Interspeech 2024
1-5 September 2024, Kos, Greece

Leveraging Large Language Models to Refine Automatic Feedback Generation at Articulatory Level in Computer Aided Pronunciation Training

Huihang Zhong[1], *Yanlu Xie*[1], *ZiJin Yao*[1]

[1]Beijing Language and Culture University, China

zhhsait@163.com, xieyanlu@blcu.edu.cn,20181051@blcu.edu.cn

Abstract

This study explores the potential of leveraging Large Language Models (LLMs) to refine automatic feedback generation in Computer-Aided Pronunciation Training (CAPT). Specifically, it evaluates the impact of two factors on the effectiveness of automatically generated pronunciation feedbacks: (1) the use of mispronunciation detection at different fine-grained levels as prompts for GPT-4 models to generate automatic feedback, and (2) the fine-tuning of GPT-4 models using specific prompt-feedback pairs aimed at optimizing feedback generation. Feedback generated through each approach is rated by second language (L2) learners in terms of comprehensibility and helpfulness. The results highlight both the potential of using LLMs for automatic feedback generation and the effectiveness of articulatory level representations. Our accessible demonstrations invite further exploration.[1]

Index Terms: mispronunciation feedback, computer-aided pronunciation training(CAPT), mispronunciation detection and diagnosis(MD&D), articulatory features

1. Introduction

Computer-Assisted Pronunciation Training (CAPT) represents a pivotal advancement in language learning, offering second language (L2) learners an unprecedented opportunity to independently hone their language skills. This method, serving as an invaluable complement to traditional classroom instruction, provides personalized pronunciation feedback, a critical aspect for mastering foreign languages. Central to CAPT's effectiveness is the Mispronunciation Detection and Diagnosis (MD&D) module, tasked with identifying and correcting pronunciation errors, thereby facilitating a more nuanced learning experience.

Historically, research in this domain, as underscored by studies such as [1, 2, 3, 4], has primarily been centered on mispronunciation detection. While these endeavors have established a robust foundation, advancements in mispronunciation diagnosis and feedback generation have not progressed at the same pace, presenting a critical challenge in the evolution of Mispronunciation Detection and Diagnosis (MD&D). Current metrics like the 'Goodness of Pronunciation' (GOP) [5] provide essential insights yet often lack the granularity and specificity required for effective second language (L2) learning. The GOP metric, for example, offers a broad evaluation but misses the detailed feedback necessary for learners to rectify specific pronunciation errors. On the other hand, the Extended Recognition Network (ERN) [6], while useful, its main limitation lies in its capability to only provide feedback on fixed read-aloud texts. The current technological landscape lacks a solution capable of delivering feedback on any arbitrary read-aloud text, which limits the scalability and adaptability of the feedback system.

In addressing the challenges of automating mispronunciation feedback, the utilization of large language models presents a promising approach. Currently, adapting these models to specific domains primarily involves methods such as fine-tuning with domain-specific datasets [7, 8, 9, 10], or employing retrieval-based systems designed for particular educational contexts [11, 12, 13]. However, the absence of extensive, open-source datasets specifically for mispronunciation feedback poses a significant limitation to the application of these methods in the realm of mispronunciation feedback. This lack of specialized data resources restricts the effectiveness of both fine-tuning and retrieval-based approaches. Consequently, we propose the integration of external information sources as an innovative solution to enhance the efficacy of feedback. This approach aims to circumvent the dataset limitations by enriching the feedback mechanism with additional, relevant data, thereby making mispronunciation feedback more precise and beneficial for learners.

Informed by the findings of Li et al. [14], which highlight the auxiliary role of articulatory features in enhancing mispronunciation detection systems, our research endeavors to incorporate these features into mispronunciation feedback mechanisms within CAPT systems. Recognizing the potential of articulatory features to provide detailed insights into pronunciation errors, we propose three distinct approaches to leverage these features for more effective feedback: (1) Viterbi Based Mispronunciation Feedback: This approach utilizes a mispronunciation detection model to identify phonemes, employs the Viterbi algorithm to pinpoint errors, analyzes these errors through articulatory features, and then synthesizes the feedback using GPT-4 for clarity and depth; (2)Prompt-based Feedback Generation: we employ prompts to guide GPT-4 in identifying error locations from phoneme inputs and use articulatory features to analyze these errors, culminating in detailed feedback; and (3) Fine-tuning Based Mispronunciation Feedback: Initially generating datasets with GPT-4, we then fine-tune other language models to achieve a feedback generation process similar to the prompt-based approach. Additionally, we introduce the Splitting Process Datasets (SPD) technique, significantly enhancing feedback effectiveness.

These methodologies, though diverse in application, share a common goal: to utilize articulatory features and LLMs in providing nuanced feedback on mispronunciations.

2. METHOD

In this section, we explore the system architecture and the methodologies employed to derive mispronunciation feedback.

[1]https://github.com/lunar333/mispronunciation-feedback

10.21437/Interspeech.2024-1005

We introduce three distinct approaches to achieving mispronunciation feedback.

2.1. Articulatory Features

By mapping various phonemes to articulatory features using an articulatory features mapping table[15], we can generate detailed feedback on articulatory mispronunciations. This process is based on understanding the significance of each dimension represented in the articulatory features, as outlined in Table 1. We derive specific articulatory mispronunciations by comparing the dimensions of these features.

For instance, consider the phonemes represented by the letters 's' and 'z'. The articulatory features of 's' are [1,2,2,3,3,3,0,0], while those of 'z' are [1,2,2,3,3,3,0,1]. A comparative analysis of these features reveals a difference only in the eighth dimension, which indicates whether the vocal cords vibrate. This kind of comparison allows us to identify precise articulatory mispronunciations.

As another example, if the phoneme 'n' is incorrectly pronounced as 'ng', we observe the following: the articulatory features for 'n' are [1,1,2,2,3,4,1,1], and for 'ng', they are [1,2,2,0,3,1,1,1]. This comparison indicates that to correct the mispronunciation, one needs to retract the tongue.

Table 1: *The articulatory feature space*

Stream	Classes	Cardinality
jaw	0:Nearly Closed, 1:Neutral, 2:Slightly Lowered, 3:Lowered	4
lip separation	0:Closed, 1: Slightly Apart, 2:Apart, 3:Wide Apart	4
lip rounding	0:Rounded, 1:Slightly Rounded, 2:Neutral, 3:Spread	4
tongue frontness	0:Back, 1:Slightly Back, 2:Neutral, 3:Slightly Front, 4:Front	5
tongue height	0:Low, 1:Mid, 2:Mid-High, 3:High	4
tongue tip	0:Low, 1:Neutral, 2:Dental, 3:Nearly Alveolar, 4:Alveolar	5
velum	0:Closed, 1:Open	2
voicing	0:Unvoiced, 1:Voiced	2

The Viterbi algorithm is an essential component in our approach to detecting mispronounced phonemes. It is a dynamic programming algorithm used in Hidden Markov Models (HMMs) to determine the most probable sequence of hidden states, which, in our case, correspond to phonemes in spoken language.

Problem Formulation: Given an observed sequence of phonemes $O = \{o_1, o_2, \ldots, o_T\}$, we define a set of states $S = \{s_1, s_2, \ldots, s_N\}$ representing possible phonemes. Our objective is to find the most probable state sequence $Q = \{q_1, q_2, \ldots, q_T\}$ that best explains the observations.

Probability Calculations: The Viterbi algorithm computes the probability of the most likely state sequence that produces the observed phonemes. The probability of a state sequence Q, given the observation sequence O, is calculated as:

$$P(Q|O) = \prod_{t=1}^{T} P(q_t|q_{t-1}) \times P(o_t|q_t) \quad (1)$$

where $P(q_t|q_{t-1})$ is the transition probability from state q_{t-1} to q_t, and $P(o_t|q_t)$ is the likelihood of observing o_t given state q_t.

Dynamic Programming Implementation: The algorithm utilizes a dynamic programming approach, involving the initialization and iterative update of a matrix of probabilities. For each state s_i at time t, it computes:

$$V_t(s_i) = \max_{q_{t-1} \in S} [P(o_t|s_i) \times P(s_i|q_{t-1}) \times V_{t-1}(q_{t-1})] \quad (2)$$

where $V_t(s_i)$ represents the probability of the most probable state sequence ending in state s_i at time t.

Backtracking for Optimal Sequence: After computing probabilities, the algorithm backtracks from the last state to determine the most probable path, identifying the sequence of phonemes.

Application to Mispronunciation Detection: This algorithm is applied to compare the predicted phoneme sequence with standard pronunciation. Discrepancies between these sequences are identified as mispronunciations, which are vital for generating accurate pronunciation feedback.

The Viterbi algorithm's role in our methodology is crucial, enabling precise identification of phonetic errors.

2.2. Viterbi Based Mispronunciation Feedback

Figure 2 outlines the architecture of Viterbi Based Mispronunciation Feedback. This method commences with a mispronunciation detection model responsible for identifying canonical phonemes within spoken input. Following this step, we employ the Viterbi algorithm to accurately locate mispronounced phonemes. These phonemes are subsequently transformed into their articulatory features. The essence of generating feedback in this method involves comparing these articulatory features as described in section 2.1. The initial feedback produced is in a textual format, which is then refined and summarized using GPT-4 to significantly improve its accessibility and user engagement.

2.3. Prompt Based Mispronunciation Feedback

This approach leverages the intrinsic capabilities of prompt-based techniques, widely adopted across various applications, tailored to the intricate demands of mispronunciation feedback. Unlike simple prompts, this task necessitates an in-depth understanding and analysis of spoken language to accurately address pronunciation errors.

The effective execution of this feedback mechanism entails a series of critical steps:

1. **Segmentation of Phonemes**: This step begins with the segmentation of both correct phonemes and canonical phonemes. Here, *correct phonemes* refer to the standard phonetic representations associated with the spoken text as per the target language's pronunciation norms. In contrast, *canonical phonemes* denote the actual phonetic outputs produced by second language (L2) learners when reading the text. The process involves breaking down the spoken text into individual words and identifying both the correct and canonical phonemes for each, thereby setting the stage for precise mispronunciation detection and feedback.

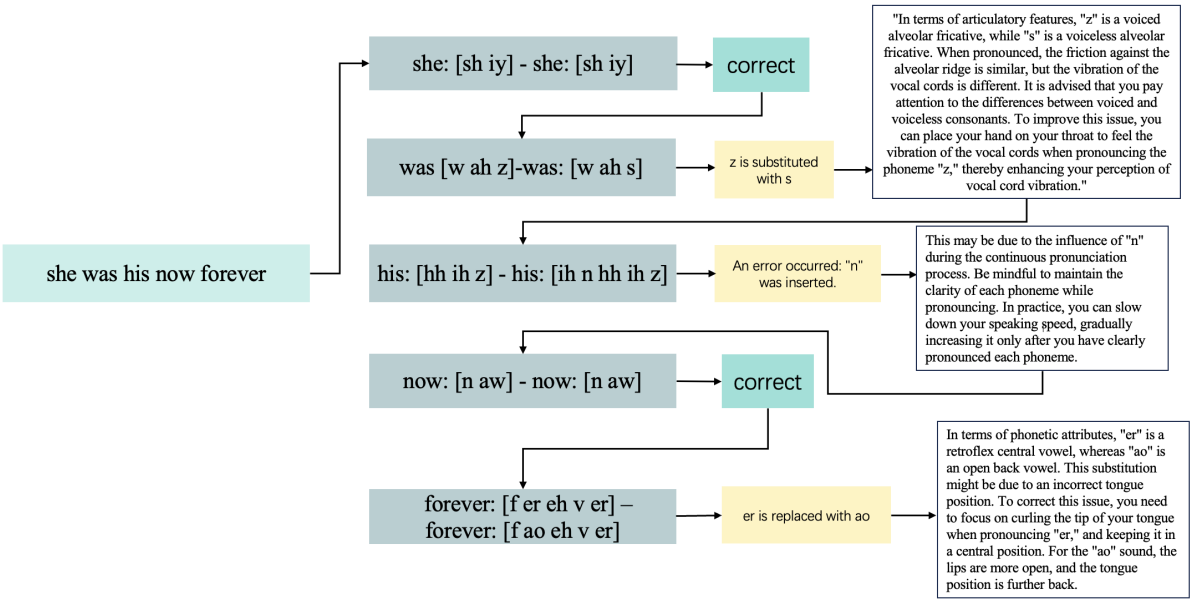

Figure 1: *GPT4 output process*

2. **Identification of Error Phonemes**: Post-segmentation, it involves contrasting the canonical phonemes with the actual ones for each word, to identify mispronounced phonemes, crucial for pinpointing exact error locations within the spoken text.
3. **Conversion to Articulatory Features**: With the error phonemes identified, they are mapped to corresponding articulatory features. This analysis, as elaborated in Section 2.2, allows for generating specific feedback, illuminating pronunciation discrepancies and suggesting improvements.

By using these steps as a foundational blueprint, we formulate prompts and examples that encapsulate the entire process. This method, **Prompt Based Mispronunciation Feedback**, enables GPT-4 to provide direct mispronunciation feedback from the given inputs, circumventing the need for intermediary steps characteristic of the Viterbi algorithm. As a result, this method markedly diminishes feedback provision latency, offering a more efficient alternative to **Viterbi Based Mispronunciation Feedback**.

2.4. Fine-tuning Based Mispronunciation Feedback

Considering that some regions may not be able to use GPT4, we have also explored fine-tuning based mispronunciation feedback. For convenience, we have directly used chatglm6b in this paper, without attempting more large language models. Our paper is not intended to explore the issue of fine-tuning large language models, but to explore how to better utilize articulatory feature to achieve mispronunciation feedback.

For large language models with parameters such as chatglm6b, using prompt directly may result in poor performance due to the limitations of the model's capabilities. Therefore, using fine-tuning would be a wise approach. The intuitive method is to use the Prompt Based Mispronunciation Feedback method in Section 2.4 to generate dataset for fine-tuning. However, our experimental results show that this method will have a relatively poor effect. We believe that this may be due to the complexity of the mispronunciation feedback step, which the model cannot achieve perfectly, so we propose the Splitting Process Datasets(SPD). The core idea of SPD is to split the dataset into multiple sub datasets, guiding the big language model to output results according to the steps.

The SPD dataset consists of two sub datasets named process datasets, we call them the problem dataset and the answer dataset respectively. The problem dataset is designed to help the model learn how to identify incorrect phonemes in the transcription text. The answer dataset is designed to help the model learn how to obtain mispronunciation feedback based on the identified incorrect phonemes. The problem dataset acquisition process is as follows: 1.Segmenting the sentence of transcription phonemes and canonical phonemes to obtain the transcription phonemes and canonical phonemes for each word. 2.Confirm words with incorrect phonemes by comparing each word's transcription phonemes and canonical phonemes. The answer dataset acquisition process is as follows: Using the problem dataset as input, require the model to output articulatory mispronunciation feedback using articulatory features mapping table. Because the later datasets are generated based on the previous dataset, we call them the problem dataset and the answer dataset, respectively. After fine-tuning the SPD dataset, we require the model to first form a problem dataset during inference and use it as input, requiring the model to output the final feedback result. This is equivalent to artificially guiding the model to generate feedback results according to the specified steps. The experiment shows that using this method will have better results than directly fine-tuning.

3. Experiment Setup

In this study, we utilized the L2-ARCTIC dataset [16], a specialized corpus of non-native English speech designed primarily for research in voice conversion, accent conversion, and mispronunciation detection. The dataset includes recordings

from twenty-four non-native English speakers with diverse phonetic backgrounds, (12 males and 12 females) non-native speakers whose L1 languages include Hindi, Korean, Spanish, Arabic, Vietnamese, and Chinese. Following prior works [17], six speakers (NJS, TLV, TNI, TXHC, YKWK, ZHAA) were selected as the test set while the rest were merged to build the training set.

The training set is used as the source for generating fine-tuning datasets, while the test set is utilized to evaluate the final performance. The specific details of the dataset generation process for fine-tuning can be found in the GitHub repository.[2] We used the gpt-4-1106-vision-preview version of GPT-4, as of April 2023, with a temperature setting of 0.

For Viterbi-based mispronunciation feedback, abbreviated as **GPT-4+Viterbi+Af** we employ two baseline methodologies to underscore the significance of both the Viterbi algorithm and articulatory features in our analysis. These baselines include the prompt-based method, abbreviated as **GPT-4**, and an integration of the Viterbi algorithm without articulatory feature analysis, denoted as **GPT-4+Viterbi**. These comparative baselines serve to highlight the distinct contributions of the Viterbi algorithm and articulatory features within our framework.

For fine-tuning based mispronunciation feedback, We employ the prompt-based method to generate datasets for direct fine-tuning, abbreviated as **GPT-4 Dataset Fine-tuning**, and utilize the SPD (Splitting Process Datasets) method for dataset generation aimed at fine-tuning, denoted as **SPD Dataset Fine-tuning**. Through comparative analysis, we aim to evaluate their respective effectiveness, thereby demonstrating the superior efficacy of the SPD method in enhancing fine-tuning outcomes.

4. RESULTS

In assessing the effectiveness of the feedback texts produced by our system, we engaged 10 second language (L2) learners, each bringing diverse experiences to the evaluation process. This diverse participant pool enhances the reliability and depth of our feedback assessment. The evaluators were instructed to rate each piece of generated feedback from two perspectives: comprehensibility (ease of understanding) and helpfulness (aid in improving pronunciation). Ratings were given on a scale from 1 ('very poor') to 5 ('excellent') for both criteria.

To ensure a thorough and unbiased evaluation, different pieces of generated feedback were assigned to various evaluators. We then calculated the Mean Opinion Score (MOS) for both comprehensibility and helpfulness, along with a 95% confidence interval for each set of evaluations, providing a nuanced insight into the effectiveness of our feedback system from the viewpoints of L2 learners.

GPT-4 served as a baseline, demonstrating the inherent capabilities of large language models in generating feedback directly. The scores for GPT-4 indicate a moderate level of effectiveness, with MOS of 3.21 for comprehensibility and 3.04 for helpfulness, suggesting that while GPT-4 can generate relevant feedback, there is room for improvement in terms of specificity and utility for pronunciation correction. The integration of the **Viterbi algorithm (GPT-4+Viterbi)** improved the feedback's stability and relevance, as evidenced by higher MOS scores of 3.51 for comprehensibility and 3.32 for helpfulness. This improvement underscores the Viterbi algorithm's role in enhancing the logical reasoning capabilities of GPT-4, providing a more stable feedback mechanism. Further incorporating **articulatory features (GPT-4+Viterbi+Af)** significantly increased the helpfulness score to 3.83, the highest among the methods tested, while slightly reducing comprehensibility to 3.45. This indicates that articulatory feature integration aids L2 learners in correcting mispronunciations more effectively, though it may introduce complexity that challenges learners with lower proficiency levels.

The fine-tuning approaches, **GPT-4 Dataset Fine-tuning** and **SPD Dataset Fine-tuning**, yielded lower effectiveness compared to the Viterbi-based methods. However, SPD Dataset Fine-tuning showed an improvement over GPT-4 Dataset Fine-tuning, with scores of 2.91 for comprehensibility and 2.71 for helpfulness. This highlights the SPD method's effectiveness in enhancing fine-tuning outcomes by segmenting the dataset into multiple sub-datasets, thereby guiding the model to output results stepwise.

In summary, the **GPT-4+Viterbi+Af** method emerged as the most effective in aiding L2 learners, demonstrating the value of combining the Viterbi algorithm with articulatory feature analysis. The results affirm the potential of tailored feedback methods in significantly improving L2 pronunciation training, with the SPD technique showing promise in refining fine-tuning processes for language models.

Table 2: *Mean Opinion Scores (MOS) of Feedback Methods*

Method	Comprehensibility	Helpfulness
GPT-4	3.21	3.0
GPT-4+Viterbi	**3.5**	3.3
GPT-4+Viterbi+Af	3.46	**3.79**
GPT-4 Dataset Fine-tuning	2.68	2.53
SPD Dataset Fine-tuning	3.0	2.7

5. CONCLUSION

This study has successfully demonstrated a significant enhancement in the effectiveness of mispronunciation feedback within Computer-Assisted Pronunciation Training (CAPT) systems through the integration of articulatory features and LLMs. Our findings reveal that incorporating articulatory features leads to a more nuanced and accurate analysis of pronunciation errors, ultimately enabling the generation of more precise and beneficial feedback for language learners.

A pivotal contribution of our research is the implementation of the Splitting Process Datasets (SPD) technique. The SPD method has proven instrumental in further refining the feedback provided by CAPT systems. By segmenting datasets into multiple sub-datasets, SPD allows for a more granular approach to feedback generation.

While our study has primarily relied on human evaluators for the assessment of feedback quality, future research could explore the development of automated evaluation metrics. Such metrics have the potential to streamline the assessment process, offering more consistent and objective measures of feedback effectiveness in CAPT systems. The adoption of automated metrics could also facilitate the scaling of the evaluation process, enabling more extensive testing and refinement of feedback mechanisms.

6. Acknowledgements

This research project is supported by Humanities and Social Sciences Research Planning Fund of the Ministry of Education

[2]https://github.com/lunar333/mis-feedback.git

(23YJA740012), Key Research Project on International Chinese Language Education (22YH49B), Key Projects of the National Language Commission (ZDI145-101). The corresponding author of the paper is Yanlu Xie.

7. References

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

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

[3] 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," *arXiv preprint arXiv:2104.08428*, 2021.

[4] D. Korzekwa, J. Lorenzo-Trueba, S. Zaporowski, S. Calamaro, T. Drugman, and B. Kostek, "Mispronunciation detection in non-native (l2) english with uncertainty modeling," in *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2021, pp. 7738–7742.

[5] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[6] A. M. Harrison, W.-K. Lo, X.-j. Qian, and H. Meng, "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training," in *International Workshop on Speech and Language Technology in Education*, 2009.

[7] 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," *arXiv preprint arXiv:2106.09685*, 2021.

[8] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, "Gpt understands, too," *arXiv preprint arXiv:2103.10385*, 2021.

[9] X. Liu, K. Ji, Y. Fu, W. L. Tam, Z. Du, Z. Yang, and J. Tang, "P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks," *arXiv preprint arXiv:2110.07602*, 2021.

[10] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," *arXiv preprint arXiv:2101.00190*, 2021.

[11] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark *et al.*, "Improving language models by retrieving from trillions of tokens," in *International conference on machine learning*. PMLR, 2022, pp. 2206–2240.

[12] D. Cai, Y. Wang, L. Liu, and S. Shi, "Recent advances in retrieval-augmented text generation," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 3417–3419.

[13] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[14] 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 *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 6135–6139.

[15] J. Tepperman and S. Narayanan, "Using articulatory representations to detect segmental errors in nonnative pronunciation," *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 8–22, 2007.

[16] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-arctic: A non-native english speech corpus." in *Interspeech*, 2018, pp. 2783–2787.

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