PRONUNCIATION ASSESSMENT WITH MULTI-MODAL LARGE LANGUAGE MODELS

Kaiqi Fu[1], Linkai Peng[2], Nan Yang[1], Shuran Zhou[1]

[1]Zuoyebang AI Research
[2]University of Connecticut

ABSTRACT

Large language models (LLMs), renowned for their powerful conversational abilities, are widely recognized as exceptional tools in the field of education, particularly in the context of automated intelligent instruction systems for language learning. In this paper, we propose a scoring system based on LLMs, motivated by their positive impact on text-related scoring tasks. Specifically, the speech encoder first maps the learner's speech into contextual features. The adapter layer then transforms these features to align with the text embedding in latent space. The assessment task-specific prefix and prompt text are embedded and concatenated with the features generated by the modality adapter layer, enabling the LLMs to predict accuracy and fluency scores. Our experiments demonstrate that the proposed scoring systems achieve competitive results compared to the baselines on the Speechocean762 datasets. Moreover, we also conducted an ablation study to better understand the contributions of the prompt text and training strategy in the proposed scoring system.

Index Terms— Pronunciation Assessment, Large language Model, Automatic Speech Recognition, Multi-modal

1. INTRODUCTION

With the development of globalization, mastering multiple foreign languages has become an essential skill in life and work. Fluent and accurate speaking, one of the major parts of mastering a language, will greatly improve communication efficiency [1]. The huge demand for spoken language learning has promoted the development of computer-assisted pronunciation training(CAPT) technology [2, 3]. Automatic speech assessment, as one of the key technologies of CAPT, can evaluate learners' speech across various dimensions such as accuracy, fluency, and completeness. Spoken language learning is typically divided into two scenarios: the open scenario and the follow-up scenario. In the follow-up scenarios, learners are required to follow and speak out a given prompt text, while the open scenarios allow learners to express opinions freely based on a posed question. In this paper, we focus on the sentence-level pronunciation scoring of accuracy and fluency in the follow-up scenario.

Most previous studies of sentence-level spoken speech evaluation fall into two categories, namely align-based [4–10] and align-free [11–13]. Align-based systems require a DNN-HMM based Automatic speech recognition (ASR) model, which is used to perform forced alignment between the learners' speech and the corresponding prompt text. The alignment will get the phoneme sequence, the start and end times of each phoneme within the speech segment, and acoustic features extracted from the ASR system (such as hidden layer features or GOP features [14] calculated by posterior probability). After obtaining the above features, a simple scoring network is built (usually Transformer [5, 6, 10] or BLSTM [7–9]) to map these features to the pronunciation scores. On the other hand, an align-free

system is built with an end-to-end network, which directly maps the frame-level acoustic features and text-related features to scores. It significantly simplifies the feature extraction process and eliminates the impact of forced alignment on scoring. Past studies on align-free systems [11, 12] have predominantly emphasized evaluating speech fluency [11, 12] and prosody [12]. In [13], the focus is on accuracy at both word and phoneme levels. Align-free system may have significant potential for sentence accuracy scoring of spoken language.

Recently, the emergence of large language models (LLMs) [15–17] has demonstrated powerful language understanding and content generation capabilities, introducing some abilities that are not present in smaller-scale language models [18], such as instruction following and in-context learning. Some previous research [19–22] has used LLMs to implement text-related scoring tasks. For example, the study [21] asks the LLM first to analyze and interpret the provided materials, and then determine the final score. This method not only generates score results but also provides detailed comments on the student's answers. In addition, the potential of LLM is not limited to text-related tasks. They can handle and understand non-text information such as audio and images well, thereby bridging the gap between different modalities. In the field of speech, the multi-modal models based on LLM usually consist of three main components: a speech encoder, a modality adapter, and a language model. Some representative multi-modal models [23–26], such as SALMONN [23], Qwen-Audio [26], combine speech features and text embedding and feed them into a decoder-only language model, demonstrating powerful capabilities in speech understanding and recognition tasks.

Inspired by the above studies on multi-modal models and the positive effects of LLMs in text-related scoring tasks, we explore the use of multi-modal models for spoken pronunciation assessment tasks in this paper. The model training process is divided into two stages. The first stage is to train the model in the speech recognition task, and then the model will be fine-tuned based on the task-specific scoring data in the second stage. When assessing a learner's speech, the speech encoder was used to map the speech into contextual features. These features contain the prosody and content information in spoken language. Then the adapter layer maps this feature to the same space as the text embedding to achieve the modality adaptation. Finally, the prompt text and the assessment task-specific prefix are embedded and concatenated with the speech features generated by the adapter layer to allow the LLM to predict the accuracy and fluency scores. The contributions of this paper are summarized as follows:

- To our knowledge, this paper is the first to use a multi-modality model based on the LLM for pronunciation assessment.

- The proposed multi-modal method belongs to the align-free system category and has achieved competitive performance
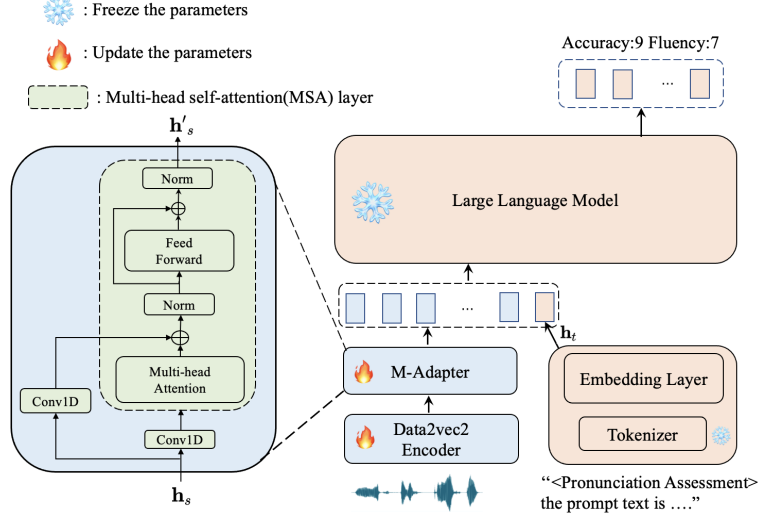
Fig. 1: The overall pipeline for L2 learner's pronunciation assessment in the proposed system.

compared to traditional align-based and align-free systems, which also verifies the effectiveness of this method.

- We contributed to a multi-modal system that achieved state-of-the-art (SOTA) results in ASR when we finished the training in the first stage.

2. PROPOSED METHOD

This section describes the proposed multi-modal pronunciation assessment method. First, we outline the three components of our system: the pre-trained acoustic encoder from data2vec2, the m-adapter as a modality adapter, and the LLM-based pronunciation assessment module. Then, we introduce the two-stage training process. The overall pipeline for pronunciation assessment is illustrated in Figure 2.

2.1. Speech Encoder

Data2vec2 [27] was employed as the audio encoder because of its versatility, efficiency, and outstanding performance in speech tasks. It comprises several convolutional blocks and multi-layer transformers. The convolutions in each block have 512 channels with strides of (5, 2, 2, 2, 2, 2, 2) and kernel sizes of (10, 3, 3, 3, 3, 2, 2). It transforms 16 kHz raw audio into a sequence of embeddings with a stride of 20 ms and a receptive field of 25 ms.

For each sample, given the raw audio $\mathbf{x_s}$, the speech features $\mathbf{h_s}$ was firstly extracted from the last hidden layer through the data2vec2 based speech encoder, which can be written as:

$$\mathbf{h_s} = SpeechEncoder(\mathbf{x_s}), \tag{1}$$

2.2. Modality adapter

The m-adapter [28] was firstly introduced for speech translation tasks. We employed it as a modality adapter in our model to connect the speech and text modalities. It consists of two pooling convolutions and a multi-head self-attention (MSA) layer. The pooling operations are used to modify the speech sequence length, while the MSA captures global information.

This work follows the paper [29] by replacing the three independent pooling modules for Q, K, and V with a shared pooling module to streamline the architecture and improve the efficiency of the model. In general, given the speech features $\mathbf{h_s}$ extracted from the speech encoder, a convolutional pooling operation is first used to adapt the speech length. These adapted features are then fed into the MSA layers to obtain the modality adapter features $\mathbf{h'_s}$, expressed as follow:

$$\mathbf{h'_s} = Adapter(\mathbf{h_s}), \tag{2}$$

The intuition of this module is that the number of features in the time domain of the speech signal is much larger than the length of the text. A scaling method needs to be applied to the speech features to balance the amount of information between the two modalities.

2.3. Pronunciation assessment with LLM

Decoder-only LLM serves as a foundational component in the proposed method. We expect it to understand the learner's speech, analyse the difference between speech and the canonical text, and then make decisions regarding accuracy and fluency scores accordingly.

For the text input of the LLM, we define the task-specific prefix as "$< Pronunciation\ Assessment >$". Additionally, in follow-up scenarios, The prompt text, as an important reference cue, can be considered as the text input of the LLM. Therefore, the final text input of LLM is defined as "$< Pronunciation\ Assessment > the\ prompt\ text\ is\ ...$", and then the text input will be transformed into embedding features denoted as $\mathbf{h_t}$. Finally, we concatenate the embedding features with adapter-modified speech features $\mathbf{h'_s}$ and then fed the concatenated features into the LLM to obtain the final prediction text $\mathbf{y}$ through a token decoding operation as equation(3), where the [;] denotes the concatenation of two vectors.

$$\mathbf{y} = LLM([\mathbf{h'_s}; \mathbf{h_t}]), \tag{3}$$

2.4. Two stages training strategy

The proposed method adopts a two-stage training strategy. Previous studies have shown that the performance of traditional align-based

Table 1: The WER results on Librispeech dataset comparing with other multi-modal systems. We briefly outlined the differences in these systems regarding speech encoders and modality adapters. Extra data indicates that these systems used additional training data besides Librespeech.

Model	Speech encoder	Modality adapter	Extra data	Test-clean	Test-other
SALMONN [23]	Whisper [30], Beats	Q-former	✓	2.4	5.4
WavLLM [24]	Whisper, WavLM	Linear	✓	2	4.8
Qwen-Audio [26]	Whisper	Linear	✓	2.04	4.19
SALM [25]	Hubert [31]	Linear	✗	1.94	3.81
Proposed	Data2vec2	M-adapter	✗	**1.73**	**3.7**

methods can be affected by the ASR system, as accuracy scoring heavily relies on speech content recognition. Therefore, in the first stage, we train the multi-modal model on the ASR task to make the connection between the two modalities more robust, which may benefit the pronunciation assessment task. We took the ASR task-specific prefix "$< transcript >$" as the text input and then utilized limited ASR data to complete the ASR task training.

In the second stage, we fine-tuned the model trained in the first stage using a limited amount of scoring data based on the "$< pronunciation\ assessment\ >$" assessment task-specific prefix and prompt text. For the ground truth, we structured it as "$accuracy : 9\ fluency : 7$", aiming for LLM to output scores in such a format to achieve multi-task scoring for both accuracy and fluency.

3. EXPERIMENTAL SETUP

3.1. Datasets

Two public datasets were used in this study. Librispeech [32], a commonly used ASR corpus, was utilized to train the multi-modal ASR in the first stage. It consists of about 1,000 hours of 16kHz read English speech derived from read audiobooks, with the dev-clean and dev-other subsets used as the validation set, and the test-clean and test-other subsets used as the test sets. To evaluate fluency and accuracy scoring in the second stage, we conducted experiments on Speechocean762 [33], an open-source speech assessment corpus consisting of 5,000 English utterances collected from 250 Chinese speakers. The whole scoring dataset is divided into 2,500 sentences for training and 2, 500 for testing. The score distribution of the test set is shown in Figure 2. The x-axis represents the human-labeled scores, and the y-axis represents the number of occurrences for each corresponding score.

3.2. Setup

The weight of the data2vec2-based speech encoder is initialized from data2vec2-large [1] and the Qwen-based LLM [17] is initialized using pre-trained weights derived from Qwen-7B [2]. For the M-adapter, we chose one MSA layer with 16 heads. The kernel size and stride of the convolution were both set to 2, aiming to halve the length of speech features extracted from the speech encoder.

We used 8 A800s for the experiments. Each stage was trained for a maximum of two epochs on the corresponding task data. The ASR and pronunciation scoring training took about 9 and 0.2 hours, respectively. During both stages, the parameters of LLM were frozen,
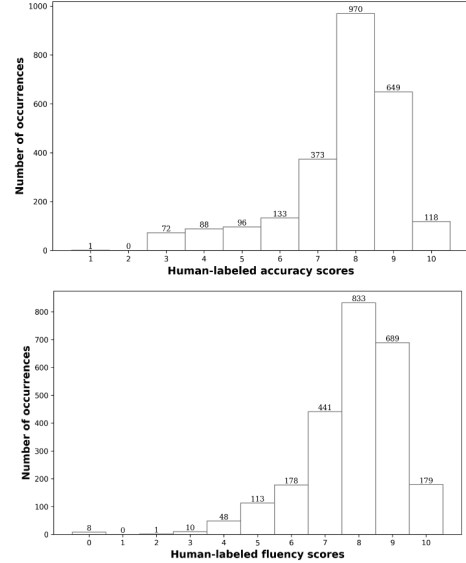

Fig. 2: The distribution of the human-labeled scores

and the parameters of the speech encoder and modality adapter were updated.

4. RESULTS AND DISCUSSION

This section first presents the ASR result of the proposed system from the first training stage. Then, we demonstrate the pronunciation assessment results after fine-tuning the model based on the ASR training. Additionally, we present ablation experiments to explore the impacts of ASR training and prompt text. The scoring system performance was evaluated using the Pearson correlation coefficient (PCC) between the predicted scores and the human-labeled scores.

4.1. Performance of ASR

ASR task was often used to evaluate the performance of multi-modal systems [23–26]. In this subsection, we selected several representative multi-modal systems with strong ASR performance as our baselines, including SALMONN [23], WavLLM [24], Qwen-Audio [26], and SALM [25].

The comparison results between the proposed system and these baseline systems are shown in Table 1. For the result of the proposed method, the predict and ground-truth texts were normalized using

---

Table 2: The PCC results on Speechocean762 dataset comparing with other baseline systems.

Type	Methods	Fluency	Accuracy
Align-based	GOPT [5]	0.753	0.714
	Multi-task [9]	0.73	0.694
	MultiPA [4]	0.772	0.705
	HiPAMA [10]	0.762	0.730
Align-free	SSL [12]	0.78	-
	SSL+IDX [11]	0.795	-
	Proposed	**0.777**	**0.713**

Table 3: The WER results of speech encoder models in the proposed method.

Method	Speech encoder	Test-clean	Test-other
Proposed	Whisper	2.28	6.34
	Hubert	1.83	**3.68**
	Data2vec2	**1.73**	3.7

Whisper's text normalization method [3] before computing WER. The results suggest that the proposed method outperforms all the baselines in terms of WER in the Librispeech test sets. This observation encourages us greatly for our upcoming pronunciation assessment tasks.

In addition, we briefly compared several different speech encoders based on the proposed method, with results shown in Table 3. The findings reveal that the data2vec2 model as the optimal choice for speech encoding performs significantly better compared to the performance of Hubert [4] and Whisper [5].

4.2. Results of pronunciation assessment

In this subsection, we present the performance of the proposed multi-modal model in pronunciation scoring, which is an align-free end-to-end approach. For a fair comparison, we selected two types of baseline systems: one is align-free end-to-end systems based on SSL features and the other one is align-based systems that rely on traditional DNN-HMM based ASR model trained with Librispeech dataset. We did not consider the align-based system using SSL features [6] as our baseline, since such an approach would make the entire align-based pipeline more complicated and less feasible for practical implementation.

We first conducted performance comparisons between the proposed multi-modal pronunciation scoring system and the align-based baseline systems presented in Table 2. The leading results validate the effectiveness of our proposed method. Moreover, compared to the complex pipelines of align-based systems, our proposed method offers a simpler way to evaluate learners' pronunciation.

Then, we evaluate the performance comparisons with the align-free methods. The proposed method demonstrates competitive results with baseline systems in fluency scoring. Additionally, it fills the gap in accuracy scoring for align-free approaches on the Speechocean762 dataset.

4.3. Ablation studies

In this subsection, we conducted experiments to determine the relative importance of ASR training in the first stage and prompt text in the proposed system for pronunciation scoring. The results are presented in Table 4.

Firstly, we can clearly see that these two factors slightly improve fluency scoring performance in system (a). The two factors focus more on the content of the speech. As verified in [7, 8], prosodic information is more important than speech content information for fluency scoring. Moreover, compared to SSL system [12], SSL+IDX [11]showed a slight performance difference due to the additional information provided by phoneme IDs, which represent speech content. This further confirms that the content does not have a significant impact on fluency scoring.

Compared to speech fluency, accuracy is more significantly affected by these factors. The results of systems (b) and (c) show that the impact of introducing the prompt text is greater than that of ASR training. Additionally, system (d) found that not using any strategies resulted in poor accuracy scoring performance.

Table 4: The PCC results of ablation studies for the proposed method.

	Factors	Fluency	Accuracy
(a)	ASR training + Prompt text	**0.777**	**0.713**
(b)	ASR training only	0.763	0.698
(c)	Prompt text only	0.768	0.665
(d)	-	0.759	0.654

5. CONCLUSIONS

This paper introduces the multi-modal approach to develop an end-to-end pronunciation assessment system for scoring fluency and accuracy of learners' L2 speech, such that it does not require any word and phone-level alignment information as input. The experiment results show that our approach achieves a competitive performance compared with previous methods, including those systems that utilize alignment information. This work can be used as the baseline system of end-to-end multi-modal approach and align-free related approach for pronunciation assessment tasks. Additionally, our proposed method achieved state-of-the-art (SOTA) performance in the ASR task, demonstrating the superior effectiveness of our multi-modal system.

In the future, we will investigate the chain-of-thought (COT) [34] ability of the proposed system to provide step-by-step explanations for the model's scoring decisions.

6. REFERENCES

[1] Alex Housen and Folkert Kuiken, "Complexity, accuracy, and fluency in second language acquisition," *Applied linguistics*, vol. 30, no. 4, pp. 461–473, 2009.

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

[3] Jonás Fouz-González, "Trends and directions in computer-assisted pronunciation training," *Investigating English pronunciation: Trends and directions*, pp. 314–342, 2015.

[4] Yu-Wen Chen et al., "Multipa: a multi-task speech pronunciation assessment system for a closed and open response scenario," *arXiv preprint arXiv:2308.12490*, 2024.

[5] Yuan Gong et al., "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.

[6] Fu-An Chao et al., "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)*. IEEE, 2022, pp. 575–582.

[7] Kaiqi Fu et al., "Using Fluency Representation Learned from Sequential Raw Features for Improving Non-native Fluency Scoring," *Proc. Interspeech 2022*, pp. 4337–4341, 2022.

[8] Kaiqi Fu et al., "Phonetic and Prosody-aware Self-supervised Learning Approach for Non-native Fluency Scoring," in *Proc. INTERSPEECH 2023*, 2023, pp. 949–953.

[9] Jeremy Heng Meng Wong et al., "Variations of multi-task learning for spoken language assessment," in *Proc. Interspeech 2022*, 2022, pp. 4456–4460.

[10] Heejin Do et al., "Hierarchical pronunciation assessment with multi-aspect attention," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[11] Wei Liu et al., "An ASR-free Fluency Scoring Approach with Self-Supervised Learning," *arXiv preprint arXiv:2302.09928*, 2023.

[12] Eesung Kim et al., "Automatic Pronunciation Assessment using Self-Supervised Speech Representation Learning," in *Proc. Interspeech 2022*, 2022, pp. 1411–1415.

[13] Yukang Liang et al., "End-to-End Word-Level Pronunciation Assessment with MASK Pre-training," in *Proc. INTERSPEECH 2023*, 2023, pp. 969–973.

[14] Wenping Hu et al., "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.

[15] OpenAI, "GPT-4 Technical Report," 2024.

[16] Hugo Touvron et al., "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[17] Jinze Bai et al., "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[18] Shervin Minaee et al., "Large language models: A survey," *arXiv preprint arXiv:2402.06196*, 2024.

[19] Stefano Bannò et al., "Can GPT-4 do L2 analytic assessment?," *arXiv preprint arXiv:2404.18557*, 2024.

[20] Kevin P Yancey et al., "Rating short l2 essays on the cefr scale with gpt-4," in *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 2023, pp. 576–584.

[21] Changrong Xiao et al., "From Automation to Augmentation: Large Language Models Elevating Essay Scoring Landscape," *arXiv preprint arXiv:2401.06431*, 2024.

[22] Shen Wang et al., "Large language models for education: A survey and outlook," *arXiv preprint arXiv:2403.18105*, 2024.

[23] Changli Tang et al., "SALMONN: Towards generic hearing abilities for large language models," in *The Twelfth International Conference on Learning Representations*, 2024.

[24] Shujie Hu et al., "Wavllm: Towards robust and adaptive speech large language model," *arXiv preprint arXiv:2404.00656*, 2024.

[25] Ziyang Ma et al., "An Embarrassingly Simple Approach for LLM with Strong ASR Capacity," *arXiv preprint arXiv:2402.08846*, 2024.

[26] Yunfei Chu et al., "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.

[27] Alexei Baevski et al., "Efficient self-supervised learning with contextualized target representations for vision, speech and language," in *International Conference on Machine Learning*. PMLR, 2023, pp. 1416–1429.

[28] Jinming Zhao et al., "M-Adapter: Modality Adaptation for End-to-End Speech-to-Text Translation," in *Proc. Interspeech 2022*, 2022, pp. 111–115.

[29] Loïc Barrault et al., "SeamlessM4T-Massively Multilingual & Multimodal Machine Translation," *arXiv preprint arXiv:2308.11596*, 2023.

[30] Alec Radford et al., "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28492–28518.

[31] Wei-Ning Hsu et al., "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.

[32] Vassil Panayotov et al., "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[33] Junbo Zhang et al., "Speechocean762: An Open-Source Non-Native English Speech Corpus for Pronunciation Assessment," in *Proc. Interspeech 2021*, 2021, pp. 3710–3714.

[34] Jason Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.