MultiPA: A Multi-task Speech Pronunciation Assessment Model for Open Response Scenarios

Yu-Wen Chen, Zhou Yu, Julia Hirschberg

Department of Computer Science, Columbia University, United States

{yuwchen, zhouyu, julia}@cs.columbia.edu

Abstract

Pronunciation assessment models designed for open response scenarios enable users to practice language skills in a manner similar to real-life communication. However, previous open-response pronunciation assessment models have predominantly focused on a single pronunciation task, such as sentence-level accuracy, rather than offering a comprehensive assessment in various aspects. We propose MultiPA, a Multitask Pronunciation Assessment model that provides sentence-level accuracy, fluency, prosody, and word-level accuracy assessment for open responses. We examined the correlation between different pronunciation tasks and showed the benefits of multi-task learning. Our model reached the state-of-the-art performance on existing in-domain data sets and effectively generalized to an out-of-domain dataset that we newly collected. The experimental results demonstrate the practical utility of our model in real-world applications.

Index Terms: pronunciation assessment, open-response scenarios, multi-task learning, real-world evaluation

1. Introduction

An automatic speech pronunciation assessment model offers a less expensive and more efficient approach to practicing and evaluating pronunciation skills in a second language (L2) [1, 2]. One common design of the pronunciation assessment model is the closed-response scenario. In the closed-response scenario, L2 learners are instructed to speak a predetermined target sentence, which is then used as the ground-truth transcript for the model. However, this design restricts evaluation to predefined sentences and fails to reflect learners' pronunciation skills in real-world communication. In contrast, the open-response scenario allows learners to speak freely or to respond to a given task or question, providing a more authentic evaluation of learners' pronunciation skills. Therefore, in this study, we aim to develop a pronunciation model that can assess learners' pronunciation skills in this open response scenario.

One of the most commonly used methods for pronunciation assessment is using Goodness of Pronunciation (GoP) features [3, 4, 5, 6, 7, 8]. These features are calculated from the posterior probability of the ground-truth phone using an automatic speech recognition (ASR) model [9]. Therefore, GoP-based models are designed and evaluated under the assumption of having accurate transcripts. As a result, their performance may experience a notable decline when directly applied to open-response scenarios where alignment between audio and accurate transcripts is not available [10]. Without access to a target sentence or ground-truth transcript, testing in the open response scenario requires a non-intrusive assessment model. Such a model can either utilize the ASR recognition results

as its transcript or operate without using a transcript for evaluation [11, 12, 13]. However, studies on non-intrusive pronunciation assessment have predominantly centered on a single pronunciation task [14, 15, 16]. For example, [17] employed deep features from a DNN-HMM-based acoustic model to obtain a sentence-level total score. [18] proposed a self-supervised learning (SSL)-based zero-shot model to assess sentence-level total proficiency. [19] fused acoustic and phoneme representations to obtain sentence-level accuracy scores. [20] developed an SSL-based ASR-free approach for fluency assessment. None of these previous studies have explored non-intrusive multitask pronunciation assessment for both sentence-level and word-level scores. In a multi-task assessment model, sentence-level accuracy, fluency, and prosody assessments provide L2 learners with a comprehensive overview of their pronunciation skills, while word-level scores pinpoint specific parts of their speech that require practice.

In this study, we propose MultiPA, a Multi-task Pronunciation Assessment model for open-response scenarios. Compared with the previous non-intrusive pronunciation assessment models, MultiPA provides a more comprehensive pronunciation assessment, including both sentence- and word-level assessments. Furthermore, we are the first to conduct a pilot study to evaluate the model's performance in real-world open-response scenarios. To do this, we collected data from L2 learners who were using an English learning chatbot to practice English and recruited experts for the annotation. The experimental results show the effectiveness of using our model in the real-world use case, and the pilot data we collected will be released for other studies to evaluate their model on multitask pronunciation assessment.

2. Our Multi-task Pronunciation Assessment model

MultiPA utilizes a pretrained SSL model (i.e., HuBERT [21]) as its main structure. Fine-tuning pre-trained SSL models has proven effective in phone-level mispronunciation detection [22], phone-level assessment [23], and sentence-level assessment [24, 25], but has not been used for assessing multi-task sentence-level and word-level scores. To provide word-level evaluation, it is necessary to identify the boundaries of individual words within the speech signal. MultiPA achieves this by first using the ASR model Whisper [26] to identify potential words in the speech signal, followed by using Charsiu [27] to obtain alignment information between the words and the speech signals. Given the potential inaccuracies of the ASR, RoBERTa [28] was employed to offer supplementary semantic feedback for the recognized words. Figure 1 shows an overview of MultiPA.
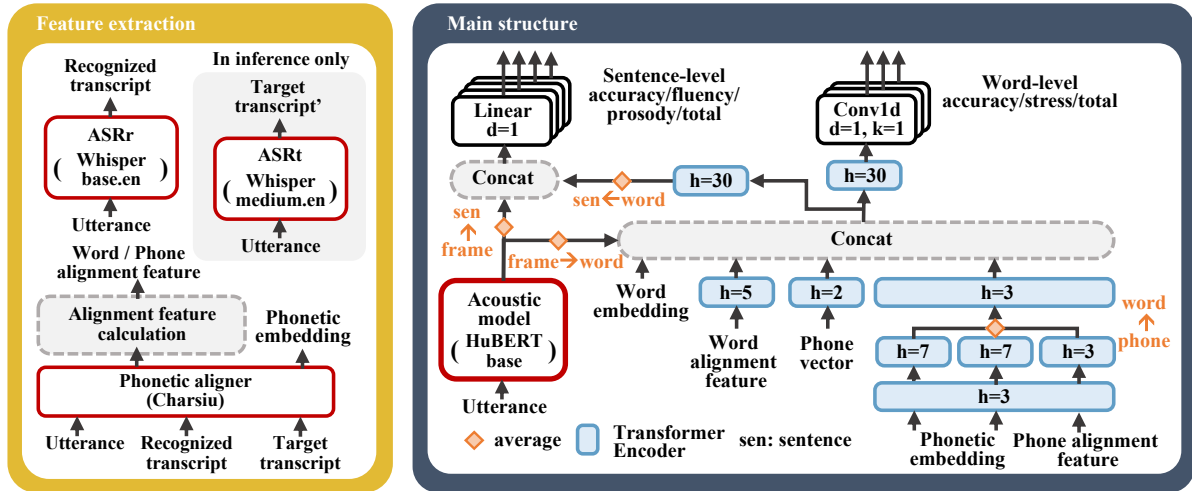
Figure 1: *Overview of MultiPA, where d in Linear and Conv1d layers refers to the output dimension, k is the kernel size, and h indicates the number of heads. The selection of h is based on empirical results.*

2.1. Auditory feature extraction

MultiPA extracts features from two transcripts: the target transcript and the recognized transcript. The target transcript represents the sentence that the learner wants to say, while the recognized transcript is how an ASR model (ASRr) recognized the speech signal. Although the target transcript is unavailable during inference, we still use it during training because word-level scores of training data are aligned with the target transcript. During inference, we use the recognition results of another ASR model (ASRt) as an alternative to the target transcript. Specifically, Whisper base.en (ASRr) generates the recognized transcript, while a larger ASR model, Whisper medium.en (ASRt), replaces the target transcript. Then, the phonetic aligner Charsiu provides word and phone alignment information between the transcript and the utterance. These aligned transcripts enable the extraction of word-level and phone-level features.

2.1.1. Word-level features

Word-level features are aligned on a word-by-word basis, with the length equal to the number of words in the target transcript. These features include word-embeddings, phone vectors, and word-alignment features. The word-embedding is the concatenation of RoBERTa [28] embeddings from target and recognized transcripts. The phone vector is the one-hot-encoded phones of the word. The word-alignment feature is computed using the alignment information with the transcripts, including duration (measuring the time each word takes), interval (indicating the time gap from the preceding word), time difference (capturing the variance in start and end times between words in the two transcripts), distance (reflecting the Levenshtein distance between words in the transcripts), aligned word count (counting the number of words aligned with the target word), phone distance (quantifying the matched phones between target word and aligned recognized words), and phone ratio (expressing the ratio of phones in the target word to those in the recognized word).

2.1.2. Phone-level features

Phone-level features are aligned on a phone-by-phone basis with the length equal to the number of phones in the target transcript. The phone-level features contain phonetic embedding and phone alignment features. The phonetic embedding is the output layer of the Charsiu, whose value indicates the probability of all possible phones. Phone alignment features include duration, interval, time difference, aligned phone count, and phone probability, where duration, interval, time difference, and aligned phone count have a similar definition as the word alignment features but are calculated on a phone basis. Lastly, the phone probability is the probability of aligning the specific targeted phone or recognized phone to the signals.

2.2. Main structure

The main structure of MultiPA is based on fine-tuning a pre-trained self-supervised-learning (SSL)-based model, HuBERT, with additional layers. MultiPA employs transformerEncoder layers for feature fusion and uses average pooling and alignment information to align features at different levels (e.g., aligning phone-level features to word-level features). Finally, linear layers and convolutional layers are used for sentence- and word-level assessment, respectively.

3. Experimental setup

3.1. Data

We utilize two datasets in our study: the open-source pronunciation assessment dataset speechocean762 [29] and our self-collected data (referred to as multiPA data). The speechocean762 dataset serves as both the training and in-domain testing, while the multiPA dataset is used for out-of-domain testing. The speechocean762 dataset was collected in a closed-response scenario with known ground-truth transcripts. However, to simulate an open-response scenario, we did not use the ground-truth transcript as model input during testing. The multiPA data was collected from real-world open-response scenarios. Detailed descriptions of the datasets are provided below.

3.1.1. speechocean762

The speechocean762 dataset contains 5,000 English utterances from 250 non-native speakers, each utterance labeled at the sentence, word, and phone level. We followed the training and testing split provided by the dataset and focused solely on the sentence and word-level labels. The sentence-level labels include the accuracy, fluency, prosody, and total scores, whereas word-level labels consist of accuracy, stress, and total scores. The utterances in the dataset are from 2 to 20 seconds long.

3.1.2. multiPA data

The multiPA dataset comprises 50 audio clips, each lasting 10 to 20 seconds, obtained from about 20 anonymous dialog chatbot users. These users were native Mandarin speakers who used the dialog chatbot to practice their English. As a real use case for automatic pronunciation assessment, users can access the system with their own headsets at their own location. We recruited five annotators with high proficiency in English to annotate the audio clips at both sentence and word levels. The sentence labels include accuracy, fluency, and prosody scores, graded on a scale from 1 (very poor) to 5 (excellent). At the word level, the focus was on intelligibility; annotators used four levels to mark segments of speech: (1) cannot understand, (2) challenging to understand but recognizable, (3) somewhat inaccurate but quite understandable, and (4) good. The target scores used for analysis were the average scores from these five annotators. We have obtained IRB approval to collect data, and we have undergone an ethics assessment.

3.2. Model details and evaluation

For all experiments, we train the model with a batch size of 2 and an SGD optimizer with learning rate 5e-5 and momentum 7e-1. All models are trained using early stopping, with a patience of 2. 10% of the training data is used as the validation set for model selection and early stopping detection. We use the Pearson correlation coefficient (PCC) as the main evaluation metric because it has often been used in previous studies and provides better interpretability when comparing the performance on in-domain and out-of-domain data. If models cannot generate assessment scores (e.g., the forced aligner fails to align the text to speech), the lowest scores in the speechocean762 training data are used as alternatives. Lastly, we repeat each experiment five times with different random seeds and report the mean and standard deviation of the results.

4. Results

4.1. Model performance

Table 1 presents a comparison of model performance without using the ground-truth transcript. The term *GT transcript free* indicates whether the original design evaluated the model under the assumption that ground-truth transcripts were unavailable. First, we conducted a comparison between our model and the GOP-based model in an open-response scenario. We selected GOPT [3] as a representative because it is the basis for several studies [30, 31, 32], and its code is open-source. To use GOPT in the open-response scenario, we used the transcript from Whisper medium.en to replace the ground-truth transcript, following the MultiPA setting. The experimental results demonstrate that MultiPA outperforms GOPT significantly. Despite both models being trained with ground-truth transcripts, MultiPA's structure is more robust for handling inaccurate tran-

scripts and can be directly applied in open-response scenarios. Note that, when evaluating in the open response scenario, there is a potential mismatch between ground-truth word labels and the assessed scores because the assessment is based on ASR-recognized words. Therefore, we used Charsiu's alignment information to force-align ground-truth words to each recognized word, and employed the average score of aligned ground-truth words as the target score for the corresponding recognized word. In this way, although the ASR-recognized words might be incorrect, both the target and predicted scores refer to similar portions of the speech signal.

We compared MultiPA with models that were evaluated without using the ground-truth transcripts. Results show that our model achieved comparable or higher performance for the single task while providing more comprehensive assessment scores for different aspects. We also include vanilla SSL as one of the baselines, which fine-tuned a pre-trained SSL model by average-pooling the SSL's output embeddings and adding a dense output layer for sentence-level scores. While vanilla SSL has the ability to provide sentence-level assessments, it lacks the capability to provide a word-level assessment due to the absence of information on word boundaries. With the additional features introduced in MultiPA, MultiPA is able to provide word-level assessment and more accurate sentence-level assessment.

4.2. Ablation study of ASR models

We conducted ablation studies to show the performance impact of using different ASR models which differ mainly in model size. In Figure 2, we observed that, while ASRr is fixed as Whisper base.en, employing medium.en produces the highest scores, whereas utilizing base.en results in the lowest. This suggests that employing ASR models with greater diversity can enhance overall performance. In addition, the improved performance of models with ASRr compared to one without reflects the effectiveness of integrating ASRs and alignment features.
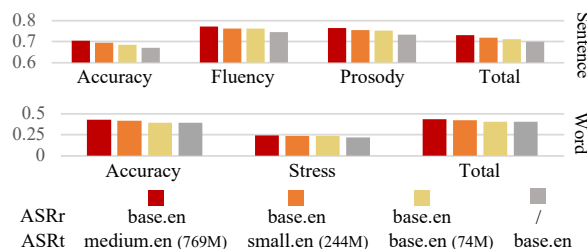

Figure 2: *Ablation studies for using different ASR models.*

4.3. Analysis of real-world open response scenario data

We first calculated correlations between different pronunciation tasks (Figure 3) in both speechocean762 and multiPA data. We observed that, in the speechocean762 dataset, the correlation between fluency and prosody is higher compared to other task pairs. This strong correlation could be attributed to the similarity in the instructions for assessing fluency and prosody. For example, the criterion for the highest fluency score is *"coherent speech, without noticeable pauses, repetition or stammering,"* and that for prosody is *"correct intonation, stable speaking speed, and rhythm."* Because the *"stable speaking speed"* implies *"without noticeable pauses, repetition or stammering"*, the resulting fluency and prosody scores could be very simi-

Table 1: *Comparison of model performance. GT transcript free indicates, under the original design, whether the model was evaluated without using ground-truth transcripts.*

	GT transcript free	Word-level score (PCC)			Sentence-level score (PCC)			
		Accuracy	Stress	Total	Accuracy	Fluency	Prosody	Total
GOPT (medium.en)	✗	0.273	0.067	0.265	0.528	0.527	0.545	0.528
Lin *et al.* [19]	✓	-	-	-	**0.725**	-	-	-
Liu *et al.* [18]	✓	-	-	-	-	-	-	0.60
Liu *et al.* [20]	✓	-	-	-	-	**0.795**	-	-
vanilla SSL	✓	-	-	-	0.692 (std:0.006)	0.757 (std:0.010)	0.757 (std:0.009)	0.714 (std: 0.006)
MultiPA	✓	**0.427** (std: 0.008)	**0.239** (std:0.025)	**0.436** (std:0.010)	0.705 (std:0.009)	0.772 (std:0.010)	**0.764** (std:0.016)	**0.730** (std:0.006)

lar. To better differentiate between fluency and prosody, we revised the instruction when collecting multiPA data. For instance, we defined the highest fluency score as *"fluent without noticeable pauses or stammering"*, following the setting of speechocean762, but modified the highest prosody score to *"excellent prosody, expressive and well-modulated tone, enhancing communication with effective pitch and rhythm variations."* As a result, in multiPA data, the correlation between prosody and fluency is lower than that in the speechocean762 dataset.
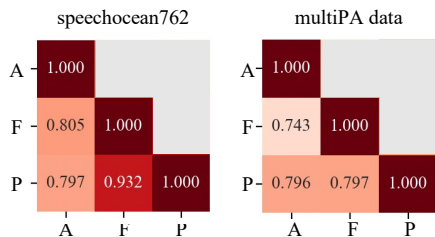

Figure 3: *Correlation between different pronunciation tasks.* A, F, P *refer to accuracy, fluency, and prosody, respectively.*

We evaluated MultiPA on the real-world data (Table 2). The results demonstrate MultiPA's ability to capture sentence-level accuracy and fluency proficiency, achieving correlations exceeding 0.6. However, we observed notably lower sentence-level prosody performance. This difference could be attributed to modifications made to the prosody assessment instructions during the collection of multiPA data, as previously discussed (Figure 3). We also examined the multitask learning approach by comparing the performance of the MultiPA structure with only sentence tasks (denoted as *-sentence*) and with only word-level tasks (denoted as *-word*). The experimental results show that multitask learning, incorporating both sentence-level and word-level assessment into the same model, can enhance performance compared to single-task learning. The word-level assessment is especially helpful for sentence-level accuracy because the word-level score mainly focuses on word accuracy.

Our analysis reveals limitations in the current model's ability to evaluate word-level scores, potentially due to the highly imbalanced word-level labels in the speechocean762 dataset [32]. To address this, we proposed a strategy that can leverage our current model effectively in real-world scenarios. We re-framed the word-level assessment from measuring the "accuracy score of a word's pronunciation" to a binary mispronunciation detection [33]: "whether the pronunciation of the

Table 2: *Performance on multiPA data (PCC).*

	Accuracy	Fluency	Prosody	Word
MultiPA	**0.62** (std:0.04)	**0.65** (std:0.05)	**0.49** (std:0.03)	**0.39** (std:0.03)
-sentence	0.57 (std:0.03)	0.60 (std:0.04)	0.48 (std:0.03)	-
-word	-	-	-	0.37 (0.02)

word needs improvement." We defined "a word that needs improvement" as a word that does not receive the highest score from all annotators. Because L2 learners might be overwhelmed by too many suggestions, we care about *precision* (i.e., whether the words suggested by the model actually need improvement) rather than recall (whether the model captured all the words that need improvement.) By setting the threshold less than *mean-std*, the model can achieve over 0.9 precision in word suggestions. This means that, if the model suggests a word with a score significantly lower than the scores from all data, then the suggested word highly likely needs improvement. Specifically, the precision is 0.926 (213/230 words) when setting the threshold as *mean-std*, and the baseline precision of suggesting all words is 0.759 (1092/1439 words). The threshold can be customized based on the pronunciation levels of all users or the sensitivity level at which the user desires feedback from the model.

5. Conclusion

We introduce MultiPA, a multi-task speech pronunciation assessment model that provides comprehensive assessment at both the sentence and word levels. Our study includes the collection of pilot data from real-world use cases, model evaluation on out-of-domain data, and analysis of correlations between different pronunciation tasks. The pilot data collected in this study will be made available for other researchers to evaluate their pronunciation assessment models. One limitation of MultiPA is that the assessed word-level scores in the open response scenario correspond to ASR-recognized words, but these words might be incorrect and vary between different runs. With the time alignment information of these words, MultiPA can still indicate specific parts of the speech signal that need improvement; however, how such information can effectively help learners to improve their pronunciation requires further investigation. In future, we plan to explore data augmentation or self-supervised methods and collect feedback from the model users to improve the model's performance further.

6. Acknowledgements

7. References

[1] F. Ehsani and E. Knodt, "Speech technology in computer-aided language learning: Strengths and limitations of a new CALL paradigm," *Language Learning & Technology*, vol. 2, pp. 45–60, 1998.

[2] K. B. Egan, "Speaking: A critical skill and a challenge," *Calico Journal*, pp. 277–293, 1999.

[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 *Proc. ICASSP 2022*, 2022, pp. 7262–7266.

[4] W. Liu, K. Fu, X. Tian, S. Shi, W. Li, Z. Ma, and T. Lee, "Leveraging phone-level linguistic-acoustic similarity for utterance-level pronunciation scoring," in *Proc ICASSP 2023*, 2023, pp. 1–5.

[5] K. Sheoran, A. Bajgoti, R. Gupta, N. Jatana, G. Dhand, C. Gupta, P. Dadheech, U. Yahya, and N. Aneja, "Pronunciation scoring with goodness of pronunciation and dynamic time warping," *IEEE Access*, vol. 11, pp. 15 485–15 495, 2023.

[6] H.-C. Pei, H. Fang, X. Luo, and X.-S. Xu, "Gradformer: A framework for multi-aspect multi-granularity pronunciation assessment," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 554–563, 2023.

[7] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, and B. Chen, "A hierarchical context-aware modeling approach for multi-aspect and multi-granular pronunciation assessment," in *Proc. INTERSPEECH 2023*, 2023, pp. 974–978.

[8] R. C. Shekar, M. Yang, K. Hirschi, S. Looney, O. Kang, and J. Hansen, "Assessment of non-native speech intelligibility using wav2vec2-based mispronunciation detection and multi-level goodness of pronunciation transformer," in *Proc. INTERSPEECH 2023*, 2023, pp. 985–988.

[9] 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.

[10] Y. Liang, K. Song, S. Mao, H. Jiang, L. Qiu, Y. Yang, D. Li, L. Xu, and L. Qiu, "End-to-end word-level pronunciation assessment with mask pre-training," in *Proc. INTERSPEECH 2023*, 2023, pp. 969–973.

[11] Z. Zhang, P. Vyas, X. Dong, and D. S. Williamson, "An end-to-end non-intrusive model for subjective and objective real-world speech assessment using a multi-task framework," in *Proc. ICASSP 2021*, 2021, pp. 316–320.

[12] R. E. Zezario, S.-w. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "MTI-Net: A multi-target speech intelligibility prediction model," in *Proc. INTERSPEECH 2022*, 2022, pp. 5463–5467.

[13] Y.-W. Chen and Y. Tsao, "InQSS: a speech intelligibility and quality assessment model using a multi-task learning network," in *Proc. INTERSPEECH 2022*, 2022, pp. 3088–3092.

[14] H. Chung, Y. K. Lee, S. J. Lee, and J. G. Park, "Spoken English fluency scoring using convolutional neural networks," in *Proc. O-COCOSDA 2017*, 2017, pp. 1–6.

[15] S. Mao, Z. Wu, J. Jiang, P. Liu, and F. K. Soong, "NN-based ordinal regression for assessing fluency of ESL speech," in *Proc. ICASSP 2019*. IEEE, 2019, pp. 7420–7424.

[16] L. Fontan, M. Le Coz, and C. Alazard, "Using the forward-backward divergence segmentation algorithm and a neural network to predict L2 speech fluency," in *Proc. Speech Prosody 2020*, vol. 2020, 2020, pp. 925–929.

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

[18] H. Liu, M. Shi, and Y. Wang, "Zero-shot automatic pronunciation assessment," in *Proc. INTERSPEECH 2023*, 2023, pp. 1009–1013.

[19] B. Lin and L. Wang, "Exploiting information from native data for non-native automatic pronunciation assessment," in *Proc. SLT 2022*. IEEE, 2022, pp. 708–714.

[20] W. Liu, K. Fu, X. Tian, S. Shi, W. Li, Z. Ma, and T. Lee, "An ASR-free fluency scoring approach with self-supervised learning," in *Proc. ICASSP 2023*, 2023, pp. 1–5.

[21] 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.

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

[23] A. Zahran, A. Fahmy, K. Wassif, and H. Bayomi, "Fine-tuning self-supervised learning models for end-to-end pronunciation scoring," *IEEE Access*, pp. 112 650–112 663, 2023.

[24] E. Kim, J.-J. Jeon, H. Seo, and H. Kim, "Automatic pronunciation assessment using self-supervised speech representation learning," in *Proc. INTERSPEECH 2022*, 2021, pp. 1411–1415.

[25] K. Fu, S. Gao, S. Shi, X. Tian, W. Li, and Z. Ma, "Phonetic and prosody-aware self-supervised learning approach for non-native fluency scoring," in *Proc. INTERSPEECH 2023*, 2023, pp. 949–952.

[26] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML 2023*. PMLR, 2023, pp. 28 492–28 518.

[27] J. Zhu, C. Zhang, and D. Jurgens, "Phone-to-audio alignment without text: A semi-supervised approach," *Proc ICASSP 2022*, pp. 8167–8171, 2022.

[28] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[29] 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," in *Proc. INTERSPEECH 2021*, 2021, pp. 3710–3714.

[30] H. Do, Y. Kim, and G. G. Lee, "Hierarchical pronunciation assessment with multi-aspect attention," in *Proc. ICASSP 2023*, 2023, pp. 1–5.

[31] 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 *Proc. APSIPA ASC 2022*, 2022, pp. 575–582.

[32] H. Do, Y. Kim, and G. G. Lee, "Score-balanced loss for multi-aspect pronunciation assessment," in *Proc. INTERSPEECH 2023*, 2023, pp. 4998–5002.

[33] H. Ryu, S. Kim, and M. Chung, "A joint model for pronunciation assessment and mispronunciation detection and diagnosis with multi-task learning," in *Proc. INTERSPEECH 2023*, 2023, pp. 959–963.

---

[1] https://labelstud.io/