ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10887601

Integrating Potential Pronunciations for Enhanced Mispronunciation Detection and Diagnosis Ability in LLMs

Minglin Wu, Jing Xu, Xueyuan Chen, Helen Meng
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
Hong Kong SAR, China
{minglinwu, jingxu.se}@link.cuhk.edu.hk, {xychen, hmmeng}@se.cuhk.edu.hk

***Abstract*—Large Language Models (LLMs) have exhibited significant potentials across various tasks. However, how to leverage the power of LLMs in the mispronunciation detection and diagnosis (MDD) task is still under-explored. In this paper, we propose a PP-ATP model, which integrates potential pronunciations covering common mispronunciations into the prompt part of LLMs, to enhance the MDD ability of LLMs in second language (L2) English. Specifically, the proposed PP-ATP model is composed of an audio encoder, an LLM decoder, and an adapter. Taking speech representations from the audio encoder as the audio prompt and reference sentence with canonical and potential pronunciations as text prompt, the LLM decoder is adapted to predict the actual pronunciation in the given L2 speech. Experiments show that our PP-ATP model achieves new state-of-the-art (SOTA) performance in MDD on CU-CHLOE corpus, confirming the effectiveness of potential pronunciation integration.**

***Index Terms*—large language model, mispronunciation detection and diagnosis, potential pronunciations, L2 English.**

I. Introduction

Mispronunciation detection and diagnosis (MDD) plays a key role in Computer-Assisted Pronunciation Training (CAPT) systems. Mispronunciation detection aims to detect incorrect pronunciations and mispronunciation diagnosis gives pinpointed feedback to second language (L2) learners about their incorrect pronunciations. Figure 1 depicts a regular MDD pattern in L2 English. First, the CAPT system prompts the learner to read a sentence. Then, the MDD system recognizes the phones articulated in the speech. Finally, the system compares the recognized phone sequence with the canonical phone sequence to detect mispronunciations and provide diagnosing information.

Pronunciation scoring and phone recognition constitute predominant methodologies in MDD. Research in pronunciation scoring either adopts different confidence measures to score pronunciations or identifies mispronunciations in an end-to-end way. For instance, [1]–[3] developed Goodness of Pronunciation (GOP) and its derivatives from acoustic models to measure pronunciations and identify pronunciations with low scores as mispronunciations. [4], [5] format mispronunciation detection as a binary classification task. They built end-to-end

This work is supported by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd., a CUHK-led under the InnoHK scheme of Innovation and Technology Commission.

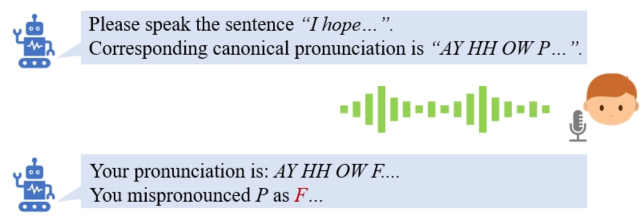

Fig. 1. An example illustrating MDD in L2 English. CMU Pronunciation Dictionary is used to represent the pronunciations.

frameworks to directly classify each phone as correct pronunciation or mispronunciation. However, pronunciation scoring methods inherently cannot diagnose mispronunciations.

To also provide diagnosing information, researchers introduced phone recognition in L2 English for MDD. [6], [7] built extended recognition networks (ERNs) which cover both canonical pronunciations and possible mispronunciations to facilitate the decoding process. However, ERNs are incapable of covering all possible pronunciations. To overcome this limitation, researchers propose to integrate free phone recognition systems in MDD. Leung *et al.* [8] designed a CNN-RNN-CTC architecture to enhance phone recognition for MDD. Feng *et al.* [9] proposed a sentence-dependent MDD system that incorporates the reference sentence as additional textual information to improve phone recognition. [10], [11] fine-tuned the pretrained wav2vec 2.0 [12] model with connectionist temporal classification (CTC) loss [13] to perform phone recognition and MDD tasks in L2 English speech. The speech representation learned by wav2vec 2.0 encodes abundant phonetic information, making the wav2vec2-CTC method perform well across multiple datasets. Kheir *et al.* [14] proposed L1-MultiMDD model which integrates an auxiliary L1-L2 network, a phoneme encoder, and a speech encoder to recognize phones in L2 speech. Wan *et al.* [15] introduced model-agnostic meta-learning (MAML) to train a phone recognition model for MDD.

Nowadays, text-centric or multi-modal large language models (LLMs), such as GPT4, Gemini, and Qwen series [16]–[18], have been assisting people in daily life and work. More and more researchers are integrating powerful LLMs in their methods to get performance enhancement in various tasks. For example, Driess *et al.* integrated visual and 3D representations

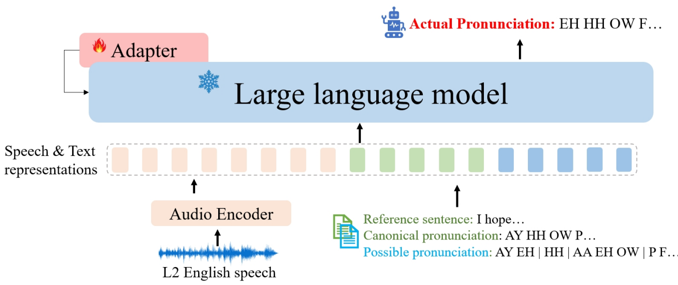

Fig. 2. Overall structure of the proposed PP-ATP model.

into PaLM [19] to develop PaLM-E [20] to facilitate robotics tasks. Su *et al.* built PandaGPT [21] by attaching a multi-modal encoder ImageBind [22] to Vicuna [23] to cope with cross-modal tasks. AudioGPT [24] utilizes the LLM as a coordinator to direct different audio foundation models to execute various tasks. Besides, some researchers [25]–[27] have tried to empower LLMs with automatic speech recognition (ASR) ability by simply attaching an audio encoder to LLMs. This makes LLMs able to process audio modality and achieve competitive performance to traditional ASR methods. In SpeechGPT [28], researchers adopted HuBERT [29] to tokenize speech into discrete tokens. Then they merge these speech tokens into the vocabulary of LLaMA [30] and adapt LLaMA to perform speech related tasks. To empower LLMs with MDD ability, Wu *et al.* proposed ATP-LLM model [31], which takes continuous speech representations obtained from an audio encoder as the audio prompt and canonical pronunciation as the text prompt, to perform phone recognition and MDD tasks in L2 English speech. ATP-LLM achieved state-of-the-art (SOTA) performance on CU-CHLOE [32] corpus.

Inspired by the idea of ERNs and the success of ATP-LLM, we propose a PP-ATP model to integrate potential pronunciations into the prompt part of LLMs to enhance the MDD ability of LLMs in L2 English.The proposed PP-ATP model is composed of an audio encoder, an LLM decoder, and an adapter. The audio encoder transforms L2 English speech into continuous speech representations. These speech representations and the reference sentence with canonical and potential pronunciations serve as audio and text prompt for the LLM decoder, making it generate actual pronunciations in the L2 English speech. With the alignment between the recognized phone sequence and the canonical phone sequence, we detect mispronunciations and give pinpointed feedback. Extensive experiments show the effectiveness of our proposed PP-ATP model. The contributions of this paper are:

- This paper explores to invoke and enhance the knowledge about pronunciations in LLMs for addressing the challenging L2 English.
- This paper proposes PP-ATP model which integrates potential pronunciations to enhance the MDD ability of LLMs. Experiments show that our proposed method achieves new SOTA performance on CU-CHLOE.
- This paper conducts extensive experiments to explore the factors influencing MDD abilities in LLMs.

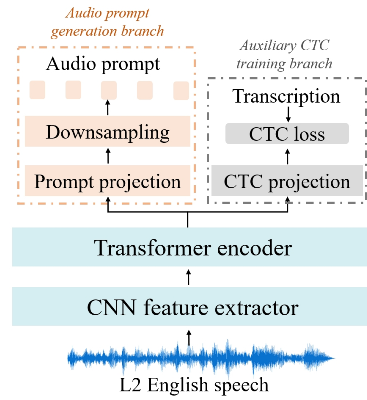

Fig. 3. The structure of the audio encoder.

This paper is organized as follows: Section II presents proposed method. Experiments and evaluations are detailed in Section III. Section IV concludes this paper.

II. Methods

A. Model architecture

As shown in Figure 2, our proposed PP-ATP model comprises an audio encoder, a large language model, and an adapter.

1) Audio encoder: The audio encoder transforms L2 English speech into continuous speech representations which are then served as the audio prompt of LLM. As shown in Figure 3, the audio encoder is composed of four components: a CNN feature extractor, a Transformer encoder, an audio prompt generation branch, and an auxiliary CTC training branch. The CNN feature extractor, consisting of multiple temporal convolutional layers, transforms raw L2 English speech into latent representations. Subsequently, the Transformer encoder transforms these latent representations into contextualized speech representations (denoted as $\mathcal{C}$), serving as input of the audio prompt generation branch, and the auxiliary CTC training branch.

The audio prompt generation branch contains a prompt projection module and a down-sampling module. The prompt projection module leverages linear layers to project $\mathcal{C}$ into speech representations whose dimension equals to the input dimension of LLMs. The down-sampling module adopts temporal convolutional layers to down-sample these speech representations and generate the final audio prompt. The auxiliary CTC training branch, composed of linear layers, projects $\mathcal{C}$ into phone probabilities, which serve as the input of the auxiliary CTC loss. The target of the CTC loss is the manually annotated phones of the L2 English speech. The auxiliary CTC training branch is removed during inference.

2) Large Language Model: We employ Qwen2, a series of decoder-only LLMs trained on trillions of tokens as our LLM. As is shown in Figure 2, The LLM decoder takes in the audio prompt and the text prompt and outputs the corresponding actual pronunciation articulated. The audio prompt is generated through the audio encoder while the text prompt is the reference sentence with canonical pronunciation and potential pronunciations.

We adopt Low-Rank Adaptation (LoRA) [33] as the adapter to efficiently adapt the LLM, keeping the parameters of the LLM frozen.

B. Potential pronunciation acquirement

To make LLMs focus on common pronunciation patterns actually occurred during L2 English learning, we derive potential pronunciations from an L2 English speech corpus. We achieve this by following steps. Take the word *hope* for instance. First, we obtain manually annotated phone sequences of the word *hope* from the L2 English corpus. For example, some of the pronunciations of it are *[HH OW P], [HH AA F], and [HH OW]*. Then, to alleviate the large computational cost caused by long input sequence, we merge phone sequences with the same length together, using | to mark the phone positions. That is, we merge *[HH OW P], [HH AA F]* together to get *[HH | OW AA | P F]*, and merge *[HH OW]* itself together to get *[HH | OW]*. Finally, we concatenate the merged phone sequences of different lengths, using comma to bind them. That is, for the word *hope*, we finally get *[HH | OW AA | P F, HH | OW]*. In this way, we obtain potential pronunciations for all words in the corpus.

C. Training strategy

During training, we jointly optimize the audio encoder and the adapter with the actual pronunciation prediction loss $\mathcal{L}_{AP}$, potential pronunciation prediction loss $\mathcal{L}_{PP}$, and the CTC loss $\mathcal{L}_{CTC}$. The overall loss is computed with Equation (1):

$$\mathcal{L} = \mathcal{L}_{AP} + \lambda_1 \mathcal{L}_{PP} + \lambda_2 \mathcal{L}_{CTC} \tag{1}$$

where λ_1 and λ_2 are hyper-parameters that manipulate the weight of the potential pronunciation prediction loss and the CTC loss, respectively.

The actual pronunciation prediction loss $\mathcal{L}_{AP}$ aims to guide LLMs to correctly recognize the desired actual articulated pronunciations in the L2 English speech. Equation (2) shows the calculation of $\mathcal{L}_{AP}$.

$$\mathcal{L}_{AP} = -\sum_{i=T_{sap}}^{T_{eap}} \log p(y_{AP}^{i} | \boldsymbol{y}_{AP}^{<i}; \boldsymbol{\theta}) \tag{2}$$

where $\boldsymbol{y}_{AP}$ is the actual pronunciation (manually annotated phone sequences) of L2 English speech. T_{sap} and T_{eap} are the starting and ending position of $\boldsymbol{y}_{AP}$. $\boldsymbol{\theta}$ represents the parameters of the entire model except those of the auxiliary CTC training branch in the audio encoder.

To invoke and enhance the knowledge about pronunciations encoded in LLMs, we also make PP-ATP predict the potential pronunciations obtained following the steps described in Section II-B. The potential pronunciation prediction loss $\mathcal{L}_{PP}$ is calculated by Equation (3):

$$\mathcal{L}_{PP} = -\sum_{i=T_{spp}}^{T_{epp}} \log p(y_{PP}^{i} | \boldsymbol{y}_{PP}^{<i}; \boldsymbol{\theta}) \tag{3}$$

TABLE I
DETAILS OF CU-CHLOE

	Training	Validation	Test	Total
Speakers	144	23	43	210
Hours	24	3.6	7	34.6

where $\boldsymbol{y}_{PP}$ is the derived potential pronunciations. T_{spp} and T_{epp} are the starting and ending position of $\boldsymbol{y}_{PP}$. The CTC loss $\mathcal{L}_{CTC}$ is calculated using Equation (4):

$$\mathcal{L}_{CTC} = -\log p(\boldsymbol{y}_{AP} | \boldsymbol{x}; \boldsymbol{\theta}_{AE}) \tag{4}$$

Here, $\boldsymbol{x}$ denotes the input L2 English speech. The term $\boldsymbol{\theta}_{AE}$ represents the parameters of the CNN feature extractor, the Transformer encoder, and CTC projection module (See Figuire 3).

III. EXPERIMENTS

A. Datasets

We adopt CU-CHLOE [32] as the L2 English corpus for all experiments. CU-CHLOE contains 34.6 hours of L2 English speech recorded by 100 Cantonese and 110 Mandarin speakers. The reference sentences speakers read include (a) The AESOP's fable "The North Wind and the Sun", (b) Confusable words, (c) Minimal pairs, and (d) Phonemic sentences. The actual articulated phones are annotated by linguists using native English phonemes. Table I details the split of CU-CHLOE.

B. Experimental setups

The structure of the CNN feature extractor and Transformer encoder is identical to that of wav2vec 2.0 base. Seven temporal convolutional layers constitute the CNN feature extractor, transforming the raw speech signal into latent feature representations at a frequency of 49 Hz. The channels, strides, and kernels of these layers are 512, (5, 2, 2, 2, 2, 2, 2), and (10, 3, 3, 3, 3, 2, 2), respectively. 12 Transformer blocks constitute the Transformer encoder. The attention heads, model dimension, and feed-forward dimension of each Transformer block are 8, 768, and 3072, respectively. We fine-tune the pre-trained wav2vec 2.0 model on CU-CHOLE corpus with CTC loss (following the recipe in [10]) to initialize the weights of the CNN feature extractor and Transformer encoder.

The prompt projection module is one linear layer with an input dimension of 768 and an output dimension of 4096, making the dimension of resulting speech representations match the input dimension of the adopted LLM. The down-sampling module is composed of one temporal convolutional layer whose input and output channel are both 4096. To achieve down-sapmling rate varying from 2x to 5x, we set its strides to (2, 5) and kernels to (3, 10). The CTC projection module is one linear layer with input and output channel being 768 and 151646, respectively.

We adopt Qwen2-0.5B as the LLM in this paper. We set the rank of LoRA to 16 and 32 to adapt the query, key, value, and

TABLE II
PERFORMANCE OF DIFFERENT APPROACHES. R REFERS TO THE RANK IN LORA, CONTROLLING THE NUMBER OF TRAINABLE PARAMETERS IN QWEN2.

	Methods	Trainable parameters	PER (%)↓	FRR (%)↓	FAR (%)↓	Precision (%)↑	Recall (%)↑	F1 (%)↑	DIAA (%)↑
Baselines	wav2vec2-CTC	-	8.51	4.63	20.82	81.78	79.18	80.46	93.81
	ATP-LLM R=16	2.16M	7.92	3.96	20.89	83.95	79.11	81.46	94.12
	ATP-LLM R=32	4.33M	7.89	3.96	20.63	83.99	79.37	81.61	94.03
Proposed	PP-ATP R=16	2.16M	7.83	3.96	20.48	84.01	79.52	81.71	94.15
	PP-ATP R=32	4.33M	**7.71**	**3.88**	**20.34**	**84.30**	**79.66**	**81.91**	**94.16**

TABLE III
INVESTIGATE THE IMPACT OF DOWN-SAMPLING RATE ON SYSTEM PERFORMANCE. R IN LORA IS SET TO 32.

	Methods	PER (%)↓	F1 (%)↑	DIAA (%)↑
Baselines	wav2vec2-CTC	8.51	80.46	93.81
	ATP-LLM (20ms)	7.89	81.61	94.03
Proposed	PP-ATP (100ms)	10.06	74.91	93.39
	PP-ATP (40ms)	7.78	81.73	94.09
	PP-ATP (20ms)	**7.71**	**81.91**	**94.16**

output projection matrices within the self-attention modules of Transformer blocks.

During training, the CNN feature extractor is kept frozen. The Transformer encoder is frozen for the initial 5000 steps. λ_1 and λ_2 in Equation (1) are set to 0.001 and 1, respectively. To avoid over-fitting, in each update, we randomly select 10% of the words from each reference sentence for the calculation of $\mathcal{L}_{PP}$ (See Equation (3)). The whole model is trained for 4 epochs with 2 A6000 GPUs with Adam optimizer. The learning rate schedule is *WarmupDecayLR* [34] with a peak learning rate of 1e-4.

C. Evaluation metrics

We adopt Phone Error Rate (PER) for evaluating the phone recognition ability in L2 English speech. Following [35], we align the canonical, transcribed, and recognized phone sequences to obtain True Acceptance (TA), False Acceptance (FA), False Rejection (FR), True Rejection (TR), Correct Diagnosis (CD), and Diagnosis Error (DE). For TA, both the transcribed and the recognized phone match the canonical phone. For FA, the recognized phone matches the canonical one, while the transcribed phone does not. FR means that the transcribed phone matches the canonical phone, but the recognized phone does not. TR means both the transcribed and recognized phones differ from the canonical one. Further, in TR cases, CD and DE indicate whether the recognized phone equals to the transcribed phone or not, respectively.

We then calculate the False Rejection Rate (FRR, FR/(TA+FR)), False Acceptance Rate (FAR, FA/(FA+TR)), Precision (TR/(TR+FR)), Recall (TR/(TR+FA)), F1 score, and Diagnosis Accuracy (DIAA, CD/(CD+DE)) to evaluate the MDD performance.

D. Main results

We show the performance of different methods on CU-CHLOE in Table II. We take wav2vec2-CTC [10] and ATP-LLM [31] methods as baselines. When the rank R of LoRA is set to 16 and 32, the trainable parameters of the adopted Qwen2-0.5B are 2.16 and 4.33 million, respectively.

From Table II, we can see that the proposed PP-ATP model with R = 32 configuration outperforms both baselines across all evaluation metrics. Specifically, our PP-ATP model achieves a PER of 7.71%, an F1 score of 81.91%, and a DIAA of 94.16%, outperforming wav2vec2-CTC and ATP-LLM (R=32) methods whose respective PERs are at 8.51% and 7.89%, F1 scores at 80.46% and 81.61%, and DIAAs at 93.81% and 94.03%.

In addition, we can see that increasing R (number of trainable parameters) leads to performance improvement for PP-ATP model. For instance, the F1 score of the proposed PP-ATP improves from 81.71% (R = 16) to 81.91% (R = 32). Notably, our PP-ATP model performs better than baselines even with a smaller R. For example when R = 16, the F1 score of PP-ATP is 81.71%, outperforming wav2vec2-CTC method and ATP-LLM (R = 32) whose F1 scores are 80.46% and 81.61%, respectively.

The SOTA performance confirms the effectiveness of our proposed PP-ATP model.

E. Ablation study

Table III investigates the influence of the down-sampling rate on system performance. The larger the down-sampling rate is, the larger the audio prompt stride. The original stride of speech representations taken as audio prompt is 20ms. When enlarging the stride of the audio prompt to 40ms, the degradation of system performance is not significant. For example, the F1s of PP-ATP (20ms) and PP-ATP (40ms) are 81.91% and 81.73%, respectively. The PP-ATP (40ms) still outperforms both baselines. When further increase the audio prompt stride to 100ms, the system suffers significant performance degradation. One possible reason is that Qwen-0.5B is a relatively small scale LLM.

IV. CONCLUSIONS

In this paper, we propose a PP-ATP model, which integrates potential pronunciations into the text prompting process, to enhance the mispronunciation detection and diagnosis (MDD) ability of LLMs in second language (L2) English. The proposed method achieves new state-of-the-art (SOTA) performance in the MDD task on the CU-CHLOE corpus, outperforming traditional SOTA method (wav2vec2-CTC) and LLM-based SOTA method (ATP-LLM). Furthermore, our proposed PP-ATP model with a small LLM decoder (Qwen2-0.5B) is relatively robust to the change of down-sampling. In the future, we will persist in finding enhanced methods of integrating linguistic information into LLMs to further improve the performance of MDD task.

REFERENCES

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

[2] H. Huang, H. Xu, Y. Hu, and G. Zhou, "A transfer learning approach to goodness of pronunciation based automatic mispronunciation detection," *The Journal of the Acoustical Society of America*, vol. 142, no. 5, pp. 3165–3177, 2017.

[3] 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 *INTERSPEECH*, 2019.

[4] 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.

[5] Z. Zhang, Y. Wang, and J. Yang, "End-to-end Mispronunciation Detection with Simulated Error Distance," in *Proc. Interspeech 2022*, 2022, pp. 4327–4331.

[6] X. Qian, H. M. Meng, and F. K. Soong, "The use of dbn-hmms for mispronunciation detection and diagnosis in l2 english to support computer-aided pronunciation training," in *INTERSPEECH*, 2012.

[7] Y.-B. Wang and L.-s. Lee, "Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 564–579, 2015.

[8] W.-K. Leung, X. Liu, and H. M. Meng, "Cnn-rnn-ctc based end-to-end mispronunciation detection and diagnosis," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8132–8136, 2019.

[9] Y. Feng, G. Fu, Q. Chen, and K. Chen, "Sed-mdd: Towards sentence dependent end-to-end mispronunciation detection and diagnosis," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3492–3496, 2020.

[10] M. Wu, K. Li, W.-K. Leung, and H. Meng, "Transformer Based End-to-End Mispronunciation Detection and Diagnosis," in *Proc. Interspeech 2021*, 2021, pp. 3954–3958.

[11] L. Peng, K. Fu, B. Lin, D. Ke, and J. Zhan, "A Study on Fine-Tuning wav2vec2.0 Model for the Task of Mispronunciation Detection and Diagnosis," in *Proc. Interspeech 2021*, 2021, pp. 4448–4452.

[12] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[13] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 369–376. [Online]. Available: https://doi.org/10.1145/1143844.1143891

[14] Y. E. Kheir, S. A. Chowdhury, and A. Ali, "L1-aware multilingual mispronunciation detection framework," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12 752–12 756.

[15] Y. Wan, Y. Shi, B. Lin, and Y. Xie, "A study of mispronunciation detection and diagnosis based on meta-learning," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12 792–12 796.

[16] OpenAI, "Gpt-4 technical report," 2024.

[17] G. Team, "Gemini: A family of highly capable multimodal models," 2023.

[18] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, and Z. Fan, "Qwen2 technical report," *arXiv preprint arXiv:2407.10671*, 2024.

[19] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, and et al, "Palm: Scaling language modeling with pathways," 2022.

[20] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence, "Palm-e: An embodied multimodal language model," 2023.

[21] Y. Su, T. Lan, H. Li, J. Xu, Y. Wang, and D. Cai, "Pandagpt: One model to instruction-follow them all," 2023.

[22] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," 2023.

[23] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," March 2023. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/

[24] R. Huang, M. Li, D. Yang, J. Shi, X. Chang, Z. Ye, Y. Wu, Z. Hong, J. Huang, J. Liu, Y. Ren, Z. Zhao, and S. Watanabe, "Audiogpt: Understanding and generating speech, music, sound, and talking head," 2023.

[25] Y. Fathullah, C. Wu, E. Lakomkin, J. Jia, Y. Shangguan, K. Li, J. Guo, W. Xiong, J. Mahadeokar, O. Kalinli, C. Fuegen, and M. Seltzer, "Prompting large language models with speech recognition abilities," 2023.

[26] Z. Ma, G. Yang, Y. Yang, Z. Gao, J. Wang, Z. Du, F. Yu, Q. Chen, S. Zheng, S. Zhang, and X. Chen, "An embarrassingly simple approach for llm with strong asr capacity," 2024. [Online]. Available: https://arxiv.org/abs/2402.08846

[27] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, C. Zhou, and J. Zhou, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[28] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, "Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities," 2023.

[29] 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 Trans. Audio, Speech and Lang. Proc.*, vol. 29, p. 3451–3460, oct 2021.

[30] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

[31] M. Wu, J. Xu, X. Wu, and H. Meng, "Prompting large language models with mispronunciation detection and diagnosis abilities," in *Interspeech 2024*, 2024, pp. 2990–2994.

[32] H. M. Meng, W. K. Lo, A. M. Harrison, P. Lee, K. H. WONG, W.-K. Leung, and F. Meng, "Development of automatic speech recognition and synthesis technologies to support chinese learners of english : The cuhk experience development of automatic speech recognition and synthesis technologies to support chinese learners of english : The cuhk experience," 2010.

[33] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[34] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, and et al, "Huggingface's transformers: State-of-the-art natural language processing," 2020.

[35] X. Qian, H. Meng, and F. Soong, "Capturing l2 segmental mispronunciations with joint-sequence models in computer-aided pronunciation training (capt)," in *2010 7th International Symposium on Chinese Spoken Language Processing*, 2010, pp. 84–88.