JCAPT: A Joint Modeling Approach for CAPT

Tzu-Hsuan Yang, Yue-Yang He, Berlin Chen

[1]National Taiwan Normal University, Taipei, Taiwan

tzuhsuan@ntnu.edu.tw, yueyanghe@ntnu.edu.tw, berlin@ntnu.edu.tw

Abstract

Effective pronunciation feedback is critical in second language (L2) learning, for which computer-assisted pronunciation training (CAPT) systems often encompass two key tasks: automatic pronunciation assessment (APA) and mispronunciation detection and diagnosis (MDD). Recent work has shown that joint modeling of these two tasks can yield mutual benefits. Our unified framework leverages Mamba, a selective state space model (SSM), while integrating phonological features and think token strategies to jointly enhance interpretability and fine-grained temporal reasoning in APA and MDD. To our knowledge, this is the first study to combine phonological attribution, SSM-based modeling, and prompting in CAPT. A series of experiments conducted on the speechocean762 benchmark demonstrate that our model consistently outperforms prior methods, particularly on the MDD task.

Index Terms: computer-assisted pronunciation training, speech attributes, Mamba, L2 speech assessment, multi-aspect scoring

1. Introduction

In the era of globalized communication, learning a second language (L2) has become increasingly essential. Computer-assisted pronunciation training (CAPT) systems have emerged as practical and scalable solutions. These systems provide learners with a low-pressure, self-directed environment to enhance their pronunciation skills through immediate, objective, and personalized feedback [1, 2]. To offer meaningful guidance, effective CAPT systems are typically composed of two integral components: automatic pronunciation assessment (APA), which delivers a broader evaluation of speaking skills; and mispronunciation detection and diagnosis (MDD), which aims to identify pronunciation errors and provide detailed diagnostic feedback.

A de-facto CAPT system typically operates in a read-aloud scenario, where L2 learners are prompted to speak predefined sentences. In this context, APA modules assess pronunciation quality across various aspects (e.g., accuracy, fluency, and completeness) and multiple linguistic granularities (e.g., phoneme-, word-, and utterance-levels) [3, 4, 5]. On a separate front, MDD focuses on identifying phonetic pronunciation errors commonly made by non-native speakers [6, 7, 8]. Such errors tend to have clear-cut distinctions between correct and incorrect pronunciations and can be systematically identified via phoneme-level mismatches including deletions, substitutions, and insertions. These two components collaboratively enable CAPT systems to deliver holistic and pedagogically relevant feedback to L2 learners for improving their pronunciation proficiency.

Recent research has shown that integrating APA and MDD into a unified modeling framework can enhance the performance of both tasks [9, 10, 11]. Such a synergy enhances not only the granularity of mispronunciation detection but also the reliability of multi-aspect pronunciation scoring. However, an underexplored, yet promising direction involves the use of articulatory attributes, which provide a linguistically grounded representation of phoneme realization, such as voicing, manner, place of articulation, and others. While these articulatory attributes have been leveraged to improve MDD performance [12, 13], their role in enhancing multi-aspect APA, in conjunction with MDD and under a joint modeling framework, remains largely unexplored.

Beyond incorporating linguistic knowledge, architectural innovations can also significantly affect the effectiveness of CAPT systems. In particular, increasing the capacity of model components to perform deeper reasoning or richer representation learning at the frame- or linguistic token-levels has shown promise. Inspired by chain-of-thought reasoning in natural language processing [14], the concept of "think tokens" has been proposed to increase the computational depth per token. This idea has been applied to Transformer-based ASR models [15], where encouraging frame-wise reflection has improved both inference quality and training efficiency. In parallel, Mamba [16], a selective state space model (SSM), has demonstrated strong potential for modeling long-range dependencies with high computational efficiency. These characteristics make it a promising candidate for fine-grained speech assessment tasks [11]. There are some research efforts exploring how prompting mechanisms can be adapted to SSM-based architectures such as Mamba [17], but to our knowledge, there is little work focusing on its applicability in CAPT scenarios, especially for phoneme-level representations.

In this paper, we present JCAPT, a **J**oint **CAPT** framework that jointly considers APA and MDD tasks using a parallel architecture, as illustrated in Figure 1. JCAPT leverages Mamba, a state space model (SSM) capable of capturing long-range temporal dependencies with high computational efficiency, for rendering phone- and word-level pronunciation characteristics. In addition to utilizing canonical phone information as in prior work [3, 18, 10], our model also integrates phonological features to enhance diagnostic precision and interpretability. To our knowledge, our work is among the first to investigate the synergistic effect of phonological attribution, think token strategies, and Mamba-based architectures within a unified modeling framework for assessing the pronunciation proficiency of L2 learners. We validate the proposed JCAPT approach through comprehensive experiments on the speechocean762 benchmark, demonstrating that JCAPT significantly outperforms existing baselines, especially with respect to MDD performance. Notably, we find that completeness—an
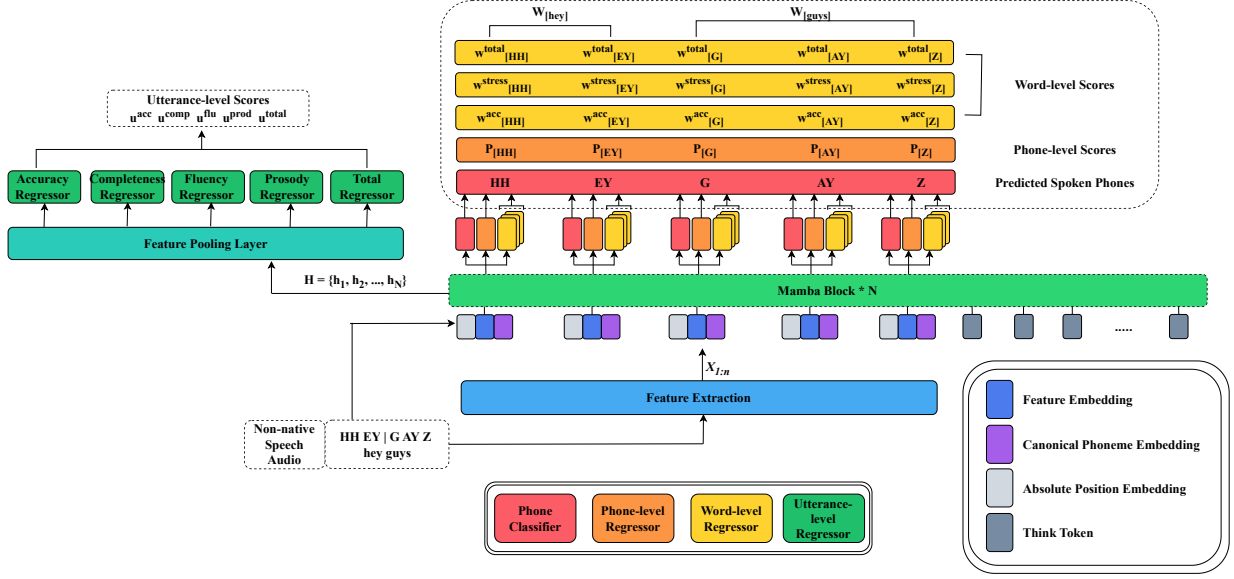
Figure 1: *The overall architecture of the multi-task learning model for pronunciation assessment and mispronunciation detection and diagnosis. The system consists of feature extraction, Mamba layers for temporal modeling, and multi-level scoring modules including phoneme-level, word-level, and utterance-level regressors.*

aspect of APA that remains particularly challenging—can be substantially improved by subtler modeling of phoneme-level pronunciation traits.

2. Methodology

Figure 1 schematically visualizes our proposed framework, JCAPT, which jointly models Automatic Pronunciation Assessment (APA) and Mispronunciation Detection and Diagnosis (MDD) through a parallel architecture. Our system consists of five key components: 1) a comprehensive feature extraction module that integrates multiple speech representations; 2) a bi-directional Mamba encoder for contextual modeling; 3) a contemplative reasoning mechanism via think tokens; 4) an attention-based pooling layer; and 5) multi-level scoring heads for APA and MDD.

2.1. Feature Extraction

Given a speech utterance from an L2 learner and the canonical phone sequence $p = \{p_1, p_2, \ldots, p_N\}$ of the corresponding text prompt, our system extracts a set of phone-level features by combining goodness of pronunciation (GOP) [19, 20, 21] with modern self-supervised representations.

Goodness of Pronunciation: GOP is a widely used feature that measures the likelihood of each phone being correctly pronounced by a speaker. We follow the standard pipeline for GOP computation, which includes forced alignment via a DNN-HMM acoustic model, canonical phoneme decoding, and posterior probability estimation. This produces phone-aligned GOP features that directly reflect pronunciation accuracy. Due to its explicit modeling of phoneme correctness, GOP is particularly effective for MDD.

Self-Supervised Representations: To capture rich contextual and articulatory information, we incorporate three self-supervised learning (SSL) models: wav2vec 2.0 [22], HuBERT [23], and WavLM [24], which are pre-trained on large-scale un-

labeled speech data. We extract frame-level hidden features from each SSL model and align them to the canonical phone boundaries using forced alignment. The aligned features are then concatenated with GOP scores to form a comprehensive phoneme-level feature vector. The resulting feature vectors of an input utterance are projected through a dense layer to obtain a sequence of phone-level embeddings, denoted as $x_{1:N}$, where N is the number of canonical phones.

Canonical Phone Embedding: In parallel, we construct canonical phone embeddings to provide symbolic linguistic supervision. For each phoneme p_i, we convert it into a one-hot vector Phn_{onehot}, and concatenate it with a phonological attribute vector Phn_{attr} that encodes articulatory properties. The resulting phoneme-level symbolic representation is projected into the same dimension as $x_{1:N}$, and later fused into the model as an auxiliary input.

2.2. Bi-directional Mamba Encoder

To model long-range dependencies with low complexity, we adopt a bi-directional Mamba encoder inspired by the Dual-Mamba architecture [25]. Its linear scaling and efficient temporal representation make it well-suited for fine-grained speech assessment.

Given the phoneme-level acoustic feature sequence $x_{1:N}$ and the canonical phoneme embedding sequence that is constructed from the concatenation of phoneme one-hot vectors and their associated phonological attributes, we fuse both inputs to form the final encoder input:

$$\hat{x}_i = x_i + c_i, \quad \text{for } i = 1, \ldots, N \quad (1)$$

where x_i is the projected acoustic feature and c_i is the symbolic canonical embedding.

The resulting sequence $\hat{X} = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N\}$ is then passed through a stack of bidirectional Mamba blocks. To enhance the encoder's reasoning capacity, we append a set of

Models	Phoneme-level		Word-level			Utterance-level					MDD Performance		
	MSE ↓	PCC ↑	Acc. ↑	Stress ↑	Total ↑	Acc. ↑	Comp. ↑	Fluency ↑	Prosody ↑	Total ↑	RE. (%) ↑	PR. (%) ↑	F1. (%) ↑
Joint-CAPT-L1 [9]	-	-	-	-	-	0.719	-	0.775	0.773	0.743	**91.40**	26.70	41.40
JAM [10]	0.076	0.664	0.622	0.241	0.638	0.773	0.205	0.831	**0.829**	0.805	34.76	64.10	45.01
JCAPT	**0.066**	**0.720**	**0.699**	**0.270**	**0.711**	**0.783**	**0.551**	**0.834**	0.824	**0.806**	40.23	**69.89**	**51.05**

Table 1: *Experimental results of different methods evaluated on speechocean762. Acc. and Comp. refers to Accuracy and Completeness, respectively.*

learnable *think tokens* to the input sequence. Their mechanism and motivation are detailed in Section 2.3.

The output of the encoder is a sequence of contextualized phoneme-level representations:

$$H = \text{BiMamba}(\hat{X}, Emb_{think}) = \{h_1, h_2, \ldots, h_N\}, \quad (2)$$

where Emb_{think} denotes the embedding of the think tokens. This enriched representation H forms the basis for the subsequent reasoning and prediction modules.

2.3. Contemplative Reasoning via Think Tokens

Inspired by contemplative prompting [15], we introduce *think tokens* to encourage deeper reasoning over each phoneme-level representation. Instead of inserting interleaved think tokens within a Transformer-based architecture, we postpend a fixed number of think tokens to the end of the input sequence, allowing the model to perform additional internal computation before making phoneme-level predictions. These tokens are implemented as learnable embeddings and are jointly optimized during training. This mechanism is particularly effective in enhancing the diagnostic capacity of MDD and improving the consistency of multi-aspect APA predictions.

2.4. Attention-based Feature Pooling

To obtain utterance-level representations for multi-aspect pronunciation assessment, we employ a set of aspect-specific attention-based pooling mechanisms. Given the encoder output $H = \{h_1, h_2, \ldots, h_N\} \in \mathbb{R}^{N \times d}$, we define a separate attention module for each assessment aspect $a \in \mathcal{A}$, where $\mathcal{A}$ denotes the set of predefined aspects (e.g., accuracy, fluency, prosody).

For each aspect a, attention weights are computed by

$$\alpha_i^{(a)} = \frac{\exp\left(w_a^\top \tanh(W_a h_i)\right)}{\sum_{j=1}^{N} \exp\left(w_a^\top \tanh(W_a h_j)\right)}, \quad (3)$$

where $W_a \in \mathbb{R}^{d_a \times d}$ and $w_a \in \mathbb{R}^{d_a}$ are learnable parameters for aspect a.

The utterance-level representation for aspect a is obtained by:

$$h_u^{(a)} = \sum_{i=1}^{N} \alpha_i^{(a)} h_i. \quad (4)$$

This design allows each aspect to focus on different parts of the input sequence, reflecting its unique contribution to overall pronunciation quality.

2.5. Multi-level Scoring Heads

To support both the APA and MDD tasks across different granularities, we design hierarchical prediction heads. For each contextualized phoneme representation h_i, a regression head estimates phoneme-level APA scores, while a classification head

predicts corresponding MDD labels. Word-level scores are derived by aggregating phoneme-level outputs based on forced alignment boundaries. For utterance-level APA, each $h_u^{(a)}$ obtained from the attention-based pooling layer is passed through an individual regression head corresponding to the aspect a, yielding holistic multi-aspect assessment scores. This multi-level design enables joint modeling of fine-grained and global pronunciation quality indicators.

2.6. Optimization

Our model is trained under a multi-task learning (MTL) framework that optimizes both the APA and MDD objectives jointly. For APA, we formulate the objective as the sum of losses across different granularity levels:

$$\mathcal{L}_{\text{APA}} = \mathcal{L}_{\text{phn}} + \mathcal{L}_{\text{word}} + \mathcal{L}_{\text{utt}}, \quad (5)$$

where each component denotes the mean squared error (MSE) loss at the phoneme-, word-, and utterance-levels, respectively.

For MDD, a phoneme-level classifier is trained using cross-entropy loss to improve mispronunciation detection:

$$\mathcal{L}_{\text{MDD}} = -\sum_{i=1}^{N} \sum_{p=1}^{P} y_{i,p} \log(\hat{y}_{i,p}), \quad (6)$$

where N is the number of training instances, P is the number of phoneme classes, $y_{i,p}$ is the one-hot ground truth, and $\hat{y}_{i,p}$ is the predicted probability for the p-th phoneme in the i-th instance.

The ultimate loss is a weighted combination of the CAPT and MDD losses:

$$\mathcal{L} = (1 - \alpha) \cdot \mathcal{L}_{\text{APA}} + \alpha \cdot \mathcal{L}_{\text{MDD}}, \quad (7)$$

where α strikes a balance between these two objectives. In our implementation, we set $\alpha = 0.3$ following [10], which was found to yield stable performance in similar settings.

3. Experiments and Results

3.1. Dataset

We conducted our experiments on the speechocean762 dataset [26], a publicly available benchmark designed for research on automatic pronunciation assessment (APA) and mispronunciation detection and diagnosis (MDD). The dataset contains 5,000 English utterances produced by 250 Mandarin-speaking L2 learners, evenly divided into training and test sets. Each utterance is annotated with human-rated pronunciation scores at the utterance, word, and phoneme levels, assessed by five expert raters using standardized rubrics. For the MDD task, the dataset provides canonical and realized phone-level transcriptions, aligned at the phoneme level. It adopts a 39-phone set, based on the CMU pronunciation dictionary [27] and extended with <del> and <unk> tokens to indicate deleted and non-categorizable phones. In particular, the data set does not include insertion errors, which facilitates cleaner alignment between canonical and observed pronunciations.

Models	Phoneme-level Score		Word-level Score (PCC)			Utterance-level Score (PCC)				
	MSE ↓	PCC ↑	Acc. ↑	Stress ↑	Total ↑	Acc. ↑	Comp. ↑	Fluency ↑	Prosody ↑	Total ↑
JCAPT	**0.066**	**0.720**	**0.699**	0.270	**0.711**	0.783	0.551	0.834	0.824	0.806
w/o phonological	**0.066**	0.716	0.689	0.239	0.701	0.775	**0.644**	**0.840**	**0.826**	**0.808**
w/o think tokens	**0.066**	**0.720**	**0.699**	**0.309**	0.710	**0.784**	0.556	0.833	0.818	**0.808**
w/o phonological, think token	0.068	0.708	0.687	0.273	0.699	0.779	0.547	0.834	0.822	**0.808**

Table 2: *Ablation Studies of the proposed method on automatic pronunciation assessment.*

Models	MDD results				Diagnosis results
	Recall (%) ↑	Precision (%) ↑	F1-score (%) ↑	PER (%) ↓	Correct Diag. (%) ↑
JCAPT	40.23	69.89	51.05	**2.66**	**54.42**
w/o phonological	**42.00**	68.98	**52.21**	2.70	52.67
w/o think tokens	39.76	69.95	50.61	2.67	54.35
w/o phonological, think token	41.27	**70.07**	51.92	2.67	52.80

Table 3: *Abalation Studies of proposed method on mispronunciation detection and diagnosis.*

3.2. Experimental Setup

Following the experimental configuration outlined in [10], we adopted the same procedures for extracting both GOP and SSL features. Additionally, we incorporated phonological attributes as introduced in [12], enriching our phoneme-level representations. To ensure robustness and reproducibility, we conducted five independent runs with different random seeds. In the following experiments, we will report the average performance in terms of Pearson Correlation Coefficient (PCC) and Mean Squared Error (MSE) across all runs.

In addition, following previous MDD studies [28], we adopt F1-score, the harmonic mean of recall (RE.) and precision (PR.), as the primary metric for evaluating mispronunciation detection performance. Furthermore, we report the phoneme error rate (PER) to reflect overall detection accuracy at the segmental level. For diagnostic evaluation, we include the correct diagnosis rate (Correct Diag.), which measures the proportion of detected errors that are correctly classified, providing insight into the interpretability and reliability of the system's feedback.

3.3. Main Results

As shown in Table 1, JCAPT consistently outperforms previous models across all evaluation levels on the speechocean762 benchmark. At the phoneme level, it achieves the lowest MSE and highest PCC, indicating more accurate phoneme-level scoring achieved. Furthermore, most word-level assessment performance is boosted, with notable gains in stress prediction and total accuracy over JAM [10]. For the utterance-level assessment, JCAPT reports higher completeness, fluency, and prosody scores, reflecting enhanced modeling of global speech characteristics. In MDD, our model significantly boosts recall and F1-score, demonstrating superior detection of mispronunciations. These results confirm that integrating phonological features, think token mechanisms, and the Mamba architecture yields more accurate and interpretable CAPT performance.

3.4. Ablation Studies

To evaluate the contribution of each module, we conducted ablation studies under three settings: (1) removing phonological features, (2) removing the think tokens, and (3) removing both. As shown in Tables 2 and 3, the full model consistently outper-forms all ablated versions in both the APA and MDD tasks.

First, removing phonological features leads to notable drops in phoneme- and word-level performance, especially in MSE, PCC, and stress prediction, highlighting their importance for modeling fine-grained articulatory patterns. In contrast, utterance-level metrics remain stable, suggesting limited influence on global prosodic traits. Second, excluding the think tokens mainly affects MDD, reducing recall and F1, which implies its effectiveness in capturing disfluency-related cues. However, minor decreases in precision indicate a potential trade-off in prediction stability. Lately, when both components are removed, performance degrades across the board, suggesting that they may play complementary roles: phonological features might provide linguistic grounding, while the think tokens may enhance cognitive sensitivity in pronunciation modeling.

4. Conclusion and Future Work

In this work, we have put forward JCAPT, a unified CAPT framework that jointly addresses APA and MDD through a parallel architecture built upon the Mamba state space model. By integrating phonological features and adopting a "think token" strategy for fine-grained temporal reasoning, JCAPT enhances both diagnostic interpretability and predictive performance. Our experimental results on the speechocean762 benchmark show that JCAPT outperforms baselines, especially in mispronunciation detection and completeness—highlighting the value of joint modeling and phoneme-level reasoning in L2 pronunciation assessment. As to future work, we will explore the generalizability of our framework across diverse learner populations, languages, and spontaneous speech scenarios. In addition, we intend to implement item-specific enhancements for scoring aspects with relatively lower correlation coefficients, particularly stress and completeness, to address their current limitations and improve the overall robustness and modeling granularity of our CAPT system.

5. Acknowledgement

6. References

[1] P. Munday, "Duolingo. gamified learning through translation," *Journal of Spanish Language Teaching*, vol. 4, no. 2, pp. 194–198, 2017.

[2] A. Kholis, "Elsa speak app: Automatic speech recognition (asr) for supplementing english pronunciation skills," *Pedagogy: Journal of English Language Teaching*, vol. 9, no. 1, pp. 01–14, 2021.

[3] Y. Gong, Z. Chen, I.-H. Chu, P. Chang, and J. Glass, "Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7262–7266.

[4] B.-C. Yan, H.-W. Wang, Y.-C. Wang, J.-T. Li, C.-H. Lin, and B. Chen, "Preserving phonemic distinctions for ordinal regression: A novel loss function for automatic pronunciation assessment," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–7.

[5] B.-C. Yan, J.-T. Li, Y.-C. Wang, H.-W. Wang, T.-H. Lo, Y.-C. Hsu, W.-C. Chao, and B. Chen, "An effective pronunciation assessment approach leveraging hierarchical transformers and pre-training strategies," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 1737–1747.

[6] H.-W. Wang, B.-C. Yan, H.-S. Chiu, Y.-C. Hsu, and B. Chen, "Exploring non-autoregressive end-to-end neural modeling for english mispronunciation detection and diagnosis," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6817–6821.

[7] W. Ye, S. Mao, F. Soong *et al.*, "An approach to mispronunciation detection and diagnosis with acoustic, phonetic and linguistic (apl) embeddings," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6827–6831.

[8] B.-C. Yan, H.-W. Wang, and B. Chen, "Peppanet: Effective mispronunciation detection and diagnosis leveraging phonetic, phonological, and acoustic cues," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 1045–1051.

[9] H. Ryu, S. Kim, and M. Chung, "A joint model for pronunciation assessment and mispronunciation detection and diagnosis with multi-task learning," *INTERSPEECH*, 2023, conference paper.

[10] Y. Y. He, B. C. Yan, T. H. Lo, M. S. Lin, Y. C. Hsu, and B. Chen, "Jam: A unified neural architecture for joint multi-granularity pronunciation assessment and phone-level mispronunciation detection and diagnosis towards a comprehensive capt system," in *Proceedings of the 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, Dec. 2024, pp. 1–6.

[11] F.-A. Chao and B. Chen, "Towards efficient and multifaceted computer-assisted pronunciation training leveraging hierarchical selective state space model and decoupled cross-entropy loss," *arXiv preprint arXiv:2502.07575*, 2025.

[12] M. Shahin and B. Ahmed, "Phonological-level mispronunciation detection and diagnosis," *Interspeech 2024*, Sep. 2024, presented at Interspeech 2024, 1–5 September 2024, Kos, Greece.

[13] B.-C. Yan, H.-W. Wang, Y.-C. Wang, and B. Chen, "Effective graph-based modeling of articulation traits for mispronunciation detection and diagnosis," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[14] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[15] T. J. Yang, A. Rosenberg, and B. Ramabhadran, "Contemplative mechanism for speech recognition: Speech encoders can think," *Proceedings of Interspeech*, pp. 3455–3459, 2024.

[16] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[17] M. Yoshimura, T. Hayashi, and Y. Maeda, "Mambapeft: Exploring parameter-efficient fine-tuning for mamba," *arXiv preprint arXiv:2411.03855*, 2024.

[18] B.-C. Yan, H.-W. Wang, Y.-C. Wang, J.-T. Li, C.-H. Lin, and B. Chen, "Preserving phonemic distinctions for ordinal regression: A novel loss function for automatic pronunciation assessment," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–7.

[19] 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.

[20] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, vol. 67, pp. 154–166, 2015.

[21] J. Shi, N. Huo, and Q. Jin, "Context-aware goodness of pronunciation for computer-assisted pronunciation training," *arXiv preprint arXiv:2008.08647*, 2020.

[22] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[23] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[24] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, J. Wu, X. Xiao, L. Zhou, C. Li, S. Ren, Y. Zhang, F. Yu, Q. Fu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[25] X. Jiang, C. Han, and N. Mesgarani, "Dual-path mamba: Short and long-term bidirectional selective structured state space models for speech separation," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[26] J. Zhang, Z. Zhang, Y. Wang *et al.*, "Speechocean762: An open-source non-native english speech corpus for pronunciation assessment," *arXiv preprint arXiv:2104.01378*, 2021.

[27] R. Weide, "The carnegie mellon pronouncing dictionary [cmudict. 0.6]," *Pittsburgh, PA: Carnegie Mellon University*, 2005.

[28] K. Li, X. Qian, and H. Meng, "Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 193–207, 2017.