (AM)								
OTTC	63.18	66.36	60.29	5.14	39.71	22.12	18.07	89.96
CROTTC	62.39	69.70	56.47	4.13	43.53	22.06	17.48	90.29
Proposed Methods (LM & LLM)								
CTC-IF	58.37	61.81	55.29	5.75	44.71	**19.98**	**13.72**	88.34
CROTTC-IF	**71.77**	76.94	67.24	**3.39**	32.76	27.47	46.52	**92.42**
CROTTC-LLaMA	56.87	54.81	59.08	8.20	40.92	21.98	15.85	86.55
CROTTC-Qwen	55.19	58.00	52.64	6.42	47.36	23.80	15.42	86.78

6.4. Experimental evaluation on L2-ARCTIC

Table 3 compares our proposed framework against recent representative MDD methods on the L2-ARCTIC benchmark. Notably, CROTTC-IF achieves a peak F1 score of 71.77%, yielding the lowest FRR and highest COR among all baselines. In the following subsections, we systematically analyze how the AM, LM, and canonical injection strategies collectively drive this superior performance.

6.4.1. Escaping the Acoustic Trap: Ablation on CROTTC

Table 4(a) details the performance differences between CTC- and OTTC-based AMs. Compared to standard CTC, CRCTC yields improvements in both F1 score and PER. For OT-based methods, the F1 score exhibits a significant improvement, increasing from 57.89% to 63.18%. Concurrently, the PER also rises from 14.45% to 18.35%, with the most notable change being a surge in insertion errors from 2.21% to 8.31%. The integration of CR effectively mitigates these insertion errors down to 7.76% and reduces the FRR to 4.13%. In MDD, FRR is a critical usability metric; incorrectly flagging a student's correct pronunciation as an error can lead to learner frustration and undermine the pedagogical value of the system. Therefore, we select CROTTC as our primary AM, albeit with a slight trade-off in the overall F1 score.

To understand these acoustic mechanisms, we illustrate the frame-wise probability distributions of argmax predictions

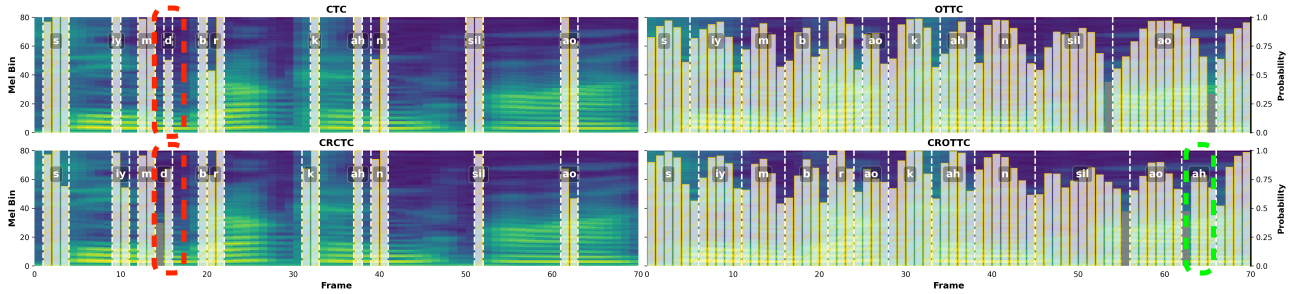

Figure 4: *Comparison of frame-level probability distributions across different AMs (blank tokens omitted). The perceived phoneme sequence is* ***/s iy m (d) b r aa k ah n sil ao ah/****, corresponding to the utterance* ***"seemed broken or" → "seem broken or-ah"****. While the CTC-based method hallucinated the* ***/d/*** *and omitted the* ***/ah/****, our CROTTC faithfully captures the actual phonetic realization.*

Table 4: *Comprehensive ablation studies on L2-ARCTIC. (a) Acoustic model variations. (b) Indirect Fusion components. (c) Canonical prompting strategies on LLM-MDD.*

(a) Ablation of CROTTC AM

Method	F1↑	P↑	R↑	FRR↓	FAR↓	PER↓	INS↓
CTC	57.89	59.40	56.45	6.49	43.55	14.45	2.21
CRCTC	58.45	61.40	55.78	5.90	44.22	**14.01**	**2.05**
OTTC	**63.18**	66.36	**60.29**	5.14	**39.71**	18.35	8.31
CROTTC	62.39	**69.70**	56.47	**4.13**	43.53	17.48	7.76

(b) Ablation of IF LM

Method	F1↑	P↑	R↑	FAR↓	EDR↓
CTC-IF	**58.37**	**61.81**	55.29	44.71	**19.98**
w/o FuN_{Enc}	57.33	60.29	54.65	45.35	21.74
w/o FuN_{Dec}	56.51	59.42	53.86	46.14	22.10
w/o Error Head	57.56	59.54	**55.71**	**44.29**	22.28
CTC-LM	54.95	59.23	51.25	48.75	22.60

(c) Ablation of LLM Canonical Prompts

Method	F1↑	P↑	R↑	FRR↓	FAR↓	EDR↓	PER↓
CTC-LLaMA	55.16	54.76	55.57	7.73	44.43	24.17	16.23
CROTTC-LLaMA	**56.87**	54.81	**59.08**	8.20	**40.92**	**21.98**	15.85
w/ cano.	40.52	**68.22**	28.83	**2.26**	71.17	32.56	**13.55**
w/ PP	42.63	54.11	35.18	5.02	64.82	35.15	14.91
w/ pos. (oracle)	**91.78**	**95.02**	**88.74**	**0.78**	**11.16**	24.72	**5.04**

across different loss functions in Fig. 4. Unlike CTC, OT-based methods assign non-blank tokens to every frame, which increases inference sensitivity but also leads to higher insertion errors during greedy decoding. By incorporating CR, the model effectively captures more subtle pronunciation deviations, successfully inheriting the alignment flexibility of OTTC while maintaining the recognition stability.

6.4.2. Escaping the Linguistic Trap: Implicit Knowledge Transfer via IF

Table 4(b) presents the ablation of the auxiliary teacher network, where CTC served as the AM backbone. Removing either FuN_{Enc} or FuN_{Dec} leads to a clear drop in the F1 score by 1.04% to 1.86%, demonstrating the effectiveness of multi-aspect monitoring in our indirect fusion strategy. Ablating the error-type classification head yields a smaller but consistent decrease in F1 alongside an increase in EDR, suggesting that this multi-view design helps to regularize the mispronunciation representations. Visualizations in Fig. 5 display the cross-attention heatmaps of the fusion network's last layer: removing these components produces blurrier, less-peaky patterns with reduced diagonal dominance, reflecting the degraded representation quality of the mispronunciation cues. Furthermore, without the teacher network's knowledge transfer, the model devolves into a standard CTC/Transformer ASR architecture (CTC-LM), resulting in a significant decrease in the F1 score. This proves that uncritically adapting general LM-based ASR models to the standalone MDD task generates suboptimal diagnostic results.

6.4.3. The Danger of Explicit Priors: Canonical Injection with LLMs

Table 4(c) presents our investigation into the effect of strong canonical priors on MDD performance. First, compared to CTC-LLaMA, CROTTC-LLaMA yields better detection performance, boosting the F1 score from 55.16% to 56.87%. This supports our hypothesis that the sparsity inherent in CTC causes MDD models to overlook transient mispronunciation cues. Second, by varying the canonical information injected into the

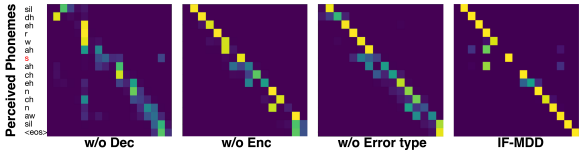

Figure 5: *Attention heatmaps of the Fusion Network under different ablation conditions. The x-axis denotes the memory, and the y-axis denotes the canonical-phoneme embeddings.*

prompt, we observe that explicitly providing canonical targets is highly detrimental to MDD, resulting in a drastic drop in the F1 score to 40.52%. While prompting with Potential Pronunciations (PP) offers the LLM slightly more flexibility, the performance recovery remains severely limited (F1 = 42.63%), consistent with the observation in [34]. These observations strongly demonstrate the "Linguistic Trap" in MDD: strong textual priors bias the LLM toward over-correction, masking objective acoustic fidelity. Third, when prompting the LLM with both PP and explicit mispronunciation positions, the F1 score naturally sees a significant improvement to 91.78%. However, a gap from perfect MDD remains: while FRR drops to 0.78%, EDR and FAR remain high at 24.72% and 11.16%. This proves the fundamental bottleneck for multimodal LLMs in MDD is not merely detecting error positions, but also the inability to utilize the fine-grained acoustic resolution required for precise diagnosis. Even with explicit positional hints, the LLM struggles to accurately identify specific substitutions and stubbornly defaults to the textual prior. Therefore, we believe that if future research aims to utilize multimodal LLMs for MDD, developing more refined acoustic processing stages is essential.

6.4.4. Balancing Acoustic Fidelity and Linguistic Context

As formulated in Eq. 17, CROTTC-IF performs decoding under a joint AM/LM strategy. In standard CTC/LM architectures for general speech recognition, a sweet spot for the AM weight (typically $\lambda \in [0.2, 0.5]$) is usually found to achieve the optimal Word Error Rate (WER). However, for the standalone MDD task, we hypothesize that a significantly higher AM weight is crucial for preserving detection sensitivity. To validate this, we

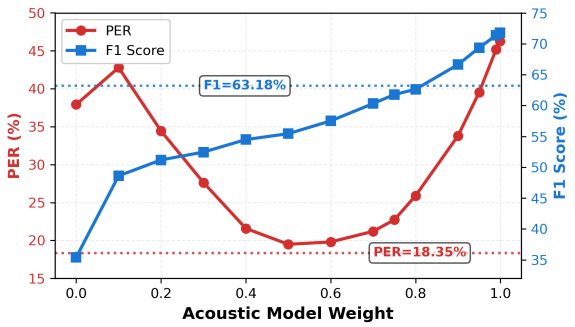

Figure 6: *Trade-off between PER and F1 on the L2-arctic over varying w_{AM} .*

Table 5: *Fine-tuned and zero-shot results on SO762 and ERJ.*

Dataset	Cond.	Model	F1 ↑	P ↑	R ↑	FRR ↓	FAR ↓	EDR ↓
SO762	Fine-tuned	Ryu2023 [67]	41.50	26.90	**91.60**	-	-	-
		JAM [65]	45.01	61.10	34.76	**0.58**	64.32	45.23
		MuFFIN [66]	**67.98**	**67.60**	68.37	1.01	31.63	58.82
		CTC-IF	46.68	31.86	87.23	6.32	12.77	50.30
		CROTTC	51.86	36.81	87.67	5.09	12.32	57.40
		CROTTC-IF	57.16	41.90	89.92	8.63	**6.53**	**10.08**
ERJ	Fine-tuned	CTC-IF	83.98	83.91	84.06	9.70	15.94	26.60
		CROTTC	85.79	85.79	85.79	8.63	14.21	29.09
		CROTTC-IF	**89.27**	**89.12**	**89.43**	**6.63**	**10.57**	**25.78**
	Zero-shot	CTC-IF	67.13	79.57	58.06	-	41.94	54.03
		CROTTC	69.17	83.37	59.10	7.16	40.90	46.92
		CROTTC-IF	**78.44**	**87.26**	**71.23**	**6.32**	**28.77**	**44.77**

conducted a spectrum-control study on the AM weight λ.

Figure 6 clearly demonstrates this divergence: while the PER reaches its optimum around $\lambda = 0.5$, the F1 score exhibits a completely different trajectory.[6] When $\lambda = 0$, the F1 drops to 36.83%, proving that relying solely on the LM's contextual priors leads to mispronunciation detection failure. As λ increases, F1 shows a monotonically increasing trend, confirming our hypothesis that acoustic fidelity is fundamentally more critical than linguistic priors in MDD. Notably, when $\lambda > 0.8$, the joint decoding achieves superior performance compared to CROTTC AM alone, demonstrating that IF-MDD provides effective and complementary soft linguistic guidance.

6.5. Generalization Study on SO762 and ERJ

Table 5 presents our experimental results on two additional L2 datasets: SO762 and ERJ. For the zero-shot experiments, we directly evaluated the models trained on L2-ARCTIC (as described in Section 6.4), and further adapted to the mentioned datasets for the fine-tuned settings. Compared to conventional baselines on SO762, CROTTC-IF achieves a highly competitive F1 score of 57.16% with significantly better FAR and EDR, without requiring the explicit fine-grained phonetic score labels as used in [65,66]. Furthermore, CROTTC-IF delivers outstanding fine-tuned and remarkably strong zero-shot results on ERJ, clearly demonstrating the robust out-of-domain (OOD) generalization of our prompt-free architecture.

7. Iqra'Eval2 Challenge

The Iqra'Eval2 Challenge [2] provides an ideal arena for our framework. Unlike standard L2 English benchmarks, Qur'anic recitation is governed by *Tajweed* rules[7], demanding precise modeling of nuanced articulatory deviations. Based on the empirical findings from our L2 English experiments, we deployed our CROTTC-IF architecture for this challenge.

Table 6: *Overview of the Iqra'Eval2 datasets.*

Dataset	Volume	Error Type	Description
`Iqra`	∼79h	None	Native speakers; golden reference
`TTS`	∼80h	Synthetic	TTS-injected mispronunciations
`Extra`	∼2h	Real errors	Real-world human mispronunciations

Table 7: *Performance of CROTTC-IF on the Iqra'Eval2 leaderboard.*

Model	F1↑	Pre. ↑	Rec.↑	PER ↓
3rd-team	71.57	67.69	75.93	4.05
CROTTC	68.77	70.07	67.52	4.11
w/ IF ($w_{\text{AM}} = 0.3$)	70.72	72.67	68.89	3.82
w/ IF ($w_{\text{AM}} = 0.9$)	**71.70**	**73.25**	**70.20**	**3.72**
1st-team	72.01	74.16	69.98	3.65

Table 6 summarizes the challenge's datasets. Among the provided corpora, we exclusively utilized the `TTS` and `extra` datasets. Specifically, we applied the `TTS` corpus for the AM's baseline training, followed by fine-tuning on the `extra` corpus, which contains genuine human pronunciation errors. For the `extra` corpus, we randomly allocated 10% of the utterances as a local test set, with the remainder used for training and validation. For vocabulary preparation, we followed [68], utilizing an inventory of 67 Arabic phonemes[8] plus 4 special tokens (`<bos>`, `<eos>`, `<sil>`, and `<blank>`), resulting in a vocabulary size of 71. Other training configurations remain consistent with those detailed in Sec 6.2. During inference, we applied a high acoustic weight ($w_{AM} = 0.9$) during joint decoding as discussed in Sec 6.4.4. As shown in Table 7, this principled approach yielded outstanding results. Operating entirely without explicit canonical prompts, our CROTTC-IF system ranked 2nd on the official Iqra'Eval2 leaderboard, achieving a highly competitive F1-score of 71.70% and a remarkably low PER of 3.72%. These results further demonstrate our core theory: decoupling acoustic modeling from explicit canonical priors can provide a robust and objective paradigm for real-world MDD tasks.

8. Conclusion and Future Work

In this paper, we proposed **CROTTC-IF**, a canonical prompt-free framework designed to overcome the inherent acoustic and linguistic bottlenecks in MDD. By integrating a dedicated acoustic frontend with an implicit knowledge transfer language model, our architecture achieves robust, state-of-the-art performance across diverse scenarios. Furthermore, through comprehensive LLM prompting experiments, we empirically demonstrated that uncritically injecting explicit canonical priors severely compromises acoustic fidelity. This highlights the critical need for careful and principled designs when incorporating canonical information. By successfully decoupling acoustic modeling from explicit textual priors, we believe this prompt-free paradigm establishes a highly objective and robust foundation for the MDD field. Future work will focus on further optimizing the detection-recognition trade-off and extending the framework to other CAPT applications.

[6] In MDD, PER is considered secondary to F1 and FRR [9,22]. High PER often reflects redundant insertions that do not degrade diagnostic accuracy, as evidenced by our robust COR at 92.42%.

[7] A strict set of phonetic rules governing Qur'anic recitation, distinct from Modern Standard Arabic (MSA).

[8] We omitted "<<", the geminated glottal stop, as it is not represented in our selected training data.

9. Generative AI Use Disclosure

Generative AI technology was employed strictly for minor grammatical corrections and stylistic improvements during the drafting process. The authors have verified all content and retain full responsibility for the accuracy and originality of this work.

10. References

[1] Y. El Kheir, A. Meghanani, H. O. Toyin *et al.*, "Iqra'eval: A shared task on qur'anic pronunciation assessment," in *ArabicNLP*, 2025, pp. 443–452.

[2] Y. E. Kheir, O. Ibrahim, A. Meghanani *et al.*, "Towards a unified benchmark for arabic pronunciation assessment: Quranic recitation as case study," 2025.

[3] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[4] A. M. Harrison, W.-K. Lo, X. Qian *et al.*, "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training." in *SLaTE*, 2009, pp. 45–48.

[5] K. Li, X. Qian, and H. Meng, "Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks," *IEEE/ACM TASLP*, vol. 25, no. 1, pp. 193–207, 2016.

[6] A. Graves, S. Fernández, F. Gomez *et al.*, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006, pp. 369–376.

[7] S. Watanabe, T. Hori, S. Kim *et al.*, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE JSTSP*, vol. 11, no. 8, pp. 1240–1253, 2017.

[8] A. Graves, "Sequence transduction with recurrent neural networks," *ICML*, 2012.

[9] W.-K. Leung, X. Liu, and H. Meng, "CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis," in *ICASSP*, 2019, pp. 8132–8136.

[10] A. Baevski, H. Zhou, A. Mohamed *et al.*, "wav2vec 2.0: a framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.

[11] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai *et al.*, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM TASLP*, vol. 29, pp. 3451–3460, 2021.

[12] M. Z. Boito, V. Iyer, N. Lagos *et al.*, "mHuBERT-147: A Compact Multilingual HuBERT Model," in *Interspeech 2024*, 2024.

[13] S. Chen, C. Wang, Z. Chen *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE JSTSP*, vol. 16, no. 6, pp. 1505–1518, 2022.

[14] L. Peng, K. Fu, B. Lin *et al.*, "A study on fine-tuning wav2vec2. 0 model for the task of mispronunciation detection and diagnosis." in *Interspeech*, 2021, pp. 4448–4452.

[15] M. Yang, K. Hirschi, S. D. Looney *et al.*, "Improving mispronunciation detection with wav2vec2-based momentum pseudo-labeling for accentedness and intelligibility assessment," in *Interspeech*, 2022, pp. 4481–4485.

[16] Y. EL Kheir, S. Chowdhury, and A. Ali, "Multi-view multi-task representation learning for mispronunciation detection," in *SLaTE 2023*, 2023, pp. 86–90.

[17] M.-S. Lin, B.-C. Yan, T.-H. Lo *et al.*, "Pg-mdd: Prompt-guided mispronunciation detection and diagnosis leveraging articulatory features," in *APSIPA ASC*. IEEE, 2024, pp. 1–6.

[18] A. Ibrahim, "Hafs2Vec: A system for the IqraEval Arabic and qur'anic phoneme-level pronunciation assessment," in *ArabicNLP*, Nov. 2025, pp. 453–456.

[19] A. Mansour, "Metapseud at iqra'eval: Domain adaptation with multi-stage fine-tuning for phoneme-level qur'anic mispronunciation detection," in *ArabicNLP*, Nov. 2025, pp. 475–479.

[20] Y. Feng, G. Fu, Q. Chen *et al.*, "Sed-mdd: Towards sentence dependent end-to-end mispronunciation detection and diagnosis," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 3492–3496.

[21] B. Lin and L. Wang, "Phoneme mispronunciation detection by jointly learning to align," in *ICASSP*, 2022, pp. 6822–6826.

[22] H. T. Tu, H. V. Khanh, T. T. Dat *et al.*, "Mispronunciation detection and diagnosis without model training: A retrieval-based approach," *arXiv preprint arXiv:2511.20107*, 2025.

[23] W. Ye, S. Mao, F. Soong *et al.*, "An approach to mispronunciation detection and diagnosis with acoustic, phonetic and linguistic (apl) embeddings," in *ICASSP*. IEEE, 2022, pp. 6827–6831.

[24] L. Peng, Y. Gao, B. Lin *et al.*, "Text-aware end-to-end mispronunciation detection and diagnosis," *arXiv preprint arXiv:2206.07289*, 2022.

[25] N. Zheng, L. Deng, W. Huang *et al.*, "Coca-mdd: A coupled cross-attention based framework for streaming mispronunciation detection and diagnosis," in *Interspeech 2022*, 2022, pp. 4352–4356.

[26] B.-C. Yan, H.-W. Wang, Y.-C. Wang *et al.*, "Effective graph-based modeling of articulation traits for mispronunciation detection and diagnosis," in *ICASSP*. IEEE, 2023, pp. 1–5.

[27] B.-C. Yan, H.-W. Wang, and B. Chen, "Peppanet: Effective mispronunciation detection and diagnosis leveraging phonetic, phonological, and acoustic cues," in *IEEE SLT*, 2022, pp. 1045–1051.

[28] Z. Ma, G. Yang, W. Chen *et al.*, "Slam-llm: A modular, open-source multimodal large language model framework and best practice for speech, language, audio and music processing," *Proc. IEEE Journal of Selected Topics in Signal Processing*, 2026.

[29] Y. Chu, J. Xu, Q. Yang *et al.*, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[30] M. Wu, J. Xu, X. Wu *et al.*, "Prompting large language models with mispronunciation detection and diagnosis abilities," in *Interspeech*, 2024, pp. 2990–2994.

[31] M. Wu, J. Xu, X. Chen *et al.*, "Integrating potential pronunciations for enhanced mispronunciation detection and diagnosis ability in LLMs," in *ICASSP*. IEEE, 2025, pp. 1–5.

[32] Z. Song, Z. Kadeer, M. Kahaer *et al.*, "Phoneme-controlled llm with self-supervised speech prompts for mispronunciation detection," *ACMMM Asia*, 2025.

[33] H. Zhong, Y. Xie, and Z. Yao, "Leveraging large language models to refine automatic feedback generation at articulatory level in computer aided pronunciation training." in *INTERSPEECH*, 2024.

[34] Y. Xie, H. Zhong, X. Lan *et al.*, "Mispronunciation detection and diagnosis based on large language models," *Computer Speech & Language*, p. 101942, 2026.

[35] N. Minematsu, C. Zhu, G. Dangtran *et al.*, "Development of shadowing speech corpora to measure instantaneous intelligibility as sequential annotation on L2 speech," in *Tech. Rep. Speech, Acoust. Soc. Jpn*, 2022, pp. 7–12.

[36] H. Geng, D. Saito, and N. Minematsu, "A pilot study of applying sequence-to-sequence voice conversion to evaluate the intelligibility of l2 speech using a native speaker's shadowings," in *Proc. APSIPA ASC*, 2024, pp. 1–6.

[37] ——, "A perception-based l2 speech intelligibility indicator: Leveraging a rater's shadowing and sequence-to-sequence voice conversion," in *Proc. Interspeech 2025*, 2025, pp. 2420–2424.

[38] A. Zeyer, R. Schlüter, and H. Ney, "Why does CTC result in peaky behavior?" *arXiv preprint arXiv:2105.14849*, 2021.

[39] J. Tian, B. Yan, J. Yu *et al.*, "Bayes risk ctc: Controllable ctc alignment in sequence-to-sequence tasks," *arXiv preprint arXiv:2210.07499*, 2022.

[40] Z. Yao, W. Kang, F. Kuang *et al.*, "Delay-penalized ctc implemented based on finite state transducer," in *Proc. Interspeech 2023*, 2023, pp. 1329–1333.

[41] Y. Kaloga, S. Kumar, P. Motlicek *et al.*, "A differentiable alignment framework for sequence-to-sequence modeling via optimal transport," *arXiv preprint arXiv:2502.01588*, 2025.

[42] Z. Yao, W. Kang, X. Yang *et al.*, "CR-CTC: Consistency regularization on ctc for improved speech recognition," in *ICLR*, 2024.

[43] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural networks*, vol. 22, no. 5-6, pp. 544–557, 2009.

[44] V. Vapnik and R. Izmailov, "Learning using privileged information: similarity control and knowledge transfer," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 2023–2049, 2015.

[45] T. Fukuda and S. Thomas, "Implicit transfer of privileged acoustic information in a generalized knowledge distillation framework." in *Interspeech*, 2020, pp. 41–45.

[46] T. Fukuda, M. Suzuki, G. Kurata *et al.*, "Efficient knowledge distillation from an ensemble of teachers." in *Interspeech*, 2017, pp. 3697–3701.

[47] X. Lu, P. Shen, Y. Tsao *et al.*, "Temporal order preserved optimal transport-based cross-modal knowledge transfer learning for asr," in *2024 SLT*, 2024, pp. 1–8.

[48] D. Lopez-Paz, L. Bottou, B. Schölkopf *et al.*, "Unifying distillation and privileged information," *ICLR*, 2016.

[49] K. Wang, G. Ortiz-Jimenez, R. Jenatton *et al.*, "Pi-dual: using privileged information to distinguish clean from noisy labels," in *ICML*, 2024.

[50] G. Ortiz-Jimenez, M. Collier, A. Nawalgaria *et al.*, "When does privileged information explain away label noise?" in *ICML*, 2023, pp. 26 646–26 669.

[51] J. Su, M. Ahmed, Y. Lu *et al.*, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127063, 2024.

[52] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *ICASSP*. IEEE, 2018, pp. 4784–4788.

[53] E. J. Hu, Y. Shen, P. Wallis *et al.*, "Lora: Low-rank adaptation of large language models." *Iclr*, vol. 1, no. 2, p. 3, 2022.

[54] A. Gulati, J. Qin, C.-C. Chiu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech*, 2020, pp. 5036–5040.

[55] A. Grattafiori, A. Dubey, A. Jauhri *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[56] Q. Team, "Qwen3 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2505.09388

[57] G. Zhao, S. Sonsaat, A. Silpachai *et al.*, "L2-arctic: A non-native english speech corpus," in *Interspeech*, 2018, p. 2783–2787.

[58] J. Zhang, Z. Zhang, Y. Wang *et al.*, "speechocean762: An open-source non-native english speech corpus for pronunciation assessment," in *Proc. Interspeech 2021*, 2021.

[59] S. D. C. of the Priority Areas Project, "Ume english speech database read by japanese students (ume-erj)," jun 2007.

[60] T. Makino and R. Aoki, "English read by japanese phonetic corpus: An interim report," *Research in Language*, vol. 9, no. 2, pp. 79–95, 2012.

[61] K. Li, X. Qian, and H. Meng, "Mispronunciation detection and diagnosis in L2 english speech using multidistribution deep neural networks," *IEEE/ACM TASLP*, vol. 25, no. 1, pp. 193–207, 2017.

[62] D. Y. Zhang, S. Saha, and S. Campbell, "Phonetic rnn-transducer for mispronunciation diagnosis," in *ICASSP*, 2023, pp. 1–5.

[63] Y. Wan, Y. Shi, B. Lin *et al.*, "A study of mispronunciation detection and diagnosis based on meta-learning," in *ICASSP*, 2024.

[64] C. Zhu, A. Wumaier, D. Wei *et al.*, "Pronunciation error detection model based on feature fusion," *Speech Communication*, vol. 156, p. 103009, 2024.

[65] Y.-Y. He, B.-C. Yan, T.-H. Lo *et al.*, "JAM: A unified neural architecture for joint multi-granularity pronunciation assessment and phone-level mispronunciation detection and diagnosis towards a comprehensive capt system," in *2024 APSIPA ASC*, 2024, pp. 1–6.

[66] B.-C. Yan, M.-K. Tsai, and B. Chen, "Muffin: Multifaceted pronunciation feedback model with interactive hierarchical neural modeling," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 4295–4310, 2025.

[67] H. Ryu, S. Kim, and M. Chung, "A Joint Model for Pronunciation Assessment and Mispronunciation Detection and Diagnosis with Multi-task Learning," in *Interspeech 2023*, 2023, pp. 959–963.

[68] N. Halabi and M. Wald, "Phonetic inventory for an Arabic speech corpus," in *LREC*, May 2016, pp. 734–738.