JAM: A Unified Neural Architecture for Joint Multi-granularity Pronunciation Assessment and Phone-level Mispronunciation Detection and Diagnosis Towards a Comprehensive CAPT System

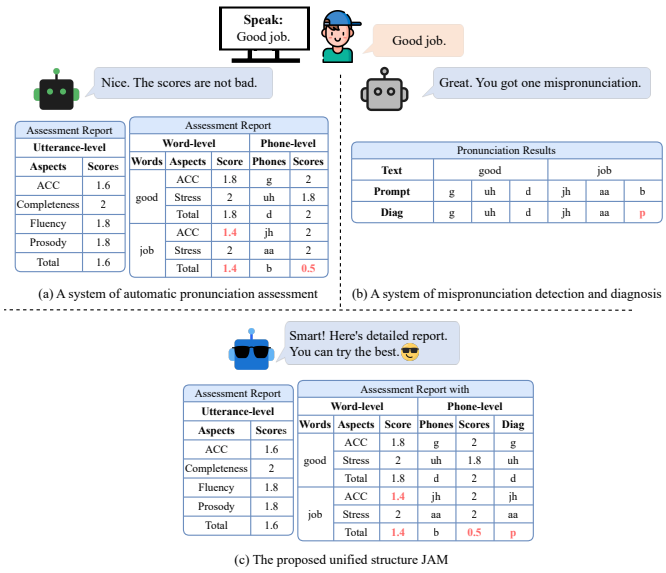
Yue-Yang He[1], Bi-Cheng Yan[1], Tien-Hong Lo[1], Meng-Shin Lin[1], Yung-Chang Hsu[2], Berlin Chen[1]
[1]National Taiwan Normal University, Taipei, Taiwan
[2]EZAI, Taiwan
E-mail: {yueyanghe, bicheng, teinhonglo, 61147077s, berlin}@ntnu.edu.tw; mic@ez-ai.com.tw

***Abstract*—Computer-assisted pronunciation training (CAPT) systems are designed for second-language (L2) learners to practice their pronunciation skills by offering objective, personalized feedback in a stress-free, self-directed learning scenario. Both mispronunciation detection and diagnosis (MDD) and automatic pronunciation assessment (APA) are indispensable components in the CAPT systems. The former is responsible for pinpointing phonetic pronunciation errors and providing the corresponding diagnostic feedback, while the latter manages to evaluate oral skills across various linguistic levels with disparate aspects. Most existing efforts typically treat APA and MDD as independent tasks, where the correlations between the assessment scores and the phonetic pronunciation errors are nearly sidelined. In light of this, we introduce JAM (a Joint neural model for APA and MDD), a novel end-to-end neural model for CAPT that streamlines the components of APA and MDD into a unified structure with a parallel pronunciation modeling architecture. To capture fine-grained pronunciation cues from L2 learners' speech, electromagnetic articulography (EMA) features are introduced for the proposed model, which portrays the movement of articulatory structures, such as the jaw, lips, and tongue. A series of experiments conducted on the speechocean762 benchmark dataset demonstrate the feasibility and effectiveness of our approach compared to several competitive baselines. Additionally, an ablation study is presented to assess the contributions of different input features and training strategies in the proposed model.**

I. Introduction

In the tide of globalization, a growing number of people are willing or being demanded to learn foreign languages. This surging need fuels the development of computer-assisted pronunciation training (CAPT) systems, which open up new possibilities for L2 (second language) to practice pronunciation skills in a stress-free and self-directed learning scenario. Beyond the immense importance in the education and learning domain, CAPT systems have achieved remarkable success across various commercial platforms and assessment services, such as Duolingo [1] and Elsa speak [2].

A de-facto archetype system for CAPT is typically instantiated with a "reading-aloud" learning scenario, where an L2 learner is presented with a text prompt and instructed to pronounce it correctly. By working in conjunction with the input speech and the presented text prompt, CAPT systems can access the learner's speaking proficiency and immediately provide instructive diagnostic feedback [3][4][5]. To offer diagnostic feedback from different learning dimensions, as decipted in Figure 1(a) and 1(b), CAPT systems consist of two indispensable components: one is mispronunciation detection and diagnosis (MDD), and the other is automatic pronunciation assessment (APA). The former aims to pinpoint erroneous pronunciation segments at phone-level and provide the corresponding diagnostic feedback [6][7][8]. The latter, in contrast, draws attention on assessing and delivering various pronunciation scores to reflect the learner's pronunciation quality on some pronunciation aspects at various linguistic granularities [9][10][11]. Most existing efforts typically treat



Assessment Report

Utterance-level	
Aspects	Scores
ACC	1.6
Completeness	2
Fluency	1.8
Prosody	1.8
Total	1.6

Assessment Report

Word-level			Phone-level	
Words	Aspects	Score	Phones	Scores
good	ACC	1.8	g	2
	Stress	2	uh	1.8
	Total	1.8	d	2
job	ACC	1.4	jh	2
	Stress	2	aa	2
	Total	1.4	b	0.5

Pronunciation Results

Text	good			job		
Prompt	g	uh	d	jh	aa	b
Diag	g	uh	d	jh	aa	p

(a) A system of automatic pronunciation assessment

(b) A system of mispronunciation detection and diagnosis

Assessment Report

Utterance-level	
Aspects	Scores
ACC	1.6
Completeness	2
Fluency	1.8
Prosody	1.8
Total	1.6

Assessment Report with

Word-level			Phone-level		
Words	Aspects	Score	Phones	Scores	Diag
good	ACC	1.8	g	2	g
	Stress	2	uh	1.8	uh
	Total	1.8	d	2	d
job	ACC	1.4	jh	2	jh
	Stress	2	aa	2	aa
	Total	1.4	b	0.5	p

(c) The proposed unified structure JAM

Fig. 1. A running example illustrates the two essential components of a CAPT system in a reading-aloud learning scenario. We present (a) phone-level mispronunciation detection and diagnosis, (b) automatic pronunciation assessment, and (c) the proposed unified structure, JAM. In this figure, '**Diag**' and '**Prompt**' represent the diagnosis result and the phone-level prompt sequence, respectively.

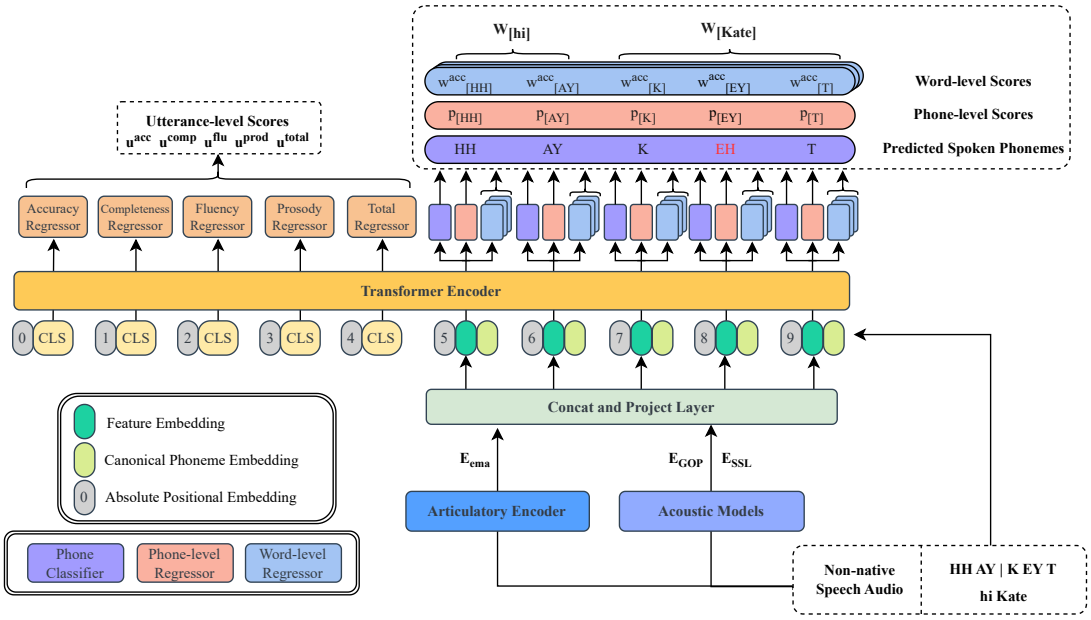


Fig. 2. Illustration of the proposed architecture of our APA method with phone-level mispronunciation detection and diagnosis. The red **'EH'** in the predicted phonemes sequence indicates that the speaker's pronunciation of this phoneme was diagnosed as a mispronunciation.

APA and MDD as independent tasks, where the correlations between the assessment scores and the phone-level pronunciation errors are nearly sidelined. However, empirical studies have revealed the negative correlations between assessment scores and the number of pronunciation errors across various aspects at diverse linguistic levels. For instance, phone-level pronunciation errors are closely related not only to utterance-level assessments, such as intelligibility [12] and fluency [13] in L2 English, but also to word-level assessments, such as stress [14] and intonation [15] in L2 Mandarin. In the follow-up work, Ryu et al. [16] ventured into integrating the tasks of MDD with APA, where the model architecture capitalized on a large-scale pre-trained speech encoder (i.e., Hubert), and then jointly optimized phone recognition and utterance-level assessment through the Connectionist Temporal Classification (CTC) [17] and cross-entropy objectives.

Towards a full-fledged CAPT system, as illustrated in Figure 1(c), this paper presents a novel end-to-end neural model, dubbed JAM (a **J**oint neural model for **A**PA and **M**DD), which leverages a unified architecture to streamline the phone dictation process and pronunciation assessments across multiple aspects (e.g., accuracy, stress, fluency) at various linguistic units (i.e., phone, word, and utterance). Specifically, JAM builds upon an iconic APA model, the goodness of pronunciation feature-based transformer (GOPT), and further integrates a phone-level predictor. When assessing a learner's speech, our model first creates timestamps for each phone segment in the phone-level text prompt using the Viterbi algorithm to align the text prompt with the learner's speech. Various pronunciation features are then extracted for each phone-level timestamp, and fed into a stack of transformer blocks to form contextualized pronunciation features. Based on these contextualized features, a predictor network generates a sequence of phonetic diagnostic feedback, while disparate regressors are employed to derive corresponding pronunciation scores from the phone-level to the utterance-level. To facilitate the modeling of fine-grained pronunciation cues and strengthen MDD performance, the proposed model incorporates electromagnetic articulography (EMA) features, which captures the movement of articulatory structures, such as the jaw, lips, and tongue. Comprehensive experiments conducted on the speechocean762 benchmark dataset show that our proposed method achieves significant and consistent improvements over several cutting-edge baselines. Additionally, an ablation study is conducted to examine the impact of different input features and training strategies on the performance of the proposed model.

II. METHODOLOGY

A. Input Features

The complete workflow is illustrated in Figure 2. We build our model based on GOPT, which processes an input utterance by first aligning the audio signals X with the corresponding text prompt T. This alignment step yields a sequence of phoneme-level goodness of pronunciation (GOP) features, denoted as $Emb_{GOP} = (Emb_{[p_1]}, Emb_{[p_2]}, ..., Emb_{[p_L]})$. Each individual GOP feature $Emb[p_l]$ is derived from a combination of log phone posterior (LPP) [19] and log posterior ratio (LPR) [20]. The resulting Emb_{GOP} is a matrix of size $84 \times L$, where L represents the length of the canonical phoneme sequence.

To enrich our model's input, we leverage pre-trained acoustic models (wav2vec 2.0 [21], Hubert [22], and WavLM [23]) to extract comprehensive acoustic features. Additionally, we

TABLE I
EXPERIMENTAL RESULTS OF DIFFERENT METHODS EVALUATED ON SPEECHOCEAN762.

Models		Phone-level Score		Word-level Score (PCC)			Utterance-level Score (PCC)					MDD Performance Score		
		MSE ↓	PCC ↑	ACC ↑	Stress ↑	Total ↑	ACC ↑	Completeness ↑	Fluency ↑	Prosody ↑	Total ↑	RE (%)	PR (%)	F1 (%)
APA methods	GOPT [9]	0.085 ±0.001	0.612 ±0.003	0.533 ±0.004	0.291 ±0.030	0.549 ±0.004	0.714 ±0.004	0.155 ±0.039	0.753 ±0.008	0.760 ±0.006	0.742 ±0.005	-	-	-
	HiPAMA [18]	0.084 ±0.001	0.616 ±0.004	0.575 ±0.004	**0.320 ±0.021**	0.591 ±0.004	0.730 ±0.002	0.276 ±0.177	0.749 ±0.001	0.751 ±0.002	0.754 ±0.002	-	-	-
	3M [3]	0.078 ±0.001	0.656 ±0.005	0.598 ±0.005	0.289 ±0.033	0.617 ±0.002	0.760 ±0.004	**0.325 ±0.141**	0.828 ±0.006	0.827 ±0.008	0.796 ±0.008	-	-	-
APA & MDD methods	Joint-CAPT-L1 [16]	-	-	-	-	-	0.719	-	0.775	0.773	0.743	**91.40**	26.70	41.40
	JAM	**0.076 ±0.002**	**0.664 ±0.001**	**0.622 ±0.012**	0.241 ±0.034	**0.638 ±0.005**	**0.773 ±0.007**	0.205 ±0.080	**0.831 ±0.004**	**0.829 ±0.004**	**0.805 ±0.004**	34.76	**64.10**	**45.01**

employ a pre-trained articulatory encoder [24] to obtain deep, EMA-like features from the last hidden layer, providing essential articulatory information for pronunciation assessment. All features are concatenated and projected into a 24-dimensional embedding Emb_{proj}.

B. Transformer-based Architecture

The projected acoustic features Emb_{proj}, canonical phone embeddings, and absolute positional embeddings are fed into a Transformer encoder. To obtain five aspects of utterance-level pronunciation scores in GOPT, five trainable [CLS] tokens are integrated with the phone-level input sequence, inspired by BERT [25].

Phoneme- and word-level regressors are added on top of the Transformer output of each corresponding phoneme. In addition, we use a classifier to predict each reference phoneme that the speaker pronounces.

C. Optimization

Our network is optimized using a Multi-Task Learning (MTL) framework. We employ Mean Squared Error (MSE) as the loss function for each granularity. The combined loss, denoted as $\mathcal{L}_{\text{APA}}$, is computed by directly summing the averaged losses at the utterance, word, and phoneme levels:

$$\mathcal{L}_{\text{APA}} = \mathcal{L}_{utterance} + \mathcal{L}_{word} + \mathcal{L}_{phoneme}, \tag{1}$$

where $\mathcal{L}_{utterance}$ and $\mathcal{L}_{word}$ are averaged utterance and word level losses of five utterance-level labels and three word-level labels, respectively; $\mathcal{L}_{phoneme}$ is the phoneme loss.

For the MDD task, the phone classifier is trained to increase the predicted probability for the correct phones by minimizing the cross-entropy loss, thus improving its accuracy in detecting mispronunciations:

$$\mathcal{L}_{\text{MDD}} = -\sum_{i=1}^{N}\sum_{p=1}^{P} y_{i,p}\log(\hat{y}_{i,p}), \tag{2}$$

where N is the number of samples, P is the number of phones, $y_{i,p}$ represents the true label for phone p in sample i (usually one-hot encoded), and $\hat{y}_{i,p}$ denotes the predicted probability of phone p for sample i.

The two loss are linearly combined as the overall objective for model learning:

$$\mathcal{L}_{CAPT} = \alpha\mathcal{L}_{MDD} + (1-\alpha)\mathcal{L}_{APA}, \tag{3}$$

where α are used to balance the two losses. α is chosen from the set of $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The value of α is set to be 0.3 based on the test set.

III. EXPERIMENTS AND RESULTS

A. Dataset

To evaluate our proposed method, we utilized the publicly available speechocean762 dataset [26] for the experiments. This dataset comprises 5,000 English speech recordings produced by 250 Mandarin L2 speakers, evenly split into training and test sets of 2,500 utterances each.

A key strength of speechocean762 is its comprehensive annotation framework. It provides human-rated scores at the utterance, word, and phoneme levels, each determined by five expert raters using consistent criteria. For training efficiency, these scores were normalized to a common 0-2 scale. Additionally, the dataset offers detailed mispronunciation transcriptions using a 39-phoneme set aligned with the CMU Pronunciation Dictionary [27]. This phoneme inventory was expanded to include <del> and <unk> symbols for omitted phonemes and non-categorial or distortion-like pronunciation, respectively.

B. Experimental Setup

Building upon the experimental setup detailed in [9], we employ the same DNN-HMM acoustic model to extract 84-dimensional GOP features. This model is based on a factorized time-delay neural network (TDNN-F) and trained on the Librispeech 960-hour dataset using the Kaldi recipe. To evaluate the efficacy of our proposed multi-view approach, we keep all training hyperparameters identical to the original GOPT. We conduct five independent experiments with different random seeds and report the mean and standard deviation of the Pearson Correlation Coefficient (PCC) and Mean Squared Error (MSE). For SSL features, we utilize pre-trained wav2vec 2.0, HuBERT, and WavLM models obtained from Hugging Face. Consistent with [3], we apply random dropout with probability *p* to each SSL feature before integration to mitigate overfitting.

TABLE II
ABLATION STUDY OF THE PROPOSED METHOD ON AUTOMATIC PRONUNCIATION ASSESSMENT.

Models	Phone-level Score		Word-level Score (PCC)			Utterance-level Score (PCC)				
	MSE ↓	PCC ↑	ACC ↑	Stress ↑	Total ↑	ACC ↑	Completeness ↑	Fluency ↑	Prosody ↑	Total ↑
JAM	**0.076**	**0.664**	**0.622**	0.241	**0.638**	**0.773**	**0.205**	**0.831**	**0.829**	**0.805**
w/o EMA feat	0.078	0.650	0.609	0.196	0.624	0.769	0.047	0.823	0.821	0.804
w/o SSL feat	0.083	0.624	0.550	**0.251**	0.568	0.724	0.155	0.783	0.780	0.754
w/o SSL, EMA feat	0.083	0.622	0.546	**0.251**	0.564	0.718	0.163	0.756	0.759	0.748

TABLE III
ABLATION STUDY OF PROPOSED METHOD ON MISPRONUNCIATION DETECTION AND DIAGNOSIS.

Models	RE (%) ↑	PR (%) ↑	F1 (%) ↑	FAR (%) ↓	FRR (%) ↓	DER (%) ↓	PER (%) ↓
JAM	**34.76**	64.10	**45.01**	**64.32**	0.58	**45.23**	2.81
w/o EMA feat.	33.61	65.48	44.38	66.39	0.51	46.47	**2.78**
w/o SSL feat.	19.23	**67.55**	29.87	80.77	**0.27**	51.54	2.79
w/o SSL, EMA feat.	20.12	66.89	30.91	79.88	0.29	52.43	2.80

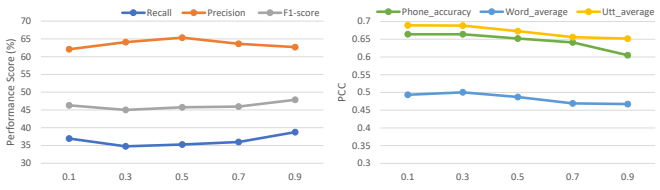


Fig. 3. Performance of different of loss weight α for MDD (left) and APA (right).

Following previous MDD work [28], the F1-score, a harmonic mean of precision (PR) and recall (RE), is commonly used as a measure of a model's accuracy. Specifically, phoneme mispronunciation detection performance is additionally evaluated by false rejection rate (FRR), false acceptance rate (FAR), and phoneme error rate (PER), as well as diagnostic error rate (DER).

C. Main Results

The performance of our framework and several state-of-the-art APA models is presented in Table I. Our proposed model JAM consistently competes with existing approaches across various evaluation metrics, demonstrating its superior ability to capture fine-grained phoneme-level errors and overall pronunciation proficiency. First, it is evident that our model consistently outperforms GOPT [9] and HiPAMA [18], which are recently proposed methods and can generalize well to the multi-aspect and multi-granular assessment, in almost metrics at the phone-level, word-level, and utterance-level, demonstrating its effectiveness on the speechocean762 dataset. Unfortunately, our proposed model exhibits poor performance in stress at the word-level and completeness at the utterance-level, likely due to data imbalance issues that need to be addressed. Secondly, while 3M [3] incorporates SSL and handcrafted features for prosody, our framework, which jointly trains APA and MDD and is augmented by EMA features, yields superior results. This underscores the benefits of integrating MDD and APA for pronunciation assessment. Finally, in the joint APA and MDD method, our proposed model JAM significantly exceeds Joint-CAPT-L1 [16] in APA performance and achieves higher F1-score. Although Joint-CAPT-L1 exhibits higher recall, its lower precision indicates a reduced ability to accurately detect correct pronunciations. Furthermore, the precision of 64% achieved by JAM indicates that our proposed model is more cautious when determining a pronunciation error. Only when the model is highly confident will it mark a pronunciation as incorrect.

D. Ablation Studies

To study the effectiveness of each feature involved in our model structure, we conducted an ablation study with the following settings: 1) removing EMA features, 2) removing self-supervised learning (SSL) features, and 3) removing EMA and SSL features simultaneously, leaving only GOP features for assessment. Table II and III shows the corresponding results. First, it can be observed that removing EMA features leads to a slight decrease in performance across most metrics in both automatic pronunciation assessment (APA) and mispronunciation detection and diagnosis (MDD). The decline indicates the importance of EMA features in capturing articulatory movements that contribute to the model's overall accuracy and effectiveness. Second, removing SSL features results in a more substantial drop in performance. This degradation is evident in the lower scores across both APA and MDD tasks. Third, when both EMA and SSL features are removed, leaving only GOP

features, the performance further declines. The results show significant decreases in various metrics, emphasizing the combined contribution of EMA and SSL features to the model's success. This highlights that relying solely on GOP features is insufficient for achieving optimal performance in both APA and MDD tasks. These findings underscore the significance of incorporating EMA and SSL features in our model to maintain high performance and accuracy in pronunciation assessment and mispronunciation detection and diagnosis.

Results on Mispronunciation Diagnosis. We further examined the diagnostic error rate (DER) in the ablation study. JAM achieved the lowest DER, indicating that EMA and SSL features enhance JAM's ability to detect the correct phoneme for erroneous pronunciation. However, JAM has a relatively lower FRR and higher FAR and PER. We trade off between FAR and FRR to increase the ability to correctly capture mispronunciations. These results indicate that the integration of EMA and SSL features is crucial for maintaining high performance and accuracy in pronunciation assessment and mispronunciation detection and diagnosis.

Effect of Loss Weight. Figure 3 illustrates the outcome of the MDD (left) and APA (right) under different loss weight α. A larger α indicates a greater loss weight assigned to the MDD. As shown in the figure, the F1-score of the MDD does not always increase monotonically with the increase of α. This suggests that simply increasing the loss weight of the MDD does not guarantee optimal performance on all metrics. In our main results, we selected $\alpha = 0.3$ as the optimal parameter, considering a trade-off between the PCC in phone accuracy and MDD effectiveness. At this setting, the model excels in phone accuracy while maintaining an acceptable performance in MDD.

IV. Conclusion

This paper presents JAM, an end-to-end neural model that integrates GOPT and a phone-level predictor, utilizing EMA features to significantly enhance speech recognition and pronunciation assessment performance. Experiments on the speechocean762 benchmark dataset show that JAM outperforms several advanced baselines, with an ablation study validating the impact of different features and strategies on its performance. In our future work, we plan to explore a hierarchical architecture and further enhance MDD performance as well as address the issues related to imbalanced data.

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] 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 *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 575–582. DOI: 10.23919/APSIPAASC55919.2022.9979979.

[4] 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.

[5] D. Korzekwa, J. Lorenzo-Trueba, T. Drugman, and B. Kostek, "Computer-assisted pronunciation training—speech synthesis is almost all you need," *Speech Communication*, vol. 142, pp. 22–33, 2022.

[6] 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. DOI: 10.1109/SLT54892.2023.10022472.

[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] D. Y. Zhang, S. Saha, and S. Campbell, "Phonetic rnn-transducer for mispronunciation diagnosis," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[9] 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.

[10] 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. DOI: 10.1109/ASRU57964.2023.10389777.

[11] W. Liu, K. Fu, X. Tian, *et al.*, "An asr-free fluency scoring approach with self-supervised learning," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[12] C. Zhu, T. Kunihara, D. Saito, N. Minematsu, and N. Nakanishi, "Automatic prediction of intelligibility of words and phonemes produced orally by japanese

learners of english," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2023, pp. 1029–1036.

[13] L. Chen, J. Tao, S. Ghaffarzadegan, and Y. Qian, "End-to-end neural network based automated speech scoring," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 6234–6238.

[14] N. F. Chen, D. Wee, R. Tong, B. Ma, and H. Li, "Large-scale characterization of non-native mandarin chinese spoken by speakers of european origin: Analysis on icall," *Speech Communication*, vol. 84, pp. 46–56, 2016.

[15] W. Li, N. F. Chen, S. M. Siniscalchi, and C.-H. 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.

[16] 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.

[17] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[18] H. Do, Y. Kim, and G. G. Lee, "Hierarchical pronunciation assessment with multi-aspect attention," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[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] 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.

[22] 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.

[23] S. Chen, C. Wang, Z. Chen, *et al.*, "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.

[24] Y. M. Siriwardena, G. Sivaraman, and C. Espy-Wilson, "Acoustic-to-articulatory speech inversion with multitask learning," *arXiv preprint arXiv:2205.13755*, 2022.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[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 *et al.*, "The carnegie mellon pronouncing dictionary," *release 0.6, www. cs. cmu. edu*, 1998.

[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. DOI: 10.1109/TASLP.2016.2621675.