IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 February 2023, accepted 6 March 2023, date of publication 10 March 2023, date of current version 24 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3255986

METHODS

A Multifaceted Approach to Oral Assessment Based on the Conformer Architecture

ZHIXING FAN[1,4], **JING LI**[2,4], **AISHAN WUMAIER**[2,4], **(Member, IEEE), ZAOKERE KADEER**[2,4], **AND ABDUJELIL ABDURAHMAN**[3,4]

[1]College of Software, Xinjiang University, Urumqi, Xinjiang 830046, China
[2]College of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang 830046, China
[3]College of Mathematics and System Sciences, Xinjiang University, Urumqi, Xinjiang 830046, China
[4]Xinjiang Key Laboratory of Multilingual Information Technology, Urumqi, Xinjiang 830046, China

Corresponding authors: Zaokere Kadeer (zuhra@xju.edu.cn) and Aishan Wumaier (hasan1479@xju.edu.cn)

This work was supported in part by the Beijing National Research Center for Information Science and Technology through the Open Topic Project "Research on Automatic Evaluation Technology of the Topic Speaking Section of Putonghua Proficiency Test" under Grant BNR2021KF02005, in part by the Xinjiang University 2022 National Undergraduate Innovation Training Program Project under Grant 202210755113, and in part by the Basic Research Program of Tianshan Talent Plan of Xinjiang under Grant 2022TSYCJU0005.

ABSTRACT Automatic speaking assessment methods are essential for helping non-native learners to learn native pronunciation. The automated speaking assessment method consists of mispronunciation detection and pronunciation quality assessment. In the past, researchers have usually focused their research on only one specific aspect of the speaking assessment task. Research on multifaceted speaking tasks has been rare, and model building has often led to reduced performance due to the omission of local feature details. In this paper, we propose a multi-width band (MB) method and apply it to the Conformer model. This method can effectively increase the ability of the model to obtain local feature information at different scales. At the same time, we used a multi-task learning approach to train a multifaceted speaking assessment model based on GOP features. We conducted experiments on a self-built monosyllabic Mandarin mispronunciation detection dataset (PSC-MonoSyllable) and an English open-source pronunciation quality assessment dataset (SpeechOcean762), respectively. The experimental results show that the method's mispronunciation detection metrics in terms of phonemes, tones, and words on the PSC-MonoSyllable dataset (F1 scores) reached 70.18%, 80.06%, and 79.82%, respectively. The results of the method on the SpeechOcean 762 dataset for the pronunciation quality assessment task also showed a certain degree of improvement in all aspects of the phoneme- and grapheme-level correlation metrics compared with the baseline model.

INDEX TERMS Computer assisted pronunciation training, mispronunciation detection, assessment of pronunciation quality, Conformer, dilation convolution.

I. INTRODUCTION

Automated speaking assessment is an essential autonomous language learning technology [1], [2] that facilitates the learning of native pronunciation (L1) by non-native speakers (L2). Compared with a traditional classroom, automatic speaking assessment is more economical and convenient, and also allows language learners to receive timely feedback on their pronunciation. Due to its usefulness, automatic speaking assessment has been extensively researched, with most of the work focusing on mispronunciation detection and pronunciation quality assessment for a single aspect, e.g., [3], [4], [5], [6], [7], [8], [9]. However, speaking assessment tasks include many other aspects, such as mispronunciation detection for phonemes (initials and finals), tones, and words; pronunciation standardization for phonemes and words; and pronunciation quality assessment for fluency and completeness of discourse. However, there is a relationship between different aspects of different tasks in speaking assessment, and jointly modeling the different aspects may lead to a more comprehensive representation of the speaking assessment model. In reality, we also want to use a single model to

The associate editor coordinating the review of this manuscript and approving it for publication was Mounim A. El Yacoubi.

carry out different aspects of the speaking assessment task simultaneously.

In this paper, based on the Goodness of Pronunciation (GOP) feature and the Conformer model architecture, we proposed a new model, called Conformer-MB, to study mispronunciation detection and pronunciation quality assessment tasks in spoken language assessment. We conduct multifaceted pronunciation quality assessment and mispronunciation detection experiments on the publicly available SpeechOcean762 dataset and the self-built PSC-MonoSyllable dataset. The SpeechOcean762 dataset contains one phoneme-level, three word-level, and five utterance-level labels, including accuracy, completeness, and fluency. The PSC-MonoSyllable dataset included speakers from universities in the Xinjiang region, and the speakers' accents are predominantly Lan Yin Guanhua and Zhongyuan Guanhua. The content of each audio in the dataset is a single Chinese character marked for pronunciation correctness by three Mandarin assessment experts in terms of phonemes, tones, and words. In the training process, we used the different labels in the above dataset for multi-task training. Furthermore, to improve the performance of the evaluation, based on the Conformer model architecture, we use several dilated convolutional networks with different dilation rates to obtain local feature information at different scales. This allows the model to acquire richer local feature information and maintain a focus on global information. In addition, as the features extracted through the encoder become increasingly abstract during the encoder accumulation process, some of the detailed information of the original features will inevitably be lost. To address this issue, we improve the computation of the attention mechanism in the encoder by using the residual structure to retain the original feature information. In summary, we propose a multifaceted speaking evaluation method based on the Conformer architecture by fully considering the problems in the current research on speaking evaluation. The evaluation assessment results obtained from experiments on different tasks with different datasets show significant improvements in various evaluation metrics.

II. RELATED WORKS

The GOP method and its variants proposed by [10] are classical speech-recognition-based spoken language evaluation methods. The GOP algorithm is based on forced alignment and the logarithmic posterior probability of a phoneme. The result is a response to the confidence level between the actual pronunciation of the phoneme and the standard pronunciation. Reference [11] proposed an improved GOP algorithm considering the HMM transfer probability in the DNN-HMM acoustic model of the enhanced GOP algorithm. After the emergence of deep learning algorithms, some improvements in GOP algorithms also used in deep learning models. Reference [12] proposed the context-dependent CaGOP algorithm, which predicts the duration of each phoneme by feeding the reference text into a self-attentive text-based encoder during GOP calculation, and uses the difference between the expected duration and the actual duration of the phoneme obtained by forced alignment as the penalty factor in the GOP calculation. In addition to the studies on spoken language evaluation tasks using GOP algorithms, there are also studies on verbal language evaluation conducted without GOP algorithms, such as wav2vec2.0-based [13], [14], [15] and deep learning feature-based methods [16], but because of the limited speech data available to L2 speakers, training with such methods usually requires the use of pre-trained models and transfer learning.

Oral language assessment studies are usually divided into automatic mispronunciation detection and automatic pronunciation quality assessment according to the task's objectives. However, in previous studies, only one aspect is generally targeted, e.g., phonemes, tones, Erhua, and words in Chinese; and phonemes, words, utterance, etc., in English. For example, in the mispronunciation detection task, [3] analyzed the pronunciation and position of each phoneme, classified the phoneme segments according to their GOP values, and trained them separately. Then, they trained an SVM classifier for each phoneme for mispronunciation detection to improve the system's ability to discriminate pronunciation quality. Studies such as [17], [18], [19], [20] perform mispronunciation detection in terms of phonemes by building speech recognition models at the phoneme level. References [4] and [5] proposed an algorithm for Chinese Mandarin tone evaluation based on the improved Fujisaki model according to Chinese Mandarin tone pronunciation characteristics, respectively. Reference [6] evaluated Mandarin tone by fusing the MSD-HMM-based embedded tone model with the GMM-based explicit tone model. This model also included an automatic evaluation of the Erhua aspects of Mandarin Chinese. Reference [7] proposed an Erhua evaluation model based on an integrated classification regression tree and discussed in detail the problem of modeling Erhua cut-offs and extracting relevant acoustic features. The above mispronunciation detection methods only model a specific aspect of Chinese or English and do not reflect the overall pronunciation of the speaker. There have only been a few previous efforts on multi-granularity pronunciation assessment [8], [9]. In these works, however, only a single score is considered for each granularity. Reference [21] first proposed a multi-granularity pronunciation quality assessment method using a single model. They used the open-source SpeechOcean762 dataset, which contains one phoneme-level, three word-level, and five utterance-level labels, including accuracy, intonation, and fluency. They performed multi-task training on the Transformer model architecture using pronunciation-quality-based features. Not only are multiple aspects of pronunciation quality evaluated, but performance is also improved in each evaluation task.

The vanilla Transformer model also has certain drawbacks, such as its inability to capture details of local features. Reference [22] proposed a Conformer model structure that combines a convolutional neural network, which is good at extracting local features, with the Transformer model, significantly improving this drawback of the Transformer model. In the speech recognition task scenario, the model constructed using the Conformer network structure was considerably enhanced compared with the Transformer model. However, the Conformer model uses a convolutional network structure with a limited ability to acquire local features, and does not improve the evaluation results in the GOP feature-based spoken language evaluation task. Our work builds a multifaceted oral assessment model based on the Conformer model architecture, using pronunciation quality features and a multitask training approach. At the same time, we further improve the convolutional module in the Conformer model. We take advantage of the fact that dilation convolution can expand the perceptual field and capture multi-scale contextual information to obtain richer feature information and improve the model's performance in different aspects.

III. METHODOLOGIES

A. CONFORMER-MB MODEL

Since the standard acoustic model uses MFCCs features as input, we first extract the MFCCs features of the audio to be evaluated. After that, as shown in Equation (1), we input the audio feature sequences (MFCCs) prepared for evaluation with the corresponding reference texts into the previously trained acoustic model. The input audio feature sequences are converted into articulatory goodness features $F_{\text{lpp+lpr}}$ according to a specific algorithm (to be described in Section III-B).

$$F_{\text{lpp+lpr}} = \text{Acoustic}(X, C) \tag{1}$$

where $X = (x_1, x_2, \ldots, x_T)$ is the input sequence and $C = (c_1, c_2, \ldots, c_U)$ is the corresponding reference text. The length of the input sequence X is T, the size of C is U, and the length of the output sequence $F_{\text{lpp+lpr}}$ is $2 \times U$. As shown in Equation (2), we use a linear layer to map the pronunciation goodness feature dimension to the same dimension, embedding_dim, as that of the text embedding layer, resulting in an output of $H_f \in R^{U \times \text{embedding\_dim}}$.

$$H_f = \text{Gop\_projection\_Layer}(F) \tag{2}$$

Since different phonemes have different textual features that can provide the model with usable textual information [16]. We use the embedding function in PyTorch to encode the different phonemes in the audio text sequence to obtain the text content features and encode the position of the phonemes to get the position of different phonemes in different text sequences. As shown in Equation (3), we encoded the reference textual content corresponding to the prepared evaluation audio in the embedding layer $H_c \in R^{U \times \text{embedding\_dim}}$.

$$H_c = \text{Phoneme\_Embedding}(C) \tag{3}$$

Reference [23] proposed an algorithm called "truncated normal distribution". As shown in Equation (4), while encoding the text content, we use the truncated normal distribution (Truncated_Normal) to initialize the position parameters of the reference text sequence, resulting in a text position encoding result $H_p \in R^{U \times \text{embedding\_dim}}$.

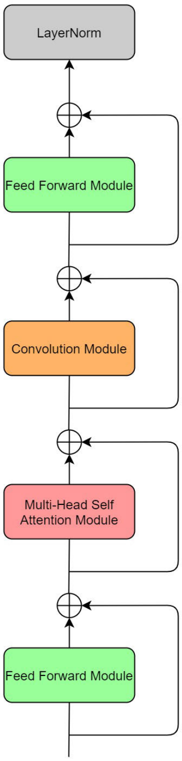


FIGURE 1. Conformer block structure [22].

$$H_p = \text{Truncated\_Normal}(\text{Positional\_Embedding}(C)) \tag{4}$$

Afterward, we add the above results to be used as input to the Conformer-MB encoder together, as shown in Equation (5).

$$H = \text{Add}(H_f, H_c, H_p) \tag{5}$$

The Conformer encoder module consists of two feed-forward neural networks, a multi-headed self-attentive mechanism module, and a convolutional module, each connected with residuals. Figure 1 illustrates the overall structure of the standard Conformer encoder module. The internal structure of the Conformer encoder module resembles a macaron structure, i.e., it consists of two identical feed-forward neural networks interspersed with a multi-headed self-attentive mechanism module and a convolutional module [22].

Compared with the structure of the Transformer encoder, the Conformer encoder is characterized by two main features: 1. the addition of a convolutional module and 2. the division of the feed-forward module into two parts. As convolutional neural networks are characterized by being good at extracting local features and weak at acquiring global representations, the Transformer model reflects the complex spatial transformations and long-range feature dependencies that constitute the global representation. The Conformer model blends the features of both, embedding local features and global representations precisely into each other, and achieves better performance in speech recognition tasks. The structure of BottleNecks consists of three layers of convolutional networks. The first layer of convolutional networks is a separable convolution with a convolutional kernel size of 1 × 1, aiming at the dimensionality reduction of the input features; the second layer of convolutional networks is a deep convolution with a convolutional kernel size of 3 × 3, aiming at deep feature extraction; the third layer of convolutional networks is a separable convolution with a convolutional kernel size of 1 × 1, aiming at dimensionality recovery. This structure is designed to reduce the computational complexity and the number of parameters while reducing the training time. However, such structures have a limited ability to obtain local feature information and a weaker ability to obtain multi-scale feature information.

Because GOP features reflect the interrelationship between the target phoneme and other phonemes, inter-feature combinations at multiple scales can better reflect the actual quality of the speaker's pronunciation. To further enhance the representation of local features, we propose a multi-width band method based on the Conformer structure, which uses convolutional kernels of the same size to capture local information at different dilation rates and then adds them together to obtain richer feature information. The so-called multi-scale band uses multiple parallel convolutional layers simultaneously to obtain different feature information under different perceptual fields. Then combine them to compensate for the limited feature information obtained using a single convolutional layer. We use the features of the Conformer structure and combine them with the Conformer model structure to build a new encoder structure, which can enhance the local information acquisition ability of the encoder while focusing on the global information. In addition, during the stacking of multiple Conformer-MB modules, the features extracted by the later modules become more and more abstract, which can lead to the loss of some of the lower-level feature information and result in reduced detection effectiveness. A cross-module strategy is added to the different Conformer-MB modules to fully retain some of the low-level fine-grained features. Specifically, the result of the calculation of the attention score in the current module is composed of two parts. The first part is the attention score in the current module multiplied by the weight coefficient, and the second part is the result of the calculation of the attention score in the previous module. As shown in the following equations ((6)-(9)), where H_i is the vector output from the feed-forward module FFN, $0 \leq \alpha \leq 1$.

$$Q_i = W_q \times H_i \tag{6}$$

$$K_i = W_k \times H_i \tag{7}$$

$$V_i = W_v \times H_i \tag{8}$$

$$\begin{aligned}\mathrm{MHSA}(Q_i, K_i, V_i) = {} & \mathrm{MHSA}(Q_{i-1}, K_{i-1}, V_{i-1}) \\ & + (1-\alpha) \times \mathrm{MHSA}(Q_i, K_i, V_i)\end{aligned} \tag{9}$$

The details of the Conformer-MB encoder are shown in Figure 2. First, the Conformer-MB encoder obtains the input feature sequence H_{i-1}, and when $i = 0$, $H_0 = H$.

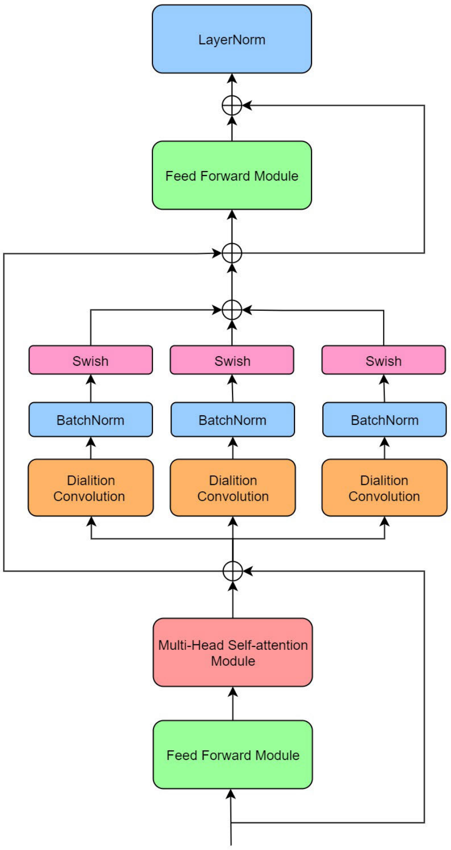


FIGURE 2. Conformer-MB encoder structure.

The H_{i-1} is passed through the feed-forward network FFN_1 of the standard Conformer model with the multi-headed attention mechanism MHSA to obtain the attention context vector H_i. We use a residual structure to retain the original feature information in this process. The final output vector obtained by this module is H_a. The specific calculation process is given in Equation (10) and Equation (11)

$$H_i = \mathrm{MHSA}(\mathrm{FFN}_1(H_{i-1})) \tag{10}$$

$$H_a = \mathrm{ResNet}(H_i, H_{i-1}) \tag{11}$$

A dilated convolutional neural network can increase the receptive field of the convolutional kernels while keeping the number of parameters constant so that the output of each convolutional kernel contains an extensive range of information. We then used three parallel dilated convolutional neural networks $\mathrm{Dilation}_{i\in1,2,3}$ to extract the input feature information at different scale sizes $D_{i\in1,2,3}$. After this, we add a BatchNorm layer and a Swish activation function [24] to speed up the training and convergence of the model. Finally, the outputs of the different inflated convolutional networks are added together, and the output vector D is obtained using the residual structure.

$$D_{i\in1,2,3} = \mathrm{Dilation}_i(H_a) \tag{12}$$

$$D_{i\in1,2,3} = \mathrm{Swish}(\mathrm{BatchNorm}(D_{i\in1,2,3})) \tag{13}$$

$$D = \mathrm{ResNet}(\mathrm{Add}(D_{i\in1,2,3}), H_a) \tag{14}$$

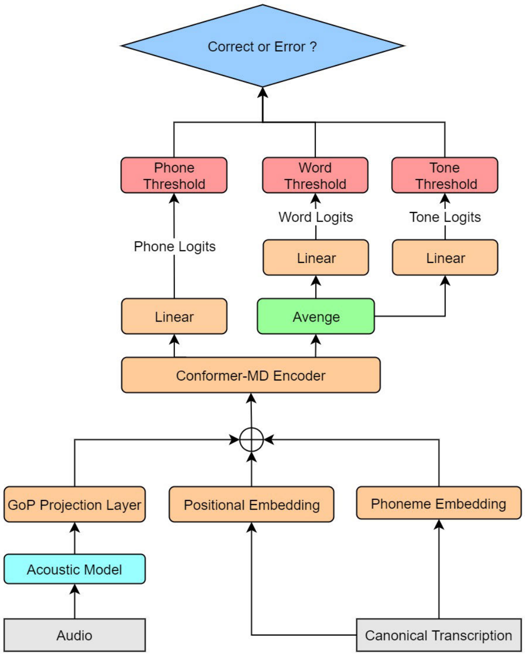


FIGURE 3. Flow chart of the Conformer-MB model under the mispronunciation detection task.

Finally, we feed D into the second network of feed-forward networks with a residual structure and use the LayerNorm function to conduct a normalization operation to obtain the final output O of the Conformer-MB encoder.

$$O = \text{LayerNorm}(\text{ResNet}(\text{FFN}_2(D), D)) \tag{15}$$

As we use the audio content of the dataset in our mispronunciation detection experiments, the audio content of the dataset is a single Chinese character, with each character consisting of an initial, a final, and a tone. Once we obtain the encoder output, we need to average the output results for the word and tone aspects as the encoding results. Afterward, the coding results of the different aspects are fed into the corresponding linear layers as well as the Sigmod function to obtain the final probability of correct pronunciation $P(\text{phoneme})$, $P(\text{word})$, $P(\text{tone})$. This part of the calculation process is given in Equation ((16)-(19)). Figure 3. shows the flow of the Conformer-MB model for the mispronunciation detection task.

$$P(\text{phoneme}) = \text{Sigmod}(\text{Linear}_{\text{phoneme}}(D)) \tag{16}$$
$$D_{\text{avg}} = \text{Avenge}(D) \tag{17}$$
$$P(\text{word}) = \text{Sigmod}(\text{Linear}_{\text{word}}(D_{\text{avg}})) \tag{18}$$
$$P(\text{tone}) = \text{Sigmod}(\text{Linear}_{\text{tone}}(D_{\text{avg}})) \tag{19}$$

We use a multi-task training approach in the model training process; specifically, a loss function is used for each aspect of the error detection task. [21] first average the losses of each granularity and then sum them up with the same weight, where $L_{\text{utterance}}$ and L_{word} are averaged utterance and word level losses of five utterance-level labels and three word-level labels, respectively. The loss values for the word and sentence tasks in [21] are obtained by averaging the loss values of their subtasks. For example, the loss value of a sentence is obtained by averaging the loss values of its five subtasks (prosodic, fluency, accuracy, completeness, and total). In this paper, the subtasks under each task are all mispronunciation detection and do not require averaging. We use the same weights (we set the weights of all losses to 1 in the experiment) to sum the losses of each aspect and obtain a unique Loss for back-propagation, i.e., $\text{Loss} = \text{Loss}_{\text{phoneme}} + \text{Loss}_{\text{word}} + \text{Loss}_{\text{tone}}$, keeping them consistent with the loss weights for each task in [21]. Reference [21] mentioned that other loss weighting methods might have worked better, but this paper proposes a new multi-scale fusion feature method for a spoken language evaluation. Hence, a more straightforward loss weighting method is convenient for comparing different models. In the model inference, we classify the probability of the correct pronunciation of the model output by setting a threshold, i.e., we consider the pronunciation correct if it exceeds the threshold and incorrect if it falls below the threshold. Except for the acoustic model, the entire network structure of the model is based on an end-to-end training approach.

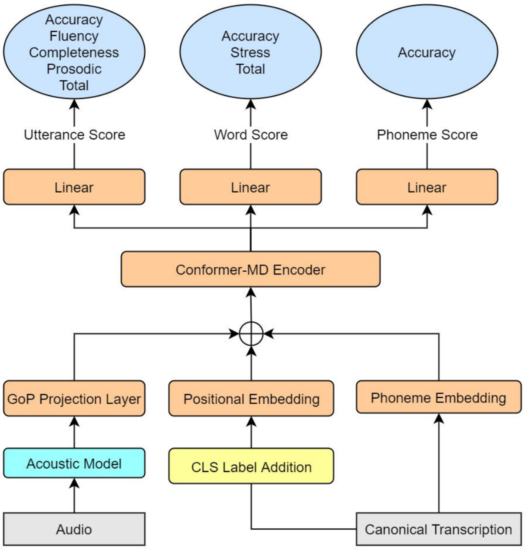


FIGURE 4. Flow chart of the Conformer-MB model for the pronunciation quality assessment task.

The flow diagram of the Conformer-MB model under the pronunciation quality assessment task is shown in Figure 4. We set five trainable CLS tokens (as shown in Equation (20)) in the input sequence of acoustic features, each corresponding to a discourse-level label. In contrast to the mispronunciation detection task, we obtained the scores corresponding to the different tags directly from the linear layer (regression task). To make a fair comparison, we used the same training method, loss function, and parameter configuration as in the original paper [21].

$$C = [C, CLS_1, CLS_2, CLS_3, CLS_4, CLS_5] \tag{20}$$

B. THE GOODNESS OF PRONUNCIATION FEATURES

In the pronunciation quality assessment task, we conducted comparative experiments on model validity using publicly available data provided by [21]. In the mispronunciation detection experiments, we train a model of standard acoustics by using the 863 Chinese speech dataset. The acoustic features required for mispronunciation detection are extracted based on this acoustic model, because acoustic models trained with both L1 and L2 speech data can produce better alignment to L2 speech data and output more accurate GOP features [25]. We used an acoustic model configuration similar to that used in other multifaceted spoken language evaluation studies to facilitate a better comparison. Therefore, we used the Kaldi tools to extract 40-dimensional MFCCs features and used them as input for the standard acoustic model. Reference [26] is suitable for training a standard acoustic model of Mandarin due to its balanced phoneme distribution and comprehensive accent coverage. We used a 12-layer factorized time-delay neural network (TDNN-F) as the network structure for this acoustic model. Because the 863 Chinese speech dataset was not divided into training, validation, and test sets, we divided the dataset randomly in the ratio of 8:1:1. In the validation set, there are eight persons of each gender, covering a total of 9822 sentences. In the test set, there are nine persons of each gender, covering 10116 sentences, and we use the remaining speech data as the training set. Table 1 shows the specific division of the validation and test sets according to the speaker profile. The final word error rate (WER) for the standard acoustic model obtained through training was 8.68%.

TABLE 1. Division of validation and test sets in the 863 dataset.

	Male	Female
Dev	M74, M44, M32, M94	F77, F70, F98, F95
	M54, M93, M36, M66	F76, F25, F68, F93
Test	M21, M97, M03, M26	F22, F28, F41, F35, F23
	M75, M51, M59, M39, M12	F61, F39, F18, F84

After obtaining the acoustic model, we used the acoustic model for GOP feature extraction and GOP score calculation. We used the log posterior probability (LPP) and log posterior ratio (LPR) of the phonemes defined in [27] as the GOP features. Furthermore, we used the GOP algorithm mentioned in [10] to calculate the GOP scores. Equation (21) and (22) show the LPP calculation process for the phoneme p.

$$\mathrm{LPP}(p) \approx \frac{1}{t_e - t_s + 1} \sum_{t=t_s}^{t_e} \log P(p|o_t) \tag{21}$$

$$P(p|o_t) = \sum_{s \in p} P(s|o_t) \tag{22}$$

where t_s and t_e are the temporal indices that make up the start and end frames of the phoneme p, o_t is the sequence of observations at the moment t, and s is all of the states corresponding to the phoneme p. $\log P(p|o_t)$ is the log posterior probability of this phoneme p within o_t. The LPR of the constituent GOP feature sequences is defined as follows, where the LPR of the phoneme p_j for the other phonemes p_i is also defined:

$$\mathrm{LPR}(p_j|p_i) = \log P(p_j|o; t_s, t_e) - \log P(p_i|o; t_s, t_e) \tag{23}$$

In previous studies, researchers have not added information related to vocal tones to the process of calculating GOP features. We wanted to reflect the articulation of tones through the [LPP, LPR] feature. Therefore, we calculated the rhymes plus tones as a separate class of phonemes. In the end, the number of phonemes in the collated phoneme lexicon was 218, and we obtained a goodness of pronunciation feature (Equation (24)) for a phoneme p with a vector size of 436 dim according to the formulae for LPP and LPR.

$$F_{\mathrm{lpp+lpr}} = [\mathrm{LPP}(p_1), \ldots, \mathrm{LPP}(p_{218}), \mathrm{LPR}(p_1|p), \ldots, \mathrm{LPR}(p_{218}|p)] \tag{24}$$

IV. EXPERIMENTS

A. DATASET

1) SpeechOcean762 DATASET

SpeechOcean762 is a free, open-source dataset released by [28], which is designed for pronunciation quality detection and evaluation and consists of 5000 English sentences from 250 non-native English speakers, half of whom are children. SpeechOcean762 contains a rich set of tagging information. Specifically, it divides pronunciation quality into utterances, words, and phonemes and provides multiple pronunciation quality scores in each area. For utterances, it scores each speaker's corpus in five areas: accuracy, fluency, completeness, prosodic, and total. For words, it provides scores in three areas: accuracy, stress, and total (from 0 to 10). For phonemes, it gives scores for the accuracy of the phoneme (0-2). Five experts independently label these scores. The training set of this dataset consists of 2500 sentences, 15849 words, and 47076 phonemes, and the test set consists of 2500 utterances, 15967 words, and 47369 phonemes. In previous work, the part-of-speech and word-level scores were rescaled to the range of 0-2 to bring them to the same level as the phoneme scores. In this work, to verify the validity of the proposed model, we use the same experimental data as published in the paper by Yuan Gong et al.

2) PSC-MonoSyllable DATASET

The PSC-MonoSyllable dataset is a Chinese speech dataset we constructed for researching Mandarin mispronunciation detection, consisting of audio for 23,428 monosyllabic Chinese characters recorded by 185 university students from the Xinjiang region, with an effective duration of 4.14 hours. The dataset consists of two main parts (PSC-MonoSyllable-115 and PSC-MonoSyllable-60). The audio data in the PSC-MonoSyllable-60 section were recorded by 60 Xinjiang University students using computer microphones in a quiet environment. Three Mandarin assessment experts marked

TABLE 2. **The correlation between the manual scoring of PSC-MonoSyllable and the words.**

	PSC-MonoSyllable-115			PSC-MonoSyllable-60		
	Expert 1	Expert 2	Expert 3	Expert 1	Expert 2	Expert 3
Expert 1	1	0.711	0.697	1	0.667	0.58
Expert 2	-	1	0.685	-	1	0.582
Expert 3	-	-	1	-	-	1

TABLE 3. **Overview of the PSC-Monosyllable dataset.**

	Phoneme							
	Train				Test			
Correct	35211				7871			
Error	2743				1031			
Count	37954				8902			
	Word							
	Train				Test			
Correct	13907				2980			
Error	5070				1471			
Count	18977				4451			
	Tone							
	T1		T2		T3		T4	
	Train	Test	Train	Test	Train	Test	Train	Test
Correct	4010	854	3840	841	3098	764	4031	848
Error	757	265	817	248	1485	299	939	332
Count	4767	1119	4657	1089	4583	1063	4970	1180

the speakers' pronunciation according to their pronunciation status in graphemes, tones, and phonemes. Table 2 shows the correlation between the manual scoring of PSC-MonoSyllable and the words.

We divided the PSC-MonoSyllable dataset into a training set and a test set in the ratio of 8:2. The training set consisted of 18,977 words, 18,977 tones, and 37,954 phonemes, and the test set consisted of 4451 words, 4451 tones, and 8902 phonemes. Table 3 shows the distribution of pronunciation correctness and error in different aspects.

B. EXPERIMENT DETAILS

In the mispronunciation detection experiments, we used an acoustic model trained on the 863 dataset and extracted pronunciation goodness features based on this acoustic model. At the same time, we used the self-built PSC-MonoSyllable dataset for training and evaluation. In our construction of the Conformer-MB model, we set the embedding layer dimension to 512, used three layers of Conformer-MB encoders, and set the number of attentional heads per encoder to one. Meanwhile, in the Conformer-MB encoder, we set the convolutional kernel size of the convolutional module to 31 × 31 and the dilation rates to 1, 5, and 9. For the actual training process, we used the Adam optimizer with an initial optimization rate of 1e-4 and a batch size of 128, using the mean square error (MSE) as the loss function. In terms of threshold setting, we first obtained the model with the smallest MSE by model training and then divided it into multiple thresholds in the range of 0.1 to 0.95, according to a spacing of 0.05, and used these thresholds to test the phonemes, tones and words tasks separately under different models. If the model output exceeded this threshold, it was classified as correct pronunciation (1). Otherwise, it was classified as incorrect pronunciation (0). This approach is similar to the threshold selection proposed by [10]. We used precision, recall, and F1-score as metrics to measure the model results, calculate the metrics, and compare them according to the two categories of correct and incorrect pronunciation. As shown in Figure 5, which illustrates the F1-score of the mispronunciation detection metric for each task at different thresholds for the Confomer-MB model obtained through training, it can be concluded from the figure that the model performs optimally when the thresholds for the phonemes, tones and words tasks reach 0.75, 0.5 and 0.5, respectively.

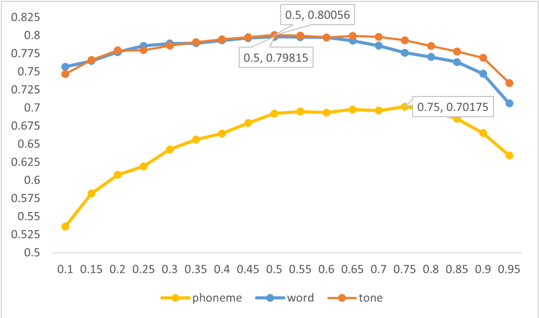


FIGURE 5. **The threshold selection process of the Conformer-MB model in the mispronunciation detection task.**

In the pronunciation quality assessment experiments, we used the same experimental data, training approach, and evaluation metrics (PCC) as in [21]. We adopted the same dimensionality of the embedding layer of the constructed pronunciation quality assessment model as that used in the original text, using the same encoder structure and parameters as in the mispronunciation detection task. In both experiments, we trained 100 epochs for each job and cut the learning rate in half every five epochs after the 20th epoch. We used a 3090 Nvidia GPU graphics card for training, CUDA version 11.6, Pytorch version 1.10.0, and AMD Ryzen 9 5900X CPU model.

1) MISPRONUNCIATION DETECTION

Research on multi-level pronunciation error detection for Mandarin is scarce. This paper's mispronunciation detection experiment is the first study on multi-level Mandarin mispronunciation detection with a multi-task learning approach. Therefore, we constructed the LSTM, Transformer model as a baseline for the pronunciation error detection task, following the model comparison approach in [21]. We also compared some traditional mispronunciation detection methods. We compared the following six models:

1) Classification using GOP scores using pre-set fixed thresholds. The threshold was set to -0.15 for phonemes, -0.10 for words, and -0.15 for tones;

TABLE 4. Results of mispronunciation detection experiments.

	Correct Pronunciation Detection			**Mispronunciation Detection**		
	Initial and Final (Phoneme)					
	PR	**RE**	**F1**	**PR**	**RE**	**F1**
GOP [10]	97.99%	39.65%	56.45%	16.91%	93.79%	28.66%
SVM [3]	94.12%	69.14%	79.72%	22.15%	67.02%	33.29%
LSTM	94.16%	98.25%	96.16%	79.97%	53.44%	64.07%
Transformer	95.78%	96.49%	96.13%	71.61%	67.51%	69.50%
Transformer-I	95.38%	97.65%	96.50%	78.01%	63.92%	70.29%
Conformer	95.66%	96.89%	96.27%	73.66%	66.44%	69.86%
Conformer-I	96.16%	95.64%	95.90%	68.03%	70.81%	69.39%
Conformer-II	95.72%	97.17%	96.44%	75.44%	66.82%	70.92%
Conformer-III	95.44%	97.76%	96.59%	79.02%	64.31%	70.91%
Conformer-MB	95.83%	96.65%	96.25%	72.61%	67.90%	70.18%
	Tone					
	PR	**RE**	**F1**	**PR**	**RE**	**F1**
GOP [10]	95.95%	19.14%	31.91%	28.67%	97.57%	44.31%
SVM [3]	90.82%	61.04%	73.01%	41.05%	81.47%	54.60%
LSTM	89.93%	93.89%	91.87%	78.86%	68.44%	73.28%
Transformer	90.48%	93.98%	92.20%	78.38%	61.20%	74.65%
Transformer-I	92.48%	89.88%	91.16%	71.97%	78.06%	74.89%
Conformer	91.24%	94.19%	92.69%	79.23%	74.46%	76.77%
Conformer-I	92.60%	92.87%	92.73%	78.40%	77.70%	78.05%
Conformer-II	93.32%	92.93%	93.13%	79.04%	80.04%	79.54%
Conformer-III	94.32%	91.08%	92.67%	75.71%	83.54%	79.44%
Conformer-MB	92.45%	**95.03%**	**93.72%**	**83.71%**	76.71%	**80.06%**
	Word					
	PR	**RE**	**F1**	**PR**	**RE**	**F1**
GOP [10]	92.46%	19.33%	31.97%	37.20%	96.81%	53.75%
SVM [3]	85.35%	67.05%	75.01%	53.46%	76.68%	63.00%
LSTM	86.95%	90.77%	88.82%	79.48%	72.40%	75.77%
Transformer	88.71%	88.89%	88.80%	77.41%	77.09%	77.25%
Transformer-I	87.79%	90.74%	89.24%	79.87%	74.44%	77.06%
Conformer	88.54%	91.28%	89.89%	81.15%	76.07%	78.53%
Conformer-I	88.48%	92.25%	90.32%	82.81%	75.66%	79.08%
Conformer-II	88.70%	92.69%	90.65%	83.70%	76.07%	79.70%
Conformer-III	88.98%	91.58%	90.26%	81.86%	77.02%	79.37%
Conformer-MB	88.76%	**92.72%**	**90.69%**	**83.78%**	76.21%	**79.82%**

2) A support vector machine classifier (SVM)-based model using the SVM classifier from the scikit-learn library with the parameters set to default;
3) An LSTM-based model;
4) A Transformer model used for training;
5) A Transformer model used for training. The output from each layer of the Transformer encoder is collected separately and stitched together to train the pronunciation error detection model (Transformer-I);
6) Training using the standard Conformer model;
7) Training using the standard Conformer model. The output from each layer of the Conformer encoder is collected separately and stitched together to train the pronunciation error detection model (Conformer-I);
8) Training using the Conformer-MB model. Convolution kernel size set to 1, 3, 5 and dilation rate set to 1 (Conformer-II);
9) Training using the Conformer-MB model. Convolution kernel size set to 1, 5, 9 and dilation rate set to 1 (Conformer-III);
10) Training using the Conformer-MB model.

Except for model 1, all models used pronunciation goodness features as input to the models. To make a fair comparison between different aspects of the detection task, models 3, 4, 5, 6, 7, 8, 9 and 10 above had the same depth, and embedding dimensions were trained with the same settings during evaluation. Also, we have tried other multi-scale feature methods for better comparison. For example, in the model 5 and 7, we extracted features of different scales from each encoder layer and stitched them together for pronunciation error detection. In model 8 and 9, we use different-sized convolution kernels to collect information at different scales and set the dilation rate to 1. Finally, we selected the best results for each model based on the test results. All of the above models are based on acoustic models trained with the same 863 data, using the same GOP features. Therefore, we conducted fair comparison experiments to demonstrate that the performance differences were not due to the GOP features.

The experimental results (Table 4) show that in the multifaceted mispronunciation detection task, models 3, 4, 5, 6, 7, 8, 9 and 10 with the multi-task learning scheme showed

TABLE 5. Results of the ablation experiment for the mispronunciation detection task.

	Correct Pronunciation Detection (F1)			Mispronunciation Detection (F1)		
	Phoneme	Tone	Word	Phoneme	Tone	Word
Training Mask						
Only Phoneme	96.54%	-	-	69.70%	-	-
Only Tone	-	93.45%	-	-	79.51%	-
Only Word	-	-	90.23%	-	-	79.20%
Joint	96.25%	**93.72%**	**90.69%**	**70.18%**	**80.06%**	**79.82%**
Attention Heads						
1	96.25%	**93.72%**	90.69%	70.18%	79.07%	**80.20%**
4	96.34%	92.52%	90.03%	71.59%	76.94%	78.65%
8	96.38%	93.04%	91.04%	70.84%	79.61%	80.12%
Embedding Dimension						
64 (337K Params)	95.91%	86.73%	87.09%	64.07%	65.70%	72.40%
128 (1216K Params)	96.06%	92.08%	89.56%	67.05%	73.30%	77.83%
256 (4595K Params)	96.17%	92.41%	89.17%	69.18%	75.90%	77.06%
512 (17840K Params)	96.25%	93.72%	**90.69%**	70.18%	**80.06%**	79.82%
1024 (70284K Params)	96.40%	93.42%	90.67%	71.04%	79.50%	80.56%
2048 (278981K Params)	96.39%	93.31%	89.66%	70.08%	78.36%	79.38%
Layers						
1 (6174K Params)	96.16%	92.95%	88.81%	69.20%	78.89%	78.53%
3 (17840K Params)	96.25%	**93.72%**	**90.69%**	70.18%	**80.06%**	79.82%
6 (35340K Params)	96.45%	92.67%	90.39%	70.43%	78.29%	79.88%
9 (52840K Params)	96.52%	92.93%	90.08%	69.35%	78.44%	78.55%

TABLE 6. Results of the pronunciation quality assessment experiment.

	Phoneme Score		Word Score (PCC)			Utterance Score (PCC)				
	MSE	PCC	Accuracy	Stress	Total	Accuracy	Completeness	Fluency	Prosodic	Total
Transformer [21]	0.084	0.613	0.537	0.320	0.551	0.718	0.194	0.758	0.766	0.747
HiPAMA [29]	0.084	0.616	0.575	0.320	0.591	0.730	0.276	0.749	0.751	0.754
Embedding(12)										
Conformer	0.084	0.617	0.565	0.299	0.580	0.721	0.110	0.750	0.750	0.747
Conformer-MB	0.083	**0.625**	**0.574**	**0.317**	**0.593**	**0.725**	0.079	**0.761**	**0.757**	**0.748**
Embedding(24)										
Conformer	0.083	0.622	0.563	0.296	0.579	0.718	0.181	0.749	0.744	0.743
Conformer-MB	0.084	0.622	**0.567**	0.279	**0.583**	**0.723**	0.14	**0.764**	**0.762**	**0.746**
Embedding(48)										
Conformer	0.084	0.615	0.565	0.300	0.580	0.724	0.112	0.754	0.752	0.747
Conformer-MB	0.084	**0.621**	0.565	0.285	0.579	0.721	0.128	**0.756**	**0.754**	0.746
Embedding(96)										
Conformer	0.084	0.62	0.569	0.290	0.584	0.727	0.030	0.757	0.757	0.75
Conformer-MB	0.084	**0.624**	**0.572**	0.276	**0.588**	0.727	0.163	**0.766**	**0.764**	**0.752**

significant improvements in mispronunciation detection compared with a single modeling approach using GOP scores and support vector machines for a particular aspect. In particular, using the Conformer-MB model was particularly effective in detecting pronunciation errors in words and tones, demonstrating that better detection can be achieved using a multi-task learning solution.

Secondly, in the experimental comparison using the multi-task learning scheme, there was a significant improvement in the detection of words and tones using the standard Conformer model structure, with a gain of 1.28% and 2.12% in the F1 metric for the mispronunciation detection category, respectively, compared with the Transformer model. This indicates that using a convolutional network that pays more attention to local information can effectively enhance the mispronunciation detection of the model.

In addition, the results of comparing Model 4 with Model 5 and Model 6 with Model 7 demonstrate that the use of multi-scale feature fusion can effectively improve the pronunciation error detection performance of the models. Also, model 8 and model 9 showed significant improvement in mispronunciation detection performance for phonemes, tones and words tasks compared to models 6 and 7. This result demonstrates that extracting features at different scales using multiple convolutional networks is more effective for mispronunciation detection tasks. Our proposed Conformer-MB model uses the dilation convolution method to obtain features at multiple scales, the final results compared to models 8 and 9, which further improve the error detection performance for the phoneme and grapheme tasks. Thus, the structure of the Conformer-MB model is more effective in extracting feature information at different scales, thus improving the overall (tones and words) performance. The model's mispronunciation detection is improved.

Finally, our proposed Conformer-MB model achieves optimal results in error detection for phonemes, tones, and words compared with all of the models mentioned above. The experimental results demonstrate the effectiveness of our proposed multi-width band approach, which allows the model to learn local feature information at different scales, significantly increasing the model's overall mispronunciation detection effectiveness.

We conducted a series of ablation studies to show the effect of various factors on performance. We set the model mentioned in Section III with three Conformer-MB encoders and an embedding layer dimension of 512 as the base model and trained it with all phoneme, tone, and grapheme evaluation tasks, then changed one factor at a time to observe performance changes. Table 5 shows the results of the ablation experiments of the Conformer-MB model in the mispronunciation detection task. Firstly, the Conformer-MB model trained using multi-task learning showed a significant improvement in mispronunciation detection in phonemes and words compared with the training method using a single task. The results of this experiment demonstrate that the multi-task learning approach allows the model to perform multiple aspects of mispronunciation detection simultaneously and improves the performance of individual tasks. Second, we compare the effect of different Conformer-MB encoder sizes on mispronunciation detection performance. We find that error detection performance is not further improved by increasing the encoder depth, suggesting that small models are preferred with relatively small datasets. In addition, we compared encoders using different numbers of attention heads, and the experimental results indicate that better results were obtained when using an attention head of 1. This experimental result is consistent with the conclusion reached in [21], demonstrating that increasing the number of attention heads in the encoder does not improve the effectiveness of model error detection. Meanwhile, although the GOP features are 436-dimensional, we experimentally demonstrate that, as the embedding layer becomes progressively more significant in terms of dimension, the encoder can capture enough rich information from it to enhance the modeling effect further. In particular, error detection is optimal for tones when the embedding size is 512 dimensions and for phonemes and words when the embedding size is 1024. However, since increasing the embedding dimension leads to an increase in the number of parameters, we choose an embedding size of 512 dimensions as the optimal result of our proposed method.

Finally, We have conducted relevant experiments in terms of inference time. We experimented with the inference times of the Transformer, Conformer, and Conformer-MB models in the mispronunciation detection task based on the test set. The results showed that the Transformer model had the shortest inference time of 1.27±0.02 seconds, and the Conformer model had the longest inference time of 1.35±0.01 seconds. The inference time of the Conformer-MB model is 1.32±0.01 seconds, which is faster than the Conformer model, but slightly slower than the Transformer model.

2) PRONUNCIATION QUALITY ASSESSMENT

In the pronunciation quality assessment task, we mainly compared the Transformer model proposed by [21]. In our experiments, we set the encoder depth and the number of attentional heads to the same as those of the baseline for the Conformer model and the Conformer-MB model. We set the embedding layer vector dimensions to 12, 24, 48, and 96 to compare the experimental results with different embedding layer vector dimensions, respectively. For a relatively fair comparison, we used the same experimental data and features as in the original paper, while keeping the training method the same.

Table 6 shows the experimental results of the pronunciation quality assessment task. First, we compared the overall scoring effect of the Conformer model with the Conformer-MB model for different embedding layer dimensions. Our proposed Conformer-MB model has a slightly higher scoring correlation than the Conformer model in all sizes and achieves optimal scores at the phoneme and word levels at an embedding dimension of 12 and an embedding dimension of 96. Secondly, we used the Conformer model structure compared with the Transformer model, which showed a slight increase in pronunciation accuracy scores at the phoneme level and pronunciation accuracy and overall scores at the word level, but a decrease in the correlation of scores in some aspects (accuracy, completeness, and rhyme) at the part-of-speech level. The reason for this is that the Transformer model, which focuses more on the global representation of the feature sequence, scores better at the part-of-speech level, while the Conformer model, which takes into account the global representation while paying more attention to local feature details, scores better at the phoneme and word levels. Our proposed Conformer-MB model solves this problem to some extent. The experimental results show that when the embedding dimension reaches 96, there is a specific improvement in pronunciation accuracy, completeness, fluency, rhythm, and overall score at the discourse level compared with the Conformer model. At the same time, the correlation between pronunciation accuracy and overall score at the word level achieved the best results, reaching 0.572 and 0.588, respectively. Thus, the effectiveness of our proposed method using multi-scale local feature fusion in pronunciation quality assessment tasks is demonstrated experimentally.

In addition to this, we also compared other multifaceted pronunciation quality assessment studies. The HiPAMA model is a multifaceted pronunciation quality assessment method proposed by [29], which uses the same experimental configuration as [21] to improve the overall assessment performance by stratifying different aspects of the assessment task through an attention mechanism. According to the comparative experimental results, the HiPAMA model improves the total scores in words and the accuracy and completeness scores in utterances, mainly due to the use of convolutional layers to capture the local information among phonemes. However, the improvement in phoneme accuracy and sentence fluency, and prosodic scores are minor because the model structure must fully utilise the global feature information sequence. Although our proposed model is slightly less effective than the HiPAMA in words score, it is higher than HiPAMA regarding phoneme accuracy, fluency, and prosodic scores under different embedding layer dimension sizes. It demonstrates that our proposed model can improve the overall performance of the model not only by using local feature information among phonemes but also by capturing global information of the whole part of speech.

V. CONCLUSION

In this paper, we proposed an improved Conformer model based on inflated convolution and applied it to the characteristics of speaking assessment tasks. The model is made more representable by obtaining different local feature information at different scales, thus improving the performance of various aspects of speaking assessment. Using different datasets for the mispronunciation detection task and the pronunciation quality assessment task, we demonstrate that our proposed approach can simultaneously evaluate multiple aspects of the speaking assessment task and improve the model performance in different aspects of such a task. The F1 indicators in the mispronunciation detection task were enhanced by 0.68% (phonemes), 5.41% (tones), and 2.57% (words) relative to the baseline model (vanilla Transformer), and by 0.32% (phonemes), 3.29% (tones), and 1.29% (words) relative to the Conformer model, respectively. At the same time, the effectiveness of our proposed model in detecting errors in words and tones is further improved with different multi-scale feature fusion methods. In the pronunciation quality assessment task, the score correlations of our proposed method were significantly higher for phonemes and graphemes compared with the baseline model (Transformer). Thus, the above experiments demonstrate the validity of our proposed method.

In addition, we propose an approach based on the traditional approach to spoken language assessment using an acoustic and a spoken language assessment model for the spoken language assessment task. First, we use an acoustic model to obtain confidence features and build a speaking evaluation model based on this feature to perform the evaluation. In this process, the final assessment results are susceptible to the performance of the acoustic model and accumulate errors in the subsequent steps, continuously amplifying them. The use of an end-to-end approach for speaking assessment is a popular method to improve this problem significantly, but most of the existing studies using end-to-end approaches for speaking assessment only model a single aspect, e.g., phoneme, tone, Etc. At the same time, some studies demonstrate the effectiveness of using wav2vec2.0 and Hubert features to improve the effectiveness of speaking assessment. Later, we will build on the model structure proposed in this paper to improve the effectiveness of spoken language assessment further using an end-to-end approach and speech features such as wav2vec2.0.

REFERENCES

[1] M. Eskenazi, "An overview of spoken language technology for education," *Speech Commun.*, vol. 51, no. 10, pp. 832–844, Oct. 2009.

[2] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Commun.*, vol. 51, no. 10, pp. 883–895, Oct. 2009.

[3] W. Li, S. M. Siniscalchi, N. F. Chen, and C.-H. Lee, "Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 6135–6139.

[4] C.-F. Shen, "Tone evaluation algorithm of Mandarin continuous speech," M.S. thesis, Soochow Univ., Suzhou, China, 2012.

[5] R. Zhang, "Research on automatic evaluation methods of Mandarin pronunciation," Ph.D. dissertation, Harbin Inst. Technol., Harbin, China, 2013.

[6] L. Zhang, "Research on automatic evaluation methods of Mandarin pronunciation quality," Ph.D. dissertation, Harbin Inst. Technol., Harbin, China 2014.

[7] L. Zhang, H.-F. Li, L. Ma, and J.-H. Wang, "Automatic detection and evaluation of the Erhua in Putonghua proficiency test," *Chin. J. Acoust.*, vol. 39, no. 5, pp. 639–646, 2014.

[8] B. Lin, L. Wang, X. Feng, and J. Zhang, "Automatic scoring at multi-granularity for L2 pronunciation," in *Proc. Interspeech*, Oct. 2020, pp. 3022–3026.

[9] T. Cincarek, R. Gruhn, C. Hacker, E. Nöth, and S. Nakamura, "Automatic pronunciation scoring of words and sentences independent from the non-native's first language," *Comput. Speech Lang.*, vol. 23, no. 1, pp. 65–88, Jan. 2009.

[10] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Commun.*, vol. 30, nos. 2–3, pp. 95–108, 2000.

[11] S. Sudhakara, M. K. Ramanathi, C. Yarra, and P. K. Ghosh, "An improved goodness of pronunciation (GoP) measure for pronunciation evaluation with DNN-HMM system considering HMM transition probabilities," in *Proc. Interspeech*, Sep. 2019, pp. 954–958.

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

[13] X. Xu, Y. Kang, S. Cao, B. Lin, and L. Ma, "Explore wav2vec 2.0 for mispronunciation detection," in *Proc. Interspeech*, Aug. 2021, pp. 4428–4432.

[14] E. Kim, J.-J. Jeon, H. Seo, and H. Kim, "Automatic pronunciation assessment using self-supervised speech representation learning," 2022, *arXiv:2204.03863*.

[15] S. Bannò and M. Matassoni, "Proficiency assessment of L2 spoken English using wav2vec 2.0," 2022, *arXiv:2210.13168*.

[16] B. Lin and L. Wang, "Deep feature transfer learning for automatic pronunciation assessment," in *Interspeech*, vol. 2021, pp. 4438–4442.

[17] W.-K. Leung, X. Liu, and H. Meng, "CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 8132–8136.

[18] K. Fu, J. Lin, D. Ke, Y. Xie, J. Zhang, and B. Lin, "A full text-dependent end to end mispronunciation detection and diagnosis with easy data augmentation techniques," 2021, *arXiv:2104.08428*.

[19] Y. Feng, G. Fu, Q. Chen, and K. Chen, "SED-MDD: Towards sentence dependent end-to-end mispronunciation detection and diagnosis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3492–3496.

[20] T.-H. Lo, S.-Y. Weng, H.-J. Chang, and B. Chen, "An effective end-to-end modeling approach for mispronunciation detection," 2020, *arXiv:2005.08440*.

[21] 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 *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7262–7266.

[22] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," 2020, *arXiv:2005.08100*.

[23] J. Burkardt, "The truncated normal distribution," Dept. Sci. Comput., Florida State Univ., Tallahassee, FL, USA, Tech. Rep., 2014, vol. 1, p. 35.

[24] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," 2017, *arXiv:1710.05941*.

[25] M. Tu, A. Grabek, J. Liss, and V. Berisha, "Investigating the role of L1 in automatic pronunciation evaluation of L2 speech," 2018, *arXiv:1807.01738*.

[26] A. Li, Z. Yin, T. Wang, Q. Fang, and F. Hu, "RASC863—A Chinese speech corpus with four regional accents," in *Proc. ICSLT-o-COCOSDA*, New Delhi, India, 2004.

[27] 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 Commun.*, vol. 67, pp. 154–166, Mar. 2015.

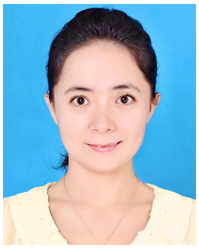
[28] J. Zhang, Z. Zhang, Y. Wang, Z. Yan, Q. Song, Y. Huang, K. Li, D. Povey, and Y. Wang, "speechocean762: An open-source non-native English speech corpus for pronunciation assessment," 2021, *arXiv:2104.01378*.

[29] H. Do, Y. Kim, and G. G. Lee, "Hierarchical pronunciation assessment with multi-aspect attention," 2022, *arXiv:2211.08102*.

ZHIXING FAN received the B.Eng. degree in electrical engineering and its automation from Shanghai Normal University. He is currently pursuing the master's degree in software engineering with Xinjiang University. His research interest includes mispronunciation detection and diagnosis (MDD).

JING LI received the B.Eng. degree in IoT engineering from Tarim University. She is currently pursuing the master's degree in computer science and technology with Xinjiang University. Her research interest includes mispronunciation detection and diagnosis (MDD).

AISHAN WUMAIER (Member, IEEE) received the Ph.D. degree in computer applied technology from Xinjiang University. He is currently a Professor with Xinjiang University, where he is also a Ph.D. Supervisor. His research interests include sentiment analysis, machine translation (MT), and mispronunciation detection and diagnosis (MDD).



ZAOKERE KADEER is currently an Experimentalist with Xinjiang University. Her research interest includes natural language processing.

ABDUJELIL ABDURAHMAN received the Ph.D. degree in operational research and cybernetics from Xinjiang University. He is currently an Associate Professor with Xinjiang University, where he is also a Ph.D. Supervisor. His research interests include the theory of differential equations and their applications, the theory of discontinuous systems, and the study of their applications.