# PG-MDD: Prompt-Guided Mispronunciation Detection and Diagnosis Leveraging Articulatory Features

Meng-Shin Lin [1], Bi-Cheng Yan[1], Tien-Hong Lo[1], Hsin-Wei Wang[1], Yue-Yang He[1], Wei-Cheng Chao[2], Berlin Chen[1]

[1] National Taiwan Normal University, Taipei, Taiwan
[2] Telecommunication Laboratories, Chunghwa Telecom Co., Ltd., Taiwan
E-mail: {61147077s, bicheng, teinhonglo, hsinweiwang, yueyanghe, berlin}@ntnu.edu.tw; weicheng@cht.com.tw

***Abstract*—Mispronunciation detection and diagnosis (MDD) manages to pinpoint phonetic errors of L2 (second-language) learners and then provides timely and informative diagnosis on erroneous pronunciation segments. Recently, dictation-based neural methods have emerged as an appealing modeling paradigm for MDD, which simultaneously identifies pronunciation errors and provides diagnostic feedback by aligning the recognized phone sequence to the corresponding canonical phone sequence of a given text prompt. Despite their decent performance in terms of F1-score, dictation-based models still struggle to accurately detect pronunciation errors with balanced precision and recall evaluations, resulting in inferior learning efficiency for L2 learners. In view of this, we propose a novel prompt-guided dictation-based MDD model, dubbed PG-MDD, that can efficiently strike a balance the precision and recall rates while maintaining a high-performing F1-score. PG-MDD first jointly optimizes the mispronunciation detection and diagnosis processes during the training phase, while aptly guiding the diagnosis process with phone-dependent thresholds in the inference phase. In addition, a novel multi-view audio encoder is introduced to render the fine-grained articulatory cues within learners' speech. A comprehensive set of empirical experiments conducted on the L2-ARCTIC benchmark dataset suggests the practical feasibility of our method in relation to several competitive baselines.**

## I. Introduction

Globalization has increased the demand for learning languages like English, Spanish, and Mandarin, driving the development of computer-assisted pronunciation training (CAPT) systems. These systems allow L2 learners to practice pronunciation independently and stress-free [1][2], supplement teachers' instruction, address the shortage of qualified teachers [3], and serve as references for professionals in assessment tasks [4][5].

Mispronunciation detection and diagnosis (MDD) is a crucial part of CAPT systems. MDD aims to identify erroneous pronunciation segments and provide diagnostic feedback to L2 learners. It can assess pronunciation proficiency at both the supra-segmental (e.g., prosody and intonation) [6][7] and segmental levels (e.g., phone and word) [8][9]. Due to the difficulty of achieving high inter-rater agreement for supra-segmental assessment, research has mainly focused on segmental assessment. Efforts in phone-level MDD typically fall into three categories: pronunciation scoring-based, dictation-based, and prompt-based methods. Pronunciation scoring-based approaches relies on a well-trained speech recognition model to derive various types of confidence measurements as indicators of mispronunciation. Commonly used indicators include, but are not limited to, phone durations [10], likelihood ratios [11], and phone posterior probabilities [12]. To better obtain informative diagnostic feedback, dictation-based methods alternatively frame MDD as a phone recognition task by employing a free-phone recognition process to dictate the most likely phone sequence uttered by an L2 learner. The erroneous pronunciation portions are then easily identified by comparing the dictation result with the corresponding canonical phone sequence [13][14]. Among others, prompt-based methods extend the aforementioned methods by simplifying the forced-alignment process, which employs an attention mechanism to calculate the soft alignment between the canonical phone sequence and the L2 learner's input speech while simultaneously generating the diagnostic feedback [15][16]. Comparing these three mainstream paradigms, dictation-based

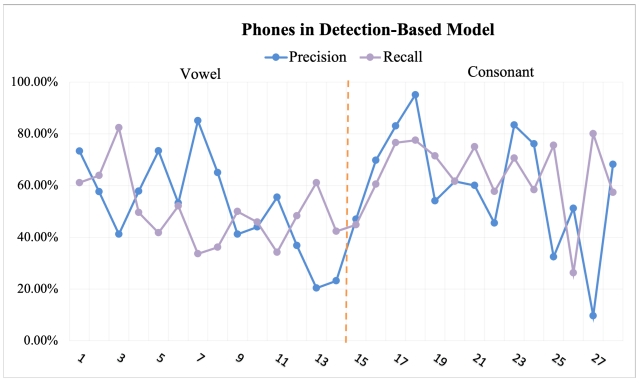
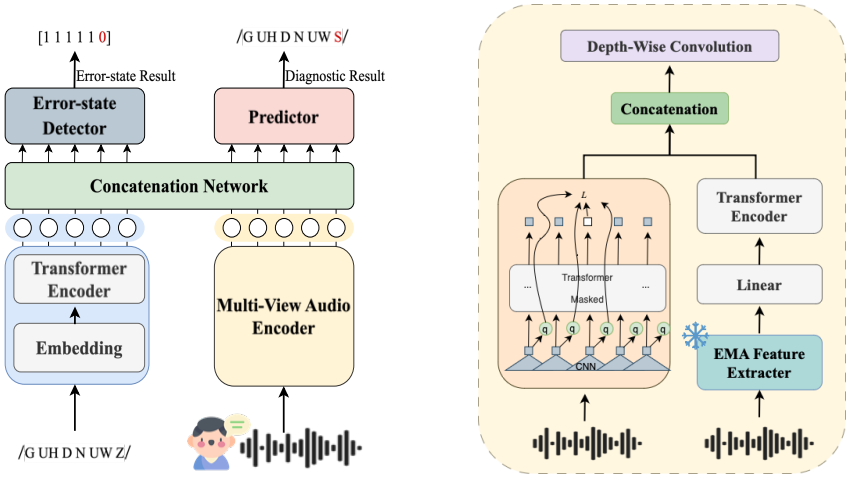



**Fig. 1**. Precision and recall distributions on different phones for a detection-based model. The orange dashed line indicates the division between vowels (on the left) and consonants (on the right)




(a) Overall Model Architecture (b) Multi-View Audio Encoder

**Fig. 2**. A schematic depiction of the model structure of our MDD method.

methods are considered to be the dominant ones, showing considerable promise [17][18].

Despite the effectiveness of existing dictation-based methods in terms of the F1-score metric, they are often faced with challenges in accurately detecting pronunciation errors with balanced precision and recall rates. As an illustration, we demonstrate the performance of mispronunciation detection for disparate phones with a dictation-based MDD model in Fig. 1 where most phones achieve only a precision of 60%, with the corresponding recall rates varying between 40% and 80% for disparate vowels and consonants. This result underscores the adverse impact that correctly pronounced phone segments are frequently misidentified as pronunciation errors by the dictation-based model, which might demotivate students, leading to lower confidence in the MDD system. What is more, failing to detect pronunciation errors potentially slows down the learning process. Building on these observations, this paper introduces a novel prompt-guided dictation-based MDD model, termed PG-MDD, with the goal of achieving a better balance between the precision and recall rates while maintaining a high-performing F1-score for the mispronunciation detection task. To aptly guide the diagnosis process of the proposed model, we first jointly optimize the mispronunciation detection and the diagnosis processes during the training phase. During the inference phase, PG-MDD determines the pronunciation errors in response to the canonical phone sequence of a text prompt using adjustable phone-dependent thresholds. Furthermore, a novel multi-view audio encoder is put forward to capture the fine-grained articulatory cues inherent in learners' speech. Comprehensive experiments conducted on the L2-ARCTIC benchmark dataset reveal that our proposed method achieves significant and consistent improvements over several strong baselines.

In summary, our contributions in this paper are at least three-fold: (1) a novel prompt-guided MDD model is proposed, dubbed PG-MDD, which can flexibly balance performance trade-offs between precision and recall rates using adjustable phone-dependent thresholds; (2) PG-MDD is able to authentically dictate the input speech of an L2 learners, benefiting from cross-modality attention between the input speech and the text prompt, which enables PG-MDD to capture the both inner- and inter-dependencies between the learner's speech and the corresponding canonical phone sequence; and (3) a novel multi-view audio encoder is introduced to exploit the fine-grained articulatory cues within learners' speech.

## II. Proposed Method

### A. Problem Definition

The proposed PG-MDD tackles the task of MDD through the following processing flow. Given an $T$-length input audio signal $\mathbf{x} = (x_1, x_2, \ldots, x_T)$ uttered by an L2 learner and an $N$-length canonical phone sequence $\mathbf{q} = (q_1, q_2, \ldots, q_N)$ converted from the text prompt via a pronunciation dictionary, the proposed model detects an error state sequence $\mathbf{e} = (e_1, e_2, \ldots, e_N)$ in response to q while simultaneously generating a sequence of phonetic diagnostic feedback $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$, where $y_n$ denotes the phone uttered by the L2 learner.

### B. Model Components

The overall architecture of our proposed model is illustrated in Fig. 2(a), comprising five primary components that work together seamlessly: (1) a prompt encoder that extracts contextualized phone embeddings from the input canonical phone sequence $\mathbf{q}$; (2) a multi-view audio encoder that leverages a pre-trained acoustic model (i.e., wav2vec 2.0) and an articulatory feature extractor to portray the learner's pronunciation from multiple views based on input audio signals $\mathbf{x}$; (3) a concatenation network that produces fused representations from the outputs of the prompt encoder and the multi-view audio encoder; (4) an error-state detector that predicts a sequence of error states $\mathbf{e}$ based on the fused representations in response to $\mathbf{q}$; and (5) a phone predictor that generates a sequence of phone-level dictation result $\boldsymbol{y}$ for yielding the phonetic diagnosis feedback.

### C. Prompt Encoder

We adopt Transformer as the backbone model for the prompt encoder. Specifically, for an input canonical phone sequence $\mathbf{q}$, the contextualized phone embeddings $\mathrm{H}_{pt}$ are derived by mapping $\mathbf{q}$ to the corresponding phone embeddings and then forward propagated through a prompt encoder.

$$\mathrm{E}_{pt} = \mathrm{Embedding}(\mathbf{q}), \tag{1}$$

$$\mathrm{H}_{pt} = \mathrm{PmtEnc}(\mathrm{E}_{pt}), \tag{2}$$

where $\mathrm{E}_{pt}$ denotes a sequence of phone embeddings and $\mathrm{PmtEnc}(\cdot)$ denotes the proposed prompt encoder which empirically adopts a single Transformer block.

### D. Multi-view Audio Encoder

Our multi-view audio encoder comprises two disparate pre-trained feature encoders, a large-scale pretrained acoustic encoder and an articulatory feature extractor. Specifically, we employ wav2vec 2.0 as the acoustic encoder to extracts acoustic features $\mathrm{H}_{ac}$ from the input audio signals $\mathbf{x}$, which are excepted to render the supra-segmental pronunciation cues, such as prosody [26] and intonation [27]. Meanwhile, the articulatory feature extractor takes $\mathbf{x}$ as input, and generates electromagnetic articulography (EMA) features $\mathrm{H}_{ar}$ to characterize the movements of articulatory structures. Here $\mathrm{H}_{ar}$ is a 12-dimensional vector sequence, with each dimension representing the x-y position of an EMA sensor for a specific articulatory structure, such as the lower incisor, upper lip, or lower lip [23].

Afterward, to tightly couple the extract features $\mathrm{H}_{ac}$ and $\mathrm{H}_{ar}$, we first map the acoustic features $\mathrm{H}_{ac}$ to the feature space as $\mathrm{H}_{ar}$ by applying a linear projection followed by a single transformer block. Next, we concatenate the intermediated acoustic representation $\mathrm{H'}_{ac}$ with $\mathrm{H}_{ar}$, and then employ a depth-wise convolution layer to refine the concatenated representations, generating a sequence of multi-view feature vectors $\mathrm{H}_{mv}$. The following equations illustrate the operations of the proposed multi-view audio encoder:

$$\mathrm{H}_{ac}, \mathrm{H}_{ar} = \mathrm{PreEnc}_{ac}(\mathbf{x}), \mathrm{PreEnc}_{ar}(\mathbf{x}) \tag{3}$$

$$\mathrm{H'}_{au} = \mathrm{TFR}(\mathrm{Proj}(\mathrm{H}_{ar})) \tag{4}$$

$$\mathrm{H}_{mv} = \text{Depth-wise Conv}([\mathrm{H}_{ac}; \mathrm{H'}_{au}]) \tag{5}$$

where $[;]$ is the concatenation operation, and Depth-wise $\mathrm{Conv}(\cdot)$ is realized as a 1-D depth-wise convolution layer with a kernel size of 3.

### E. Concatenation Network

The concatenation network is designed to fuse the textual features $\mathrm{H}_{pt}$ with multi-view acoustic features $\mathrm{H}_{mv}$, enabling the model to attend to the full context of the input phone representations both within and between the textual and acoustic information via the self-attention mechanism.

$$[\overline{\mathrm{H}}_{pt}; \overline{\mathrm{H}}_{au}] = \mathrm{ConcatNet}([\mathrm{H}_{pt}; \mathrm{H}_{au}]) \tag{6}$$

where $\mathrm{ConcatNet}(\cdot)$ is a stack of 2 Transformer blocks.

### F. Error-state Detector

The error-state detector is a sequential binary labeling model that receives $\overline{\mathrm{H}}_{pt} = (\overline{\mathrm{h}}_1^{pt}, \overline{\mathrm{h}}_2^{pt}, \ldots, \overline{\mathrm{h}}_N^{pt})$ from the concatenation network, and then outputs a sequence of error states $\mathbf{e} = (e_1, e_2, \ldots, e_N)$. Here each element in $\mathbf{e}$ represents the error state of $n$-th phone in the canonical phone sequence $\mathbf{q}$, with 1 indicating that phone $q_n$ is pronounced incorrectly and 0 indicating correct pronunciation.

For each error state $e_n$ of the canonical phone $q_n$, the probability of error detection is defined as:

$$\mathcal{P}_{\mathrm{dec}}(e_n = 1|\mathbf{x}, \mathbf{q}) = \sigma\big(\mathrm{proj}_d(\overline{\mathrm{h}}_n^{pt})\big) \tag{7}$$

where $\mathcal{P}_{\mathrm{dec}}(e_n = 1|\mathbf{x}, \mathbf{q})$ is the conditional probability which represents how likely the phone $q_n$ is mispronounced, $\sigma$ represents the nonlinear function which we used sigmoid function, $\mathrm{proj}_d$ is the linear projection layer which transforms the hidden dimension to a single scaler. Furthermore, to estimate the error states in a fine-grained manner, we determine the optimal threshold value for each canonical phone $q_n$ in the development set.

### G. Predictor

The phone predictor is responsible for dictating the L2 learner' speech, which takes $\overline{\mathrm{H}}_{au} = (\overline{\mathrm{h}}_1^{au}, \overline{\mathrm{h}}_2^{au}, \ldots, \overline{\mathrm{h}}_T^{au})$ as input and generates a sequence of dictated phones as phonetic diagnosis feedback. The prediction process is optimized with the connectionist temporal classification (CTC) loss which manages the alignment between the input features and the phone sequence $\boldsymbol{y}$. This CTC loss is denoted as follows:

$$\begin{aligned}\mathcal{P}_{diag}(\boldsymbol{y}|\mathbf{x}, \mathbf{q}) &= \sum_{A \in \mathcal{B}_{CTC}^{-1}(\boldsymbol{y})} \mathcal{P}(\boldsymbol{y}|A, \mathbf{x}, \mathbf{q}) \mathcal{P}(A|\mathbf{x}, \mathbf{q}) \\ &= \sum_{A \in \mathcal{B}_{CTC}^{-1}(\boldsymbol{y})} \mathcal{P}(A|\mathbf{x}, \mathbf{q})\end{aligned} \tag{8}$$

The joint probability $\mathcal{P}(A|\mathbf{x}, \mathbf{q})$ is further factorized using the probabilistic chain rule accompanied by a conditional independence assumption.

$$\begin{aligned}\mathcal{P}(A|\mathbf{x}, \mathbf{q}) &= \prod_{t=1}^{T} \mathcal{P}(\mathbf{a}_t|\mathbf{a}_1, \ldots, \mathbf{a}_{t-1}, \mathbf{x}, \mathbf{q}) \\ &\approx \prod_{t=1}^{T} \mathcal{P}(\mathbf{a}_t|\mathbf{x}, \mathbf{q})\end{aligned} \tag{9}$$

The conditional probability $\mathcal{P}(\mathbf{a}_t|\mathbf{x}, \mathbf{q})$ in Eq. (11) is computed as

$$\mathcal{P}(\mathbf{a}_t|\mathbf{x}, \mathbf{q}) = \mathrm{Softmax}(\mathrm{Proj}_{\mathrm{diag}}(\overline{\mathrm{H}}_{au})) \tag{10}$$

where $\mathrm{Proj}_{\mathrm{diag}}(\cdot)$ is a linear projection converting hidden dimension to the vocabulary size.

### H. Training objective

The proposed model is trained in an end-to-end manner, with the learning process designed to optimize two objectives: pronunciation error detection and phonetic diagnosis.

**Pronunciation Error Detection.** The objective function of pronunciation error detection $\mathcal{L}_{det}$ is defined by the negative log likelihood of Eq. (8) over all error states:

$$\mathcal{L}_{det} = -\log \sum_{n=1}^{N} \mathcal{P}_{\text{det}}(e_n|\mathbf{x}, \mathbf{q}) \quad (11)$$

**Phone Predictor.** The objective function of phone predictor is defined by the negative log-likelihood of Eq. (10) over all possible alignments:

$$\mathcal{L}_{diag} = -\log \sum_{A \in \mathcal{B}_{\text{CTC}}^{-1}(\mathbf{y})} \prod_{t=1}^{T} \mathcal{P}(\mathbf{a}_t|\mathbf{x}, \mathbf{q}) \quad (12)$$

These two modules are linearly combined as the overall objective for model learning:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{det} + \mathcal{L}_{diag} \quad (13)$$

where $\alpha \in [0,1]$ is a controllable parameter: a larger $\alpha$ value means a higher weight on the mispronunciation diagnosis subtask and the value of $\alpha$ is set to be 0.3 based on the development set.

### *I. Inference*

At inference time, a prompt-guided mechanism is put forward for the PG-MDD to strike a balance between recall and precision evaluations. We first identify erroneous pronunciation segments with the outputs of the error-state detector and the phone-dependent thresholds. Meanwhile, the N-best diagnosis results are obtained from the phone predictor, which are then aligned to the canonical phone sequence with Levenshtein distance, individually. Afterward, based on the identified mispronunciation segments, the diagnosis results are obtained by checking where the aligned candidates differ from the canonical phone sequence.

## III. EXPERIMENTS

### *A. Speech Datasets*

**TIMIT** [21] is a widely used English read speech dataset that was designed for the development and evaluation of automatic speech recognition systems. TIMIT contains recordings of 630 speakers from 8 major dialects of American English.

**L2-ARCTIC** [20] is a recently released L2-English speech dataset, which was compiled for research on CAPT, accent conversion, and others. L2-ARCTIC contains both correctly pronounced and mispronounced utterances from 24 non-native speakers (12 males and 12 females), whose mother-tongue languages include Hindi, Korean, Mandarin, Spanish, Arabic, and Vietnamese.

### *B. Experimental Setup*

**Experimental Settings.** Before the joint optimization of mispronunciation detection and phone recognition task, we fine-tune the acoustic encoder and phone predictor using the TIMIT corpus. We then perform joint optimization of the proposed model using the L2-ARCTIC corpus. Furthermore, to unify the phone dictionaries used in TIMIT and L2-ARCTIC, we map the 61-phone set of the TIMIT corpus to the 39-phone set used in the L2-ARCTIC corpus. In our experimental setup, each attention head in the transformer block is configured as a single-head attention. All model components are optimized using the Adam optimizer [27] for 1000 steps, with an initial learning rate of $7\times10^{-5}$and a batch size of 64.

**Implementation Details of the Multi-view Audio Encoder.** We adopt wav2vec 2.0-large as the acoustic encoder, which consists of a feature encoder, a context network, and a quantization module. Each component of wav2vec 2.0 is jointly learned through a combination of mask prediction and contrastive learning. Furthermore, wav2vec 2.0 achieves state-of-the-art results in phone recognition, which is highly relevant to CAPT. For the articulatory feature extractor, we adopt an acoustic-to-articulatory inversion (AAI) pre-trained model. This pre-trained model utilizes self-supervised learning techniques to generalize across unseen speaker data, mapping speech signals to an articulatory space that captures the mechanisms of speech production. The model has been trained on the Electromagnetic Articulography (EMA) dataset and demonstrates the interpretability of its feature representations by validating estimated feature representations against actual speech production behaviors.

TABLE I
Experimental results of different methods on mispronunciation detection.

| Model | PR (%)↑ | RE (%)↑ | F1 (%)↑ |
|---|---|---|---|
| GOP [9] | 69.97 | 32.54 | 44.42 |
| CNN-RNN-CTC [13] | **74.78** | 36.76 | 49.29 |
| APL-2 [25] | 52.79 | 54.49 | 53.62 |
| CoCA-MDD [19] | 64.65 | 51.01 | 57.03 |
| MDDGCN [16] | 51.90 | **61.97** | 56.49 |
| RNN-T [24] | 63.40 | 55.30 | 59.10 |
| PG-MDD | 60.15 | 60.06 | **60.10** |

**Evaluation Metrics.** We assess the performance of our proposed method using Precision (PR), Recall (RE), and F1-score. For canonical phonemes, true acceptance (TA) and true rejection (TR) denote correctly identified phonemes, while false acceptance (FA) and false rejection (FR) denote misclassified phonemes. These metrics are used to comprehensively evaluate our model's effectiveness.

## IV. EXPERIMENTAL RESULTS

### *A. Comparison with Prior Arts*

At the outset, we report on the mispronunciation detection results in Table I to compare the proposed method with several cutting-edge baselines. There are three noteworthy points to these results. First, all the dictation-based methods (CNN-RNN-CTC, APL-2, CoCA-MDD, and RNN-T) significantly outperform the conventional scoring-based method, GOP, in terms of F1-score, demonstrating the promising potential of dictation-based models for MDD tasks. Second, most detection-based methods exhibit high precision but yield inferior results in recall. This result underscores the challenge that dictation-based MDD models face in striking a balance between precision and recall when detecting pronunciation errors. Conversely, the prompt-based model (MDDGCN) tends to aggressively detect pronunciation errors in L2 learners' pronunciations, leading to a higher recall while resulting in a decrease in precision. Finally, the proposed PG-MDD model integrates a multi-view encoder and employs a prompt-guided mechanism, which not only outperforms other state-of-the-art methods in terms of F1-score but also mitigates the limitations in both precision and recall encountered by previous models.

TABLE II
Ablation studies on PG-MDD.

| Model | Canonicals | | Mispronunciations | | | RE (%)↑ | PR (%)↑ | F1 (%)↑ | PER(%)↓ |
|---|---|---|---|---|---|---|---|---|---|
| | | | | True Rejection | | | | | |
| | TA (%)↑ | FR (%)↓ | FA (%)↓ | Corr Diag. (%)↑ | Diag. Error (%)↓ | | | | |
| PG-MDD | 93.36 | 6.63 | **39.94** | 76.87 | 23.13 | **60.06** | 60.15 | **60.10** | 13.92 |
| -w/o PG | **94.87** | **5.13** | 44.33 | **80.20** | **19.80** | 55.67 | **64.45** | 59.74 | **12.80** |
| -w/o ES | 94.57 | 5.43 | 45.19 | 79.46 | 20.54 | 54.81 | 62.74 | 58.51 | 13.20 |
| -w/o ES and PmtEnc | 93.69 | 6.31 | 43.14 | 78.20 | 21.80 | 56.86 | 60.07 | 58.42 | 13.97 |
| -w/o ES, PmtEnc, and ArEnc | 94.05 | 5.95 | 45.63 | 78.35 | 21.65 | 54.37 | 60.38 | 57.22 | 13.98 |

### *B. Ablation Studies*

Next, a series of model ablation experiments are conducted, as shown in Table II, to demonstrate the importance of each component in our PG-MDD architecture. This ablation study reports the following settings: 1) inference without the prompt-guided mechanism (-w/o PG) 2) removing the error-state detector (-w/o ES), 3) removing the error-state detector and the prompt encoder (- w/o ES and PmtEnc), and 4) removing error-state detector, the prompt encoder, and the articulatory feature extractor from the multi-view audio encoder (- w/o ES, PmtEnc, and ArEnc). The main motivation of this paper is to balance the performance of recall and precision while preserving the high F1-score. If PG-MDD inference is conducted without the proposed prompt-guided mechanism, the model performance deteriorates in F1-score, leading to an imbalance between recall and precision, even though a more accurate phone recognition rate is achieved. Next, when the error-state detector is removed, performance declines across all evaluation metrics compared to PG-MDD. This result suggests that integrating an error-state detector into the dictation-based MDD model enhances the effectiveness of the mispronunciation detection and diagnosis task. Finally, we assess the impact of articulatory features by comparing the last two rows in Table II. We observe that when the articulatory feature extractor is removed, performance declines in terms of recall while achieving a comparable phone error rate. We attribute this to the fact to articulatory features that are beneficial for the task of mispronunciation detection.

### *C. Impact of Articulatory Features on MDD*

Going one step further, we delve into the impact of articulatory features by reporting the performance changes in the common mispronunciation patterns [20]. From Table III, it is evident that the articulatory features provide more subtle pronunciation cues for the proposed PG-MDD, thereby enhancing the performance of mispronunciation diagnosis for the frequently mispronounced phone pairs, such as (/z/, /s/) and (/ih/, /iy/). Notably, articulatory features show significant improvement for the mispronunciation pattern (/v/, /f/). On the other hand, some mispronunciation patterns result in inferior performance with articulatory features, such as (/p/, /b/).

TABLE III
Impact of articulatory features on phone-level substitution error detection and diagnosis.

| Error Type | w/o Articulatory Feats | | w/ Articulatory Feats | |
|---|---|---|---|---|
| | Correct Detection | Correct Diag. | Correct Detection | Correct Diag. |
| z->s | 74.62 | 72.61 | **79.64** | **77.63** |
| v->f | 11.86 | 11.86 | **28.81** | **27.11** |
| ih->iy | 28.30 | 27.67 | **30.81** | **30.18** |
| ow->ao | 11.60 | 10.71 | **16.96** | **11.39** |
| er->ah | **73.46** | **60.20** | 66.36 | 47.95 |
| dh->d | **89.90** | **84.86** | 70.64 | 67.20 |
| d->t | 34.24 | 27.39 | **38.35** | **35.61** |
| p->b | **31.65** | **29.68** | 17.18 | 15.62 |
| th->t | **84.61** | **75.00** | 80.76 | 67.30 |

## V. CONCLUSION

In this paper, we have proposed a novel prompt-guided dictation-based mispronunciation detection and diagnosis (MDD) model, termed PG-MDD. Our model integrates the strengths of a multi-view audio encoder and a prompt-guided mechanism within a unified framework, which jointly optimizes the processes of mispronunciation detection and diagnosis. This approach effectively addresses the common issue of how to properly strike a balance between precision and recall rates, facing dictation-based models. A comprehensive set of empirical results on the L2-ARCTIC benchmark dataset shows that our proposed architecture significantly outperforms several competitive baselines. PG-MDD achieved the highest F1-score, demonstrating robust performance in both precision and recall. This balanced performance is particularly crucial for providing reliable and comprehensive feedback in practical applications, ensuring more effective learning outcomes for L2 learners.

## REFERENCES

[1] Nancy F. Chen, and Haizhou Li. “Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning,” in proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 1–7, 2016.

[2] Alistair V. Moere and Ryan Downey, “Technology and artificial intelligence in language assessment,” Handbook of second language assessment, pp. 341–358, 2016.

[3] Keelan Evanini, Xinhao Wang, “Automated speech scoring for non-native middle school students with multiple task types,” in Proceedings of Interspeech (INTERSPEECH), pp. 2435–2439, 2013.

[4] Keelan Evanini, Maurice Cogan Hauck, Kenji Hakuta, “Approaches to automated scoring of speaking for K–12 English language proficiency assessments,” ETS Research Report Series, pp. 1–11, 2017.

[5] Ramsey Cardwell, Ben Naismith, Geoffrey T. LaFlair, and Steven Nydick. “Duolingo English test: technical manual,” Duolingo Research Report, 2022.

[6] Wei Li, Nancy F. Chen, Sabato Marco Siniscalchi, and Chin-Hui Lee, “Improving mispronunciation detection of mandarin tones for non-native learners with soft-target tone labels and BLSTM-based deep tone models,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 12, pp. 2012–2024, 2019.

[7] Kun Li, Shaoguang Mao, Xu Li, Zhiyong Wu, and Helen Meng, “Automatic lexical stress and pitch accent detection for L2 English speech using multi-distribution deep neural networks,” Speech Communication, vol. 96, pp. 28–36, 2018.

[8] Daniel Korzekwa, Jaime Lorenzo-Trueba, Thomas Drugman, Shira Calamaro, and Bozena Kostek, “Weakly-supervised word-level pronunciation error detection in non-native English speech,” in proceedings of Interspeech (INTERSPEECH), pp. 4408–4412, 2021.

[9] Silke M. Witt and S. J. Young, “Phone-level pronunciation scoring and assessment for interactive language learning,” Speech Communication, vol. 30, pp. 95–108, 2000.

[10] Quy-Thao Truong, Tsuneo Kato, and Seiichi Yamamoto, “Automatic assessment of L2 English word prosody using weighted distances of F0 and intensity contours,” in proceedings of Interspeech (INTERSPEECH), pp. 2186-2190, 2018.

[11] Jiatong Shi, Nan Huo, and Qin Jin, “Context-aware goodness of pronunciation for computer-assisted pronunciation training,” in proceedings of Interspeech (INTERSPEECH), pp. 3057–3061, 2020.

[12] Shaoguang Mao, Zhiyong Wu, Runnan Li, Xu Li, Helen Meng, and Lianhong Cai, “Applying multitask learning to acoustic-phonemic model for mispronunciation detection and diagnosis in L2 English speech,” in proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 6254–6258, 2018.

[13] Wai-Kim Leung, Xunying Liu, and Helen Meng, “CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis,” in proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8132-8136, 2019.

[14] Bi-Cheng Yan, Meng-Che Wu, Hsiao-Tsung Hung, Berlin Chen, “An end-to-end mispronunciation detection system for L2 English speech leveraging novel anti-phone modeling,” in proceedings of Interspeech (INTERSPEECH), pp. 3032–3036, 2020.

[15] Bi-Cheng Yan, Hsin-Wei Wang, and Berlin Chen, “Peppanet: Effective mispronunciation detection and diagnosis leveraging phonetic, phonological, and acoustic cues,” in Proceedings of IEEE Spoken Language Technology Workshop (SLT), pp. 1045–1051, 2023.

[16] Bi-Cheng Yan, Hsin-Wei. Wang, Yi-Cheng Wang, and Berlin Chen, “Effective graph- based modeling of articulation traits for mispronunciation detection and diagnosis,” in proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5, 2023.

[17] Linkai Peng, Kaiqi Fu, Binghuai Lin, Deng Feng Ke, and Jinsong Zhan, “A study on fine-tuning wav2vec2.0 model for the task of mispronunciation detection and diagnosis,” in proceedings of Interspeech (INTERSPEECH), pp. 4448–4452, 2021.

[18] Minglin Wu, Kun Li, Wai-Kim Leung, and Helen Meng, “Transformer based end-to-end mispronunciation detection and diagnosis,” in proceedings of Interspeech (INTERSPEECH), pp. 3954–3958, 2021.

[19] Nianzu Zheng, Liqun Deng, Wenyong Huang, Yu Ting Yeung, Baohua Xu, Yuanyuan Guo, Yasheng Wang, Xiao Chen, Xin Jiang, and Qun Liu, “Cca-mdd: A coupled crossattention based framework for streaming mispronunciation detection and diagnosis,” arXiv preprint arXiv:2111.08191, 2021.

[20] Guanlong Zhao, Sinem Sonsaat, A. Silpachai, I. Lucic, E. C.-Hudilainen, J. Levis and R. G.-Osuna, “L2-ARCTIC: A Non-native English Speech Corpus,” in proceedings of Interspeech (INTERSPEECH), pp. 2783–2787, 2018.

[21] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, “The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM,” Linguistic Data Consortium, 1993.

[22] Baevski, Alexei, et al. “wav2vec 2.0: A framework for self-supervised learning of speech representations,” Advances in neural information processing systems 33, pp. 12449-12460, 2022.

[23] Peter Wu, Li-Wei Chen, Cheol Jun Cho, Shinji Watanabe, Louis Goldstein, Alan W Black, Gopala K. Anumanchipalli, “Speaker-Independent Acoustic-to-Articulatory Speech Inversion,” in proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5, 2023.

[24] Daniel Yue Zhang, Soumya Saha and Sarah Campbell , “Phonetic RNN-Transducer for Mispronunciation Diagnosis,” in proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5, 2023.

[25] Wenxuan Ye, Shaoguang Mao, Frank Soong, Wenshan Wu, Yan Xia, Jonathan Tien, Zhiyong Wu "An Approach to Mispronunciation Detection and Diagnosis with Acoustic, Phonetic and Linguistic (APL) Embeddings," in proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6827-6831, 2022.

[26] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, and B. Chen, “3M: An effective multi-view, multi-granularity, and multi-aspect modeling approach to English pronunciation assessment,” in proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 575-582, 2022.

[27] S. Bannò, and M. Matassoni, “Proficiency assessment of L2 spoken English using wav2vec 2.0,” in IEEE Spoken Language Technology Workshop (SLT), pp. 1088-1095, 2022.

[28] C. McGhee, K. Knill, and M. Gales, “Towards acoustic-to-articulatory inversion for pronunciation training,” In Workshop on Speech and Language Technology in Education (SLaTE), pp. 66-70, 2023.

[29] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2017.

[30] Konstantinos Kyriakopoulos, Kate M. Knill, and Mark J.F. Gales, “Automatic detection of accent and lexical pronunciation errors in spontaneous nonnative English speech,” in proceedings of Interspeech (INTERSPEECH), pp. 3052–3056, 2020.

[31] Yassine El Kheir, Shammur Absar Chowdhury, and Ahmed Ali, “Multi-view multi-task representation learning for mispronunciation detection,” in arXiv preprint arXiv:2306.01845, 2023.

[32] Daniel-Yue Zhang, Soumya Saha, and Sarah Campbell, “Phonetic RNN-transducer for mispronunciation diagnosis,” in proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5, 2023