IF-MDD: INDIRECT FUSION FOR PROMPT-FREE MISPRONUNCIATION DETECTION AND DIAGNOSIS

Haopeng Geng, Daisuke Saito, Nobuaki Minematsu

Graduate School of Engineering, The University of Tokyo

ABSTRACT

Mispronunciation detection and diagnosis (MDD) plays a vital role in computer-assisted language learning (CALL). Although recent approaches have achieved promising performance by leveraging canonical phonemes as auxiliary inputs, this reliance constrains their applicability in spontaneous language learning scenarios. In this work, we propose IF-MDD, an indirect fusion model that integrates canonical phoneme and error-related information as privileged information during training while obviating the requirement of text-prompting at inference. Despite being trained on limited data, IF-MDD achieved competitive diagnostic performance, reaching an F1 score of 60.67% and an error diagnosis rate of 19.98% on the L2-ARCTIC. Furthermore, experiments show that IF-MDD generalizes reliably to unseen speakers with diverse L1 backgrounds. These findings underscore the potential of IF-MDD as a scalable and practical solution for language learners. Our implementation[1] and demo[2] are publicly available.

Index Terms— Mispronunciation detection and diagnosis, computer-assisted language learning, learning using privileged information, self-supervised speech representation, multi-task learning.

1. INTRODUCTION

Mispronunciation detection and diagnosis (MDD) aims to automatically identify pronunciation errors made by second-language (L2) learners and provide phoneme-level feedback that indicates where and how their pronunciations deviate from canonical forms. Conventional approaches typically leverage ASR-based likelihoods or goodness-of-pronunciation (GOP) scores [1], with extended recognition networks (ERNs) further modeling error patterns for phoneme-level classification [2, 3].

Benefiting from the end-to-end structure, recent MDD approaches can be directly optimized as an independent task. Modern MDD broadly falls into two categories: dictation-style and text-prompting style. Dictation-style methods aim to recognize the uttered phoneme sequence exclusively from acoustic features. For example, [4] introduced a connectionist temporal classification (CTC) framework and demonstrated the feasibility of capturing L2 phoneme information from acoustic features alone. More recent studies have leveraged self-supervised learning (SSL) speech representations, such as wav2vec 2.0 [5] and WavLM [6], which provide robust phonemic context features and have proven effective for both detection and diagnosis. For instance, [7] employed a fine-tuned wav2vec 2.0 model with a CTC decoder, achieving promising results on diagnosti c accuracy. To address the scarcity of expert-annotated data, a semi-supervised training method using pseudo-labeling was also applied and yielded further improvements in diagnostic performance [8]. In addition to acoustic features alone, auxiliary information implicitly predicted from acoustic signals has also been developed to strengthen free phoneme recognition. For example, [9] proposed a dual-path SSL framework that focused on differences between mono- and multilingual representations, while incorporating a detection task to predict manner-of-articulation from the acoustic signal. Similarly, [10] leveraged an electromagnetic articulography (EMA) dataset to jointly train an acoustic-to-articulatory inversion (AAI) model with the acoustic encoder, which also yielded promising improvements.

Meanwhile, canonical phoneme sequences are usually available in reading-aloud MDD scenarios. Therefore, it is natural to integrate this prior knowledge in MDD modeling. For example, Zheng et al. proposed a coupled-cross attention mechanism to fuse acoustic and canonical phoneme embeddings, then thresholded the phonemes with low likelihood as mispronunciations [11]. Yan et al. observed that errors often occur among similar phonemes (e.g., /t/, /d/, /th/) and therefore incorporated a phoneme lookup table or a graph-based network alongside the canonical phonemes [12, 13] . More recently, Wu et al. prompted large language models (LLMs) with canonical phonemes and candidate errors [14, 15]. In general, text-prompting methods outperform dictation-style approaches.

However, despite the promising performance, the text-prompting methods face two key limitations. First, the requirement for canonical phoneme sequences at inference is impractical in spontaneous or reading-after (e.g., shadowing [16]) scenarios. Second, as mispronunciation segments are usually sparse and correctly pronounced segments are acoustically close to the canonical feature, simplistically fusing acoustic and canonical representations biases the embedding feature toward the canonical pattern, causing the model to overlook mispronunciation cues. As a result, the system may become overly tolerant of atypical pronunciations, which limits its diagnostic performance, especially for elementary L2 learners.

To address these concerns, in this work, we propose an indirect fusion framework, IF-MDD, to adequately integrate the canonical information with acoustic feature. Our contributions are summarized as follows:

- **Framework design**: We introduce a novel MDD framework that leverages canonical cues as privileged information through indirect fusion during the training phase only, thereby enabling a text-prompt-free system at inference.
- **Performance**: IF-MDD achieves state-of-the-art diagnostic performance among prompt-free methods, while maintaining competitive recognition and diagnostic accuracy compared to leading text-prompting approaches.
- **Generalizability**: Beyond same-L1 test conditions, IF-MDD

[1] https://github.com/Secondtonumb/IF-MDD

[2] https://secondtonumb.github.io/publication_demo/ICASSP_2026/

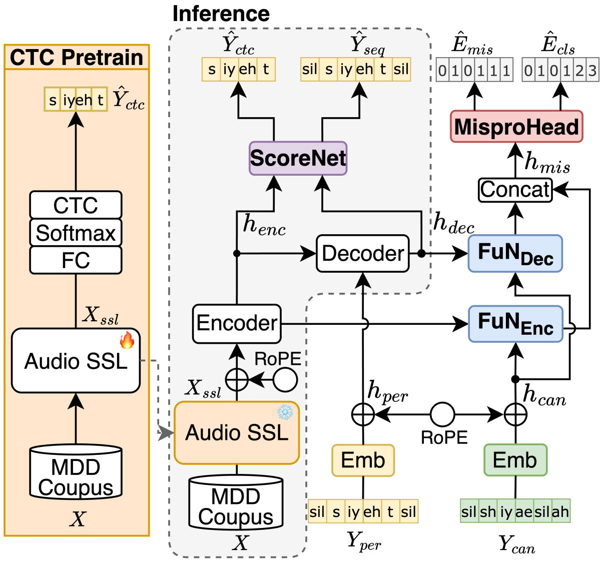

Fig. 1: Overview of IF-MDD

demonstrates robustness to out-of-domain L2 speakers, highlighting its strong generalizability and practical applicability.

2. PROPOSED METHOD

2.1. Learning using privileged information

Learning using privileged information (LUPI) is a training paradigm that leverages additional information available only during the training phase to implicitly transfer prior knowledge to the model [17]. Such strategies have been shown to be effective across various machine learning tasks particularly when training data is noisy or labels are scarce [18, 19]. A related and well-known approach is knowledge distillation [20], where a student model learns to approximate teacher outputs via a distillation loss, thereby implicitly transferring teacher knowledge to the student model. This paradigm has also been widely applied and validated in speech recognition tasks [21, 22].

2.2. IF-MDD model

We adopt the LUPI strategy to mitigate the text-prompt leakage problem in MDD. Specifically, we treat canonical phonemes and mispronunciation cues as privileged information available only during training, so that the canonical and mispronunciation cues guide the hidden representation via backpropagation. As illustrated in Fig. 1, our IF-MDD consists of five components: 1) a pre-trained speech SSL feature extractor; 2) an encoder–decoder backbone; 3) two feature-fusion networks; 4) two mispronunciation detection heads; and 5) a ScoreNet for final decoding.

2.2.1. Pre-trained speech SSL feature extractor

Instead of using frozen SSL outputs, we fine-tuned WavLM-Large[3] on L2 speech with a shallow fully connected projection and a CTC objective. The adapted model serves as the front end of IF-MDD, providing L2-specific phonetic representations and coarse segmentation evidences to downstream modules.

2.2.2. Encoder-decoder backbone

To better predict the perceived L2 phonemes, we employ an encoder–decoder architecture as IF-MDD's backbone. Let the input waveform $X = [x^0 \dots x^{T_s-1}]$ and the perceived phoneme sequence $Y_{\text{per}} = [y^0_{per} \dots y^{T_p-1}_{per}]$, where T_s and T_p denote the lengths of speech signal and phoneme sequence, respectively. The encoder-decoder backbone is defined as:

$$\mathbf{X}_{\text{ssl}} = \text{SSL}(\mathbf{X}), \tag{1}$$

$$\mathbf{h}_{\text{per}} = \text{RoPE}(\mathbf{Y}_{\text{per}}), \tag{2}$$

$$\mathbf{h}_{\text{enc}} = \text{Enc}\big(\text{RoPE}(\mathbf{X}_{\text{ssl}})\big), \tag{3}$$

$$\mathbf{h}_{\text{dec}} = \text{Dec}\big(\mathbf{h}_{\text{per}},\, \mathbf{h}_{\text{enc}}\big), \tag{4}$$

where Enc$(\cdot)$ is a 2-layer Conformer encoder and Dec$(\cdot)$ is a 2-layer Transformer decoder [24]. Rotary position embeddings RoPE$(\cdot)$ are applied to both $\mathbf{h}_{\text{enc}} \in \mathbb{R}^{T_s \times D}$ and $\mathbf{h}_{\text{dec}} \in \mathbb{R}^{T_p \times D}$ to enhance positional modeling [25]. Note that the SSL$(\cdot)$ is kept frozen here to ensure stable training.

2.2.3. Fusion Network

To explicitly model the mispronunciation information that is indicated by the misalignment between acoustic and canonical features, we introduced two auxiliary fusion networks (FuN). Given the canonical phoneme sequence $Y_{\text{can}} = [y^0_{can} \dots y^{T_p-1}_{can}]$, the fusion networks are defined as:

$$\mathbf{h}_{\text{can}} = \text{RoPE}(Y_{\text{can}}) \tag{5}$$

$$\mathbf{h}^{\text{enc}}_{\text{mis}} = \text{FuN}_{\text{enc}}\big(\mathbf{h}_{\text{can}},\, \text{DS}(\mathbf{h}_{\text{enc}})\big), \tag{6}$$

$$\mathbf{h}^{\text{dec}}_{\text{mis}} = \text{FuN}_{\text{dec}}\big(\mathbf{h}_{\text{can}},\, \mathbf{h}_{\text{dec}}\big), \tag{7}$$

$$\mathbf{h}_{\text{mis}} = \text{Concat}\big(\mathbf{h}^{\text{enc}}_{\text{mis}},\, \mathbf{h}^{\text{dec}}_{\text{mis}}\big) \tag{8}$$

where $\mathbf{h}_{\text{mis}} \in \mathbb{R}^{T_p \times D}$ denotes the fused representation. Here we employed two parallel branches: an encoder-side FuN$_{\text{enc}}(\cdot)$ and a decoder-side FuN$_{\text{dec}}(\cdot)$, each implemented with stacked Transformer-decoder blocks to model multi-aspect feature interactions in $(\mathbf{h}_{\text{can}}, \mathbf{h}_{\text{enc}})$ and $(\mathbf{h}_{\text{can}}, \mathbf{h}_{\text{dec}})$, respectively. In both branches, FuN$(\cdot)$ uses $\mathbf{h}_{\text{can}}$ as the query and $\mathbf{h}_{\text{enc}}$ or $\mathbf{h}_{\text{dec}}$ as the memory. In particular, a 1D CNN downsampler DS$(\cdot)$ is applied within FuN$_{\text{enc}}(\cdot)$, which we found improves alignment in our experiments.

2.2.4. Mispronunciation-detection head

Prior works often relied on a single binary error indicator to capture mismatches between acoustic and canonical cues [12,26]. However, we argue that a more comprehensive diagnosis should not only detect mispronunciation positions but also identify their error types. Thus, we extend the mispronunciations head with an additional error-type head. Given the fused mispronunciation representation $\mathbf{h}_{\text{mis}}$, we apply a shared CNN trunk followed by two task-specific projection heads:

$$\mathbf{u} = \text{CNN}(\mathbf{h}_{\text{mis}}), \tag{9}$$

$$\hat{E}_{\text{mis}} = \sigma\big(\text{CNN}_{\text{bin}}(\mathbf{u})\big), \tag{10}$$

$$\hat{E}_{\text{cls}} = \text{Softmax}\big(\text{Linear}_{\text{cls}}(\mathbf{u})\big). \tag{11}$$

where $\hat{E}_{\text{mis}} \in \{0, 1\}^{T_p}$ denotes an error position indicator, and $\hat{E}_{\text{cls}} \in \{0, 1, 2, 3\}^{T_p}$ denotes the distribution over $\{$*correct*, *substitution*, *deletion*, *insertion*$\}$.

[3] https://huggingface.co/microsoft/wavlm-large

Table 1: MDD results of conventional methods and IF-MDD. Methods are categorized by whether they use extra data or text prompts at inference. Metrics with ↑ are *higher-is-better* and ↓ are *lower-is-better*. All metrics are reported as percentages.

Models	Extra Data	Text Prompt	**Recognition Accuracy**		**Diagnostic Accuracy**		**Detection Metrics**			PER (↓)
			FRR (↓)	FAR (↓)	CDR (↑)	EDR (↓)	P (↑)	R (↑)	F1 (↑)	
MDDGCN [12]	Yes	Yes	9.18	38.03	74.76	25.27	51.90	**61.97**	56.49	-
Peppanet [13]	Yes	Yes	9.61	**36.47**	74.12	25.88	**64.53**	51.38	56.81	-
PG-MDD [10]	Yes	Yes	6.65	39.94	76.87	23.13	60.06	60.15	60.10	13.92
MPL-MDD [8]	Yes	No	5.60	48.80	77.29	22.71	60.39	51.20	55.42	14.36
RNN-T [23]	Yes	No	**5.30**	44.70	-	-	63.40	55.30	59.10	15.47
w2v2-CTC [7]	Yes	No	5.70	41.80	70.72	29.28	62.86	58.57	60.44	16.20
MV-w2v2 [9]	No	No	-	-	-	-	59.23	61.43	60.31	14.13
IF-MDD-CTC (Ours)	No	No	6.12	40.64	78.43	21.57	62.03	59.36	**60.67**	14.87
IF-MDD-Seq (Ours)	No	No	5.75	44.71	**80.02**	**19.98**	61.81	55.29	58.37	**13.72**

2.2.5. *ScoreNet*

During inference, the model has no access to Y_{can} or Y_{per}. To combine the complementary cues from the CTC and decoder branches, we employ a ScoreNet module under the hybrid CTC/Seq2Seq scheme [27]. Specifically, with the decoder's autoregressive decoding denoted as $\text{Dec}^{\text{AR}}(\cdot)$, two hypotheses are generated:

$$\mathbf{p}_{\text{ctc}} = \text{Softmax}(\text{FC}(\mathbf{h}_{\text{enc}})), \tag{12}$$

$$\hat{Y}_{\text{ctc}} = \text{CTC-Decode}(\mathbf{p}_{\text{ctc}}), \tag{13}$$

$$\hat{Y}_{\text{seq}} = \text{BeamSearch}(\text{Dec}^{\text{AR}}(\mathbf{h}_{\text{enc}}),\ \mathbf{p}_{\text{ctc}}), \tag{14}$$

where $\hat{Y}_{\text{ctc}}$ is produced by a non-autoregressive CTC decoder, and $\hat{Y}_{\text{seq}}$ is generated via beam search augmented with CTC posteriors $\mathbf{p}_{\text{ctc}}$. Thus, $\hat{Y}_{\text{ctc}}$ reflects indirect fusion from acoustic features only, whereas $\hat{Y}_{\text{seq}}$ incorporates both acoustic information and contextual cues from $\text{Dec}(\cdot)$.

2.3. Training criteria

We define the multi-task loss function as:

$$\mathcal{L}_{\text{MDD}} = \alpha\mathcal{L}_{\text{ctc}} + (1-\alpha)\mathcal{L}_{\text{seq}} \tag{15}$$

$$\mathcal{L}_{\text{err}} = \mathcal{L}_{\text{mis}} + \mathcal{L}_{\text{cls}} \tag{16}$$

where $\mathcal{L}_{\text{ctc}}$ is the CTC loss and L_{seq} is the cross-entropy loss on $\mathbf{h}_{\text{dec}}$ for phoneme prediction.[4] The focal loss $\mathcal{L}_{\text{mis}}$ and the cross-entropy loss $\mathcal{L}_{\text{cls}}$ correspond to mispronunciation detection and error-type classification, respectively. [28]. Additionally, to encourage monotonic alignments in $\text{FuN}(\cdot)$, we apply guided-attention loss $\mathcal{L}_{\text{ga}}$ [29]. This is crucial for $\mathbf{h}_{\text{mis}}$ to capture discontinuities that characterize mispronunciation cues. Thus, the overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{MDD}} + \beta\mathcal{L}_{\text{err}} + \gamma\mathcal{L}_{\text{ga}} \tag{17}$$

where hyperparameters α, β, and γ are set to 0.3, 1.0, and 10, respectively.

3. EXPERIMENTAL EVALUATION

3.1. Dataset

We used the L2-ARCTIC corpus [30] as the sole resource for building IF-MDD. Following prior work, six speakers with diverse L1 backgrounds (NJS, TLV, TNI, TXHC, YKWK, and ZHAA) were reserved for testing, yielding 900 utterances [8]. The remaining 2,647 mispronunciation-annotated recordings were used for training and validation. For phoneme clustering, we adopted a 41-unit ARPAbet-style inventory (39 phonemes + silence + 1 error class).

3.2. Evaluation metrics

We followed the hierarchical evaluation protocol of [31]. 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). The corresponding rates are defined as:

$$\text{FAR} = \frac{FA}{FA+TR}, \quad \text{FRR} = \frac{FR}{FR+TA}, \quad \text{EDR} = \frac{ED}{ED+CD}.$$

MDD performance was then summarized with the F1 score, treating TR as the positive class:

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

We also reported phoneme error rate (PER) as a measure of recognition accuracy:

$$\text{PER} = \frac{S+D+I}{N},$$

where S, D, and I denote the numbers of substitutions, deletions, and insertions, and N is the total number of perceived phonemes, with alignment artifacts removed but silence tokens retained.

3.3. Experimental evaluation and analysis

Table 1 compares IF-MDD against competitive baselines [7–10, 12, 13, 23], which are categorized by whether they used extra data or text-prompting during inference. IF-MDD-CTC and IF-MDD-Seq denote the two outputs described in Sec. 2.2.5.

Among all systems, IF-MDD-CTC achieves the highest overall detection quality with F1 = 60.67%, surpassing the best text-prompted baseline PG-MDD (60.10%). It also delivers a balanced P/R = 62.03%/59.36%, indicating fewer false alarms on correct phones and fewer misses on mispronounced phones, which is desirable for practical CAPT deployments. While slightly lower in F1 (58.37%), IF-MDD-Seq attains the lowest PER (13.72%) and EDR (19.98%) among all methods, indicating more consistent phoneme recognition and improved diagnosis feedback.

[4] Strictly speaking, the target for $\mathcal{L}_{\text{ctc}}$ is not Y_{per}, but a refined version with alignment artifacts removed, which is generally shorter than Y_{per}.

Table 2: Ablation Study on IF-MDD and fusion strategies.

Factors	FAR (↓)	EDR (↓)	P (↑)	R (↑)	F1 (↑)
IF-MDD	44.71	19.98	61.81	55.29	58.37
-w/o FuN$_{\mathrm{Enc}}$	45.35	21.74	60.29	54.65	57.33
-w/o FuN$_{\mathrm{Dec}}$	46.14	22.10	59.42	53.86	56.51
-w/o Error-Type Head	44.29	22.28	59.54	55.71	57.56
Direct Fusion	71.52	29.38	66.09	28.48	39.80
Reverse Fusion	7.84	83.32	15.90	92.16	27.10

Table 3: IF-MDD zero-shot and fine-tuned results on ERJ.

Cond.	Model	FAR(↓)	EDR(↓)	P(↑)	R(↑)	F1(↑)
Zero-shot	MPL-MDD	44.54	**51.25**	78.24	55.46	64.91
	IF-MDD-CTC	**39.51**	54.73	**79.68**	**60.49**	**68.77**
	IF-MDD-Seq	41.94	54.03	79.57	58.06	67.13
Fine-tuned	MPL-MDD	17.16	37.24	80.88	82.84	81.85
	IF-MDD-CTC	**11.82**	30.22	81.92	**87.18**	**84.47**
	IF-MDD-Seq	15.94	**26.60**	**83.91**	84.06	83.98

Compared to IF-MDD, earlier text-prompted methods (MDD-GCN [12], Peppanet [13]) exhibit relatively higher FRR. Since these systems embedded prior mispronunciation knowledge into the model, they tend to be overly sensitive to correct pronunciations, such that even minor deviations from the canonical sequence are flagged. Their unbalanced P/R further confirms this limitation. By contrast, PG-MDD [10] relaxed such priors and thus achieved a more balanced P/R and better F1, though it still falls short of IF-MDD-CTC.

As for prompt-free systems [7, 8, 23], these approaches often rely on additional training data (e.g. TIMIT [32]) to ensure adequate phoneme recognition accuracy. However, when extra L1 phoneme corpora are incorporated alongside L2 data without differentiation, the model's ability to capture L2-specific mispronunciation cues can be weakened. This results in higher FAR and degraded diagnostic accuracy. In contrast, IF-MDD achieves a favorable FAR/FRR balance, demonstrating the effectiveness of our indirect fusion strategy.

3.4. Ablation study for IF-MDD

To quantify the contribution of fusion network and error-type classifier in Sec. 2.2.3 and Sec 2.2.4, we conducted an ablation study and the results are summarized in Table 2. For brevity, we report IF-MDD-Seq only. Notably, removing either FuN$_{\mathrm{enc}}$ or FuN$_{\mathrm{dec}}$ leads to a clear drop in F1 by 1.03–1.85 points, demonstrating the effectiveness of multi-aspect monitoring in our indirect fusion. Ablating the error-type classification head yields a smaller but consistent decrease in F1 and increase in EDR, which suggests that this auxiliary task helps regularize the mispronunciation representation. Additionally, Fig. 2 shows the last layer cross-attention heat maps of the fusion network. Removing IF-MDD components leads to blurrier, less-peaky attention maps with reduced diagnostic power, indicating degraded representation quality.

To further demonstrate our indirect fusion strategy, we evaluated two alternative fusion settings. As illustrated in Fig. 3, in direct fusion, $\mathbf{h}_{\mathrm{enc}}$ is injected into $\mathbf{h}_{\mathrm{can}}$ whereas in reverse fusion $\mathbf{h}_{\mathrm{can}}$ is merged into $\mathbf{h}_{\mathrm{enc}}$. As shown in the last two rows in Table 2, direct fusion substantially increases FAR and reduces F1 despite relatively high precision. This indicates an over-reliance on the canonical sequence. Reverse fusion also yielded clear F1 degradation, as the injected canonical information acts as noise and thus hinders the modeling of L2-phoneme information. These results further confirm the effectiveness of our indirect fusion strategy.

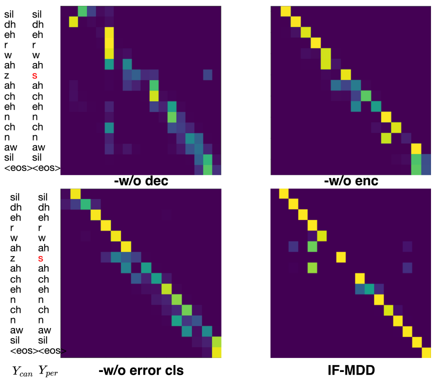

Fig. 2: Attention heatmaps of Fusion Network's under different ablation conditions; the x-axis denotes the memory and the y-axis denotes the canonical-phoneme embeddings.

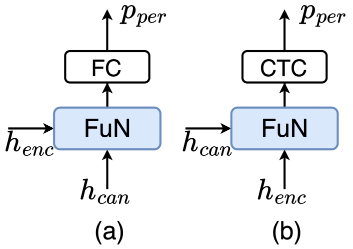

Fig. 3: Illustration of (a) direct fusion and (b) reverse fusion.

3.5. Generalization to out-of-domain speakers

A practical MDD system should be robust across diverse accents and speakers. To assess this, we evaluated IF-MDD on a Japanese L2 English corpus, ERJ [33], where Japanese speakers were not included in the training data for IF-MDD. Following [34], we converted the IPA-style mispronunciation annotations of 800 utterances into the ARPAbet inventory described in Sec. 3.1. We then randomly selected 50 utterances from 12 speakers of varying proficiency as the test set, with the remainder used for training and validation. We selected MPL-MDD [8] as our baseline and evaluated IF-MDD under two conditions: zero-shot and fine-tuned. As shown in Table 3, in the zero-shot setting, IF-MDD-CTC and IF-MDD-Seq achieved higher F1 scores than MPL-MDD, demonstrating stronger generalization to unseen speakers. After fine-tuning, all systems improved, and IF-MDD-CTC still attained the best overall performance (F1 = 84.47) and IF-MDD-Seq achieved the lowest EDR (26.60%), which is consistent with the trends observed in Table 1. These findings indicate that IF-MDD is effective for both zero-shot inference and limited-data domain adaptation, making it a practical solution for robust MDD across speakers and accents.

4. CONCLUSION AND FUTURE WORK

We presented IF-MDD, a prompt-free MDD framework that leverages canonical cues as training-only privileged information via indirect fusion, opening a new path to for prompt-free L2 speech assessment. In future work, we will strengthen canonical guidance without inducing over-reliance and explore semi-/self-supervised learning and unsupervised domain adaptation to mitigate the scarcity of L2 annotations.

5. REFERENCES

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

[2] 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.

[3] K. Li, X. Qian, 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.

[4] W.-K. Leung, X. Liu, H. Meng, “Cnn-rnn-ctc based end-to-end mispronunciation detection and diagnosis,” in *ICASSP*, 2019, pp. 8132–8136.

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

[6] 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.

[7] 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.

[8] 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.

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

[10] 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, pp. 1–6.

[11] 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.

[12] 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.

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

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

[15] 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.

[16] 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.

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

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

[19] G. Ortiz-Jimenez, M. Collier, A. Nawalgaria et al., “When does privileged information explain away label noise?,” in *ICML*, 2023, pp. 26646–26669.

[20] G. Hinton, O. Vinyals, J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.

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

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

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

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

[25] J. Su, M. Ahmed, Y. Lu et al., “Roformer: Enhanced transformer with rotary position embedding,” *Neurocomputing*, vol. 568, pp. 127063, 2024.

[26] Z. Lin, R. Takashima, D. Saito et al., “Shadowability annotation with fine granularity on l2 utterances and its improvement with native listeners’ script-shadowing.,” in *Interspeech*, 2020, pp. 3865–3869.

[27] 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.

[28] T.-Y. Lin, P. Goyal, R. Girshick et al., “Focal loss for dense object detection,” in *ICCV*, 2017, pp. 2999–3007.

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

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

[31] K. Li, X. Qian, 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.

[32] J. S. Garofolo, L. F. Lamel, W. M. Fisher et al., “Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1,” *NASA STI/Recon technical report n*, vol. 93, pp. 27403, 1993.

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

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