

MISPRONUNCIATION DETECTION AND DIAGNOSIS WITHOUT MODEL TRAINING: A RETRIEVAL-BASED APPROACH

Huu Tuong Tu^{1,2} Ha Viet Khanh¹ Tran Tien Dat¹ Vu Huan³ Thien Van Luong³
 Nguyen Tien Cuong² Nguyen Thi Thu Trang^{1*}

¹Hanoi University of Science and Technology ²VNPT AI, VNPT Group
³National Economics University

ABSTRACT

Mispronunciation Detection and Diagnosis (MDD) is crucial for language learning and speech therapy. Unlike conventional methods that require scoring models or training phoneme-level models, we propose a novel training-free framework that leverages retrieval techniques with a pre-trained Automatic Speech Recognition model. Our method avoids phoneme-specific modeling or additional task-specific training, while still achieving accurate detection and diagnosis of pronunciation errors. Experiments on the L2-ARCTIC dataset show that our method achieves a superior F1 score of 69.60% while avoiding the complexity of model training.

Index Terms— Mispronunciation detection and diagnosis, retrieval-based methods, training-free framework, automatic pronunciation assessment

1. INTRODUCTION

Mispronunciation Detection and Diagnosis is a fundamental task in Computer-Assisted Pronunciation Training (CAPT). Given a reference text and a learner’s spoken utterance, an MDD system must determine whether the pronunciation is correct, localize mispronounced units, and provide diagnostic feedback. The earliest systems were built on the Goodness-of-Pronunciation (GOP), which use acoustic models to compute phoneme scores and use a threshold to detect mispronunciation [1]. While GOP offers a simple and efficient approach for scoring pronunciation accuracy, it lacks detailed diagnostic feedback for learners.

With the rise of end-to-end Automatic Speech Recognition (ASR), MDD research has shifted toward phoneme recognition models. The phoneme sequences recognized from the learner’s audio by these models are aligned with the canonical phoneme sequence from the reference text to perform detection and diagnosis. One notable ASR-based model is CNN-RNN-CTC architecture [2]. This model processes input audio through a deep learning framework trained with the Connectionist Temporal Classification (CTC) [3]

loss function, directly mapping raw acoustic features to pronounced phoneme sequences. While it achieves superior results over former methods, this approach does not use prior text information that learners were expected to read aloud.

To address this limitation, [4] proposed an end-to-end architecture for sentence-dependent MDD, combining acoustic features with a predefined canonical text sequence via cross-attention [5]. However, mismatches between the reference text and predicted phoneme sequence often cause inconsistencies. To overcome this issue, Fu et al. [6] replaced the reference text with canonical phoneme sequences and applied data augmentation to improve input diversity and robustness.

Pretrained self-supervised speech models and ASR models have become widely used in MDD systems [7, 8]. By leveraging large-scale unlabeled or transcribed speech corpora, these models can encode raw audio into robust phonetic embeddings through transfer learning, providing a rich representation of speech. Such embeddings significantly improve the ability of MDD systems. In parallel, specially designed linguistic encoders, such as text-aware and graph-based methods, have been explored to enhance the linguistic branch [9, 10]. These approaches capture structured dependencies among canonical phonemes, phonological rules, and articulatory attributes, enabling MDD systems to incorporate prior linguistic knowledge more effectively.

More recently, multitask approaches have gained considerable attention in MDD. Instead of relying solely on the phoneme recognition, [11] proposed a multilingual MDD framework that leverages the speaker’s native language (L1) and target language (L2) information. This design explicitly models cross-lingual phonological disparities, thereby improving detection robustness. In addition, several studies have shown the benefits of joint training for MDD and Automatic Pronunciation Assessment, demonstrating that the two tasks are highly correlated and that their integration leads to more reliable error detection and diagnosis [12]. Other works have explored multi-view speech representation approaches, where multiple pretrained audio encoders are fused to strengthen phonetic representations, achieving improved generalization across diverse groups of L2 learners [13].

*Corresponding author.

In this work, we find that modern ASR systems already encode sufficient linguistic information to support mispronunciation detection. Therefore, inspired by retrieval-based methods, we discover that MDD can be performed without phoneme-specific modeling or additional task-specific training, while still providing the necessary detail for accurate detection and diagnosis of pronunciation errors. Particularly, our contributions are summarized as follows:

1. We propose a training-free MDD framework that eliminates the need for phoneme-level modeling.
2. We introduce the first retrieval-based strategy for MDD, leveraging pretrained ASR models in a RAG-inspired manner.
3. We show through experiments on public L2 English datasets that our approach achieves competitive detection accuracy with minimal complexity.

2. PROPOSED METHOD

Retrieval-based strategies have recently demonstrated strong effectiveness across a variety of domains, including enhancing large language models [14] and speech processing [15, 16]. In this paper, we propose a retrieval-based pipeline for MDD, which is named Phoneme Embedding Retrieval MDD (PER-MDD). The overall pipeline illustrates in figure 1.

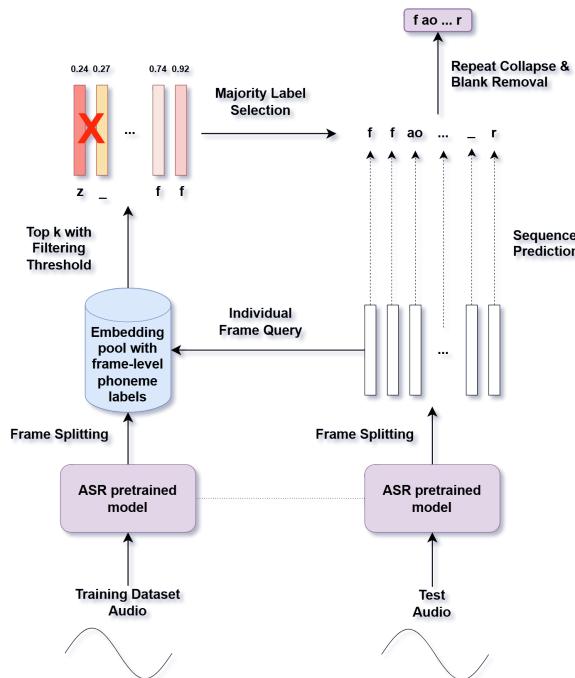


Fig. 1: Illustration of our proposed PER-MDD method.

2.1. Phoneme embedding pool construction

Let $\mathcal{D}_{\text{train}}$ be a labeled training dataset where each utterance has phoneme-level time alignments. Each utterance is segmented into frames $\{x_t\}_{t=1}^T$, and a pretrained ASR model $f(\cdot)$ maps each frame to an embedding:

$$e_t = f(x_t), \quad e_t \in R^d. \quad (1)$$

Since each frame corresponds to a specific time in the audio, and the dataset provides phoneme labels with start and end times, we can assign each frame a phoneme label $y_t \in \mathcal{V} \cup \text{blank}$, where \mathcal{V} is the phoneme vocabulary.

The phoneme embedding pool is then defined as:

$$\mathcal{P} = (e_t, y_t)_{t=1}^N, \quad (2)$$

with N total frame-level pairs.

To construct the pool, we can select embeddings in different ways for each phoneme span (from start time s to end time e of embedding audio with phoneme labels):

- **All-frame:** include every frame embedding e_t .
- **Middle-frame:** include only the embedding of the middle frame of the span, $e_{\lfloor(s+e)/2\rfloor}$.
- **Mean-frame:** compute the average embedding over the span $\bar{e} = \frac{1}{e-s} \sum_{t=s}^e e_t$.

This pool \mathcal{P} serves as the reference for retrieval-based phoneme prediction during inference.

2.2. Inference procedure

For a test utterance, we extract frame-level embeddings $q_t = f(x_t^{\text{test}})$. Each query q_t is compared against the pool \mathcal{P} using cosine similarity:

$$s(q_t, e) = \frac{q_t \cdot e}{\|q_t\| \|e\|}. \quad (3)$$

Candidate Retrieval. The top- k nearest neighbors are selected:

$$\mathcal{N}_k(q_t) = \text{Top-}k(\{(e, y) \in \mathcal{P} \mid s(q_t, e)\}). \quad (4)$$

Then, a filtering threshold τ is applied:

$$\mathcal{N}^*(q_t) = \{y \mid (e, y) \in \mathcal{N}_k(q_t), s(q_t, e) \geq \tau\}. \quad (5)$$

Label Assignment. The predicted label at frame level is:

$$\hat{y}_t = \begin{cases} \text{blank}, & \text{if } \mathcal{N}^*(q_t) = \emptyset, \\ \text{mode}(\mathcal{N}^*(q_t)), & \text{otherwise.} \end{cases} \quad (6)$$

Post-processing. The sequence $\{\hat{y}_t\}$ is refined by: (i) collapsing consecutive duplicates, and (ii) removing blanks. This produces a final phoneme sequence \hat{Y} that represents the predicted pronunciation, which is aligned with the canonical phoneme sequence to detect and diagnose mispronunciations.

Table 1: Performance comparison between our proposed model and its baselines

Model	MDD Metric							ASR metric	
	FRR↓	FAR↓	DER↓	PRE↑	REC↑	F1↑	DA↑	PER↓	COR↑
PHN-M2 [17]	6.33	45.37	25.12	64.51	54.63	59.16	86.88	17.12	-
L1-MultiMDD [11]	4.60	-	-	-	-	57.40	-	12.55	-
w2v2-XLSR [8]	5.70	41.80	29.28	62.86	58.20	60.44	-	16.20	-
Joint-Align [18]	-	-	-	77.12	53.31	63.04	-	-	-
MDDGCN [9]	9.18	38.03	25.24	51.90	61.97	56.49	-	-	-
MVmuli-MTseq [13]	-	-	-	61.43	59.23	60.31	-	14.13	-
PER-MDD (Ours)	4.43	32.44	37.77	71.78	67.56	69.60	91.57	104.08	90.42

↓ lower is better ↑ higher is better

3. EXPERIMENTS

3.1. Datasets

We evaluate our model’s performance using the publicly available L2-ARCTIC dataset [19], which contains non-native English speech from speakers of various native languages, including Hindi, Korean, Mandarin, Spanish, Arabic, and Vietnamese (24 speakers in total). L2-ARCTIC is specifically designed for CAPT tasks and provides canonical phoneme sequences from the reference text, corresponding audio recordings, and the actual phonemes produced by each speaker. Following prior work [4, 6, 18], we use six speakers (“NJS”, “TLV”, “TNI”, “TXHC”, “YKWK”, “ZHA”) to construct the test set, and 12 speakers to build the training set.

3.2. Experimental setup

We use the publicly available HuBERT model [20], fine-tuned on a large-scale ASR dataset¹. We sample 500 training audio files to build the phoneme embedding pool, representing each span with its middle frame (known as mid-frame pooling). The similarity threshold is 0.7, and the retrieval top- k is 10.

Table 2: MDD evaluation example

Canonical phoneme	ae	d	v	ay	s
Human-annotated phoneme	ae	t	v	ey	sh
Predicted phoneme	ae	d	f	ey	z
Evaluation result	TA	FA	FR	TR CD	TR DE

3.3. Evaluation metrics

Consistent with previous studies [2, 4, 6, 7], we evaluate our models using both ASR and MDD performance metrics. For ASR, we use the phone error rate (PER) and Correctness (COR) to assess performance. The MDD evaluation process

involves categorizing the model’s predictions into distinct groups: true acceptance (TA), true rejection (TR), false acceptance (FA), and false rejection (FR). Within TR, we further divide results into correct diagnosis (CD) and error diagnosis (DE). Metrics such as detection accuracy (DA), diagnosis error rate (DER), recall (REC), precision (PRE), and F1 score are computed to assess the model’s performance. An illustrative example of the MDD evaluation is presented in Table 2. A green phoneme is one that is the same as the canonical phoneme, while a red phoneme is one that is different. For more details on how all the metrics are computed, refer to [7].

3.4. Performance analysis

3.4.1. Performance comparison with baselines

In MDD systems, misjudging a large number of correctly pronounced phones as mispronunciations can frustrate learners and negatively impact their learning experience. Therefore, FRR is regarded as the most critical metric in MDD. As shown in Table 1, our proposed model achieves superior performance with an FRR of only 4.43%. Moreover, our model also delivers significant improvements in F1, FAR, and recall, providing significant performance gains of around 6%, compared to state-of-the-art (SOTA) baselines such as MD-DGCN [9] and Joint-Align [18]. Note from this table that since some baseline models are not open-source, we report the official results released by the authors. The parameters of our scheme shown in Table 1 are provided in Subsection 3.2.

It is worth noting from Table 1 that our retrieval-based MDD method tends to produce a relatively large number of insertion errors compared to human-annotated transcripts, leading to considerably higher PER than baselines. However, these insertion errors do not significantly affect MDD performance. As described in the metrics of [2], MDD systems primarily evaluate whether each phoneme in the canonical phoneme sequence is pronounced correctly, focusing on diagnosing errors such as whether a phoneme is pronounced as another, rather than detecting insertions.

To further illustrate this, the Correctness, computed similarly to $(1 - \text{PER})$ but considering only substitution and dele-

¹<https://huggingface.co/facebook/hubert-large-ls960-ft>

Table 3: Ablation studies on our PER-MDD

ASR model	Top-k	Pool size	Threshold	Strategy	PER\downarrow	REC\uparrow	PRE\uparrow	F1\uparrow	FRR\downarrow	DER\downarrow
Data2vec	10	500	0.7	Mid	188.20	75.25	55.57	63.93	10.04	47.85
Wav2vec2					173.28	74.46	64.58	69.17	6.81	43.54
Hubert					104.08	67.56	71.78	69.60	4.43	37.77
Hubert	5 6 7 8 9 10	500	0.7	Mid	135.27	71.01	72.39	71.69	4.52	39.19
					123.36	69.59	72.71	71.11	4.36	38.31
					116.74	69.10	72.56	70.79	4.36	38.28
					112.36	69.03	72.53	70.73	4.36	37.81
					107.36	68.21	71.67	69.90	4.50	37.38
					104.08	67.56	71.78	69.60	4.43	37.77
Hubert	10	100 200 500 1800	0.7	Mid	140.12	73.55	57.38	64.47	9.11	40.94
					123.49	71.20	65.28	68.11	6.32	40.39
					104.08	67.56	71.78	69.60	4.43	37.77
					84.63	62.83	77.49	69.40	3.04	36.91
Hubert	10	500	No 0.6 0.7 0.8 0.9	Mid	102.10	67.65	71.64	69.59	4.47	37.75
					102.32	67.63	71.74	69.63	4.44	37.77
					104.08	67.56	71.78	69.60	4.43	37.77
					101.90	69.28	67.23	68.24	5.63	41.47
					76.60	73.18	37.14	49.27	20.66	55.16
					All Mean Mid	76.17 101.02 104.08	61.90 65.81 67.56	63.94 67.43 71.78	62.90 66.61 69.60	5.82 5.30 4.43

↓ lower is better ↑ higher is better

tion errors, reaches 90.42%, indicating that our model predicts phonemes largely in agreement with human annotations.

3.4.2. Ablation study

We conduct an ablation study by modifying several key components to evaluate their individual contributions to the performance of our PER-MDD, as summarized in Table 3.

First, we examine the effect of different ASR models on performance by comparing the default HuBERT configuration with two additional models: Data2vec [21]² and Wav2vec2 [22]³. Among them, HuBERT achieves the best performance on this task.

Next, we vary the threshold to quantify its influence on performance. The results show only minor differences between no threshold and a threshold of 0.7, as most top- k candidates already have cosine similarity scores above this level. When the threshold is set higher, at 0.9, PER decreases but F1 is considerably reduced, indicating a necessary trade-off.

We then explore the impact of the top- k value. Decreasing k generally worsens PER, but MDD performance shows slight improvements, suggesting a balance between retrieval precision and coverage.

²<https://huggingface.co/facebook/data2vec-audio-large-960h>

³<https://huggingface.co/facebook/wav2vec2-large-960h-lv60>

We also evaluate different pooling strategies for constructing the vector database, including all-frame pooling, mean-frame pooling, and mid-frame pooling. Among them, the mid-frame strategy achieves the best performance. In addition to better results, it also reduces the pool size, leading to faster query times.

Finally, we vary the pool size, using 100 random utterances, 200 utterances, 500 utterances, and the full set of 1,800 utterances from the training dataset. The results clearly show that larger pool sizes reduce insertion errors and improve MDD metrics overall.

4. CONCLUSIONS

This paper explores the use of a retrieval-based approach to enhance MDD systems. We proposed a novel framework that integrates an ASR model with a retrieval-based method, thereby eliminating the need to train an additional phoneme recognition model. Our approach achieves significant improvements in both FRR and F1 score, reaching an FRR of only 4.43%, the best among compared methods, and delivering a 6.56% F1 gain over SOTA models. These findings underscore the effectiveness of retrieval-based strategies for advancing MDD. However, our model still struggles with insertion errors, which lead to a relatively high PER. In future work, we aim to develop more robust mechanisms for handling insertions and to further optimize retrieval efficiency.

5. REFERENCES

- [1] Silke M Witt and Steve J Young, “Phone-level pronunciation scoring and assessment for interactive language learning,” in *Speech Communication*. 2000, vol. 30, pp. 95–108, Elsevier.
- [2] Wai-Kim Leung, Xunying Liu, and Helen Meng, “Cnn-rnn-ctc based end-to-end mispronunciation detection and diagnosis,” *ICASSP 2019*, pp. 8132–8136, 2019.
- [3] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proc. ICML 2006*, 2006.
- [4] Yiqing Feng, Guanyu Fu, Qingcai Chen, and Kai Chen, “Sed-mdd: Towards sentence dependent end-to-end mispronunciation detection and diagnosis,” *ICASSP 2020*, pp. 3492–3496, 2020.
- [5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, NIPS’17.
- [6] Kaiqi Fu, Jones Lin, Dengfeng Ke, Yanlu Xie, Jinsong Zhang, and Binghuai Lin, “A full text-dependent end to end mispronunciation detection and diagnosis with easy data augmentation techniques,” in *Proc. Interspeech 2021*.
- [7] Wenxuan Ye, Shaoguang Mao, Frank Soong, Wenshan Wu, Yan Xia, Jonathan Tien, and Zhiyong Wu, “An approach to mispronunciation detection and diagnosis with acoustic, phonetic and linguistic (apl) embeddings,” in *ICASSP*, 2022, pp. 6827–6831.
- [8] Linkai Peng, Kaiqi Fu, Binghuai Lin, Dengfeng Ke, and Jinsong Zhan, “A Study on Fine-Tuning wav2vec2.0 Model for the Task of Mispronunciation Detection and Diagnosis,” in *Proc. Interspeech 2021*, pp. 4448–4452.
- [9] Bi-Cheng Yan, Hsin-Wei Wang, Yi-Cheng Wang, and Berlin Chen, “Effective graph-based modeling of articulation traits for mispronunciation detection and diagnosis,” in *ICASSP*, 2023, pp. 1–5.
- [10] Peng Linkai, Yingming Gao, Binghuai Lin, Dengfeng Ke, Yanlu Xie, and Jinsong Zhang, “Text-aware end-to-end mispronunciation detection and diagnosis,” 06 2022.
- [11] Yassine El Kheir, Shammur Absar Chowdhury, and Ahmed Ali, “L1-aware multilingual mispronunciation detection framework,” in *ICASSP*, 2024.
- [12] Hyungshin Ryu, Sunhee Kim, and Minhwa Chung, “A joint model for pronunciation assessment and mispronunciation detection and diagnosis with multi-task learning,” in *Proc. Interspeech*, 2023, pp. 4476–4480.
- [13] Yassine EL Kheir, Shammur Chowdhury, and Ahmed Ali, “Multi-view multi-task representation learning for mispronunciation detection,” in *9th Workshop on Speech and Language Technology in Education (SLaTE)*, 2023, pp. 86–90.
- [14] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang, “Retrieval-augmented generation for large language models: A survey,” 2024.
- [15] Anas Mohammed Alhumud, Muhammad AL-Qurishi, Yasser Omar Alomar, Ali Alzahrani, and Riad Souissi, “Improving automated speech recognition using retrieval-based voice conversion,” in *The Second Tiny Papers Track at ICLR 2024*, 2024.
- [16] Matthew Baas, Benjamin van Niekerk, and Herman Kamper, “Voice conversion with just nearest neighbors,” in *Interspeech 2023*, 2023, pp. 2053–2057.
- [17] Xing Wei and Wenwei Dong and Catia Cucchiarini and Roeland van Hout and Helmer Strik, “Leveraging Articulatory Information to Enhance End-to-End Models for L2 Mispronunciation Detection and Diagnosis,” in *10th Workshop on Speech and Language Technology in Education (SLaTE)*, 2025, pp. 76–80.
- [18] Binghuai Lin and Liyuan Wang, “Phoneme mispronunciation detection by jointly learning to align,” in *ICASSP 2022*, pp. 6822–6826.
- [19] Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna, “L2-ARCTIC: A Non-native English Speech Corpus,” in *Proc. Interspeech*, 2018.
- [20] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, pp. 3451–3460, Oct. 2021.
- [21] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli, “Data2vec: A general framework for self-supervised learning in speech, vision and language,” in *ICML*, 2022, pp. 1298–1312.
- [22] Alexei Baevski, Yuhao Zhou, and Michael Auli Abdelrahman Mohamed, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *NeurIPS*, vol. 33, 2020.