# PHONEME MISPRONUNCIATION DETECTION BY JOINTLY LEARNING TO ALIGN

*Binghuai Lin*[1], *Liyuan Wang*[1]

[1]Smart Platform Product Department, Tencent Technology Co., Ltd, China

## ABSTRACT

Phoneme mispronunciation detection plays an important role in Computer-Assisted Pronunciation Training. Traditional methods either rely on phone recognition, which has the limitation of incapability of detecting mispronounced phonemes out of the dictionary, or the need of external phoneme alignment for extracting acoustic features. In this paper, we propose a method for phoneme mispronunciation detection by jointly learning to align. Specifically, we first obtain acoustic and canonical phoneme representations utilizing acoustic and phoneme encoders. Second, we utilize an attention mechanism to fuse acoustic features of each frame and phoneme representations. Finally, a convolutional neural network (CNN)-based layer following the fused representations is utilized for better exploring local context. The network is jointly optimized for phoneme mispronunciation and phoneme alignment based on a multi-task learning framework. Experimental results based on a public dataset L2-ARCTIC show the state-of-the-art (SOTA) performance with an F1-score of 63.04%. It is also found that optimizing the phoneme alignment can further improve the performance of phoneme mispronunciation detection.

***Index Terms***— Mispronunciation detection, alignment, attention, multi-task learning, phoneme classification

## 1. INTRODUCTION

Non-native speakers are heavily influenced by their native tongue (L1) when learning the target language (L2). The common approach to tackle this problem is through Computer-Assisted Pronunciation Training (CAPT). A system for CAPT should provide detailed feedback on pronunciation errors for language learners [1].

Commonly, there are two different approaches for phoneme mispronunciation detection: (1) automatic speech recognition (ASR)-based methods for phone recognition; (2) feature-based mispronunciation classification depending on extra phoneme alignment. Many previous studies have proposed to adopt ASR-based methods for phoneme mispronunciation detection. The phonological rules are incorporated to Extended Recognition Network (ERN) in traditional speech recognizers to capture the mispronounced phonemes [2]. Other methods such as APM (acoustic-phonemic model) utilized acoustic features and phonetic information as input and calculated the phone-state posterior probabilities based on deep neural network (DNN) to generate the recognized phoneme sequence for phoneme mispronunciation detection [3]. With the development of end-to-end (E2E) ASR, some E2E mispronunciation detection and diagnosis (MDD) models have been developed. A convolutional neural network (CNN), recurrent neural network (RNN), and connectionist temporal classification (CTC) (CNN-RNN-CTC)-based speech recognition for MDD has been proposed and achieved great improvement than previous work [4]. A hybrid CTC-Attention model together with a novel anti-phone modeling technique has also been proposed for MDD tasks [5]. Recently, a self-supervised pre-training (SSP) model called wav2vec 2.0 [6] optimized by CTC loss has been found effective in mispronunciation detection [7]. Though they can achieve promising results, these ASR-based methods need to be guaranteed to cover all mispronunciations in the dictionary. Also, inconsistency of optimization goals between ASR and mispronunciation detection might have a negative impact on the performance [8]. Recent works tackled these problems by optimizing both ASR and mispronunciation detection [9, 10].

Feature-based methods for mispronunciation detection commonly focus on the optimization of feature extraction. A variation of the posterior probability ratio called Goodness of Pronunciation (GOP) was proposed for pronunciation evaluation and error detection [11]. With the development of DNN, feature learning and pronunciation mispronunciation detection can be optimized jointly. Long short-term memory recurrent network (LSTM) was adopted to extract feature representations for features such as speech attributes and phone features for mispronunciation detection [12]. A CNN was used to extract features for pronunciation mispronunciation detection based on the MLP classifier [13]. Wav2vec 2.0 relying on external alignment was explored for phoneme mispronunciation as well. These feature-based methods commonly rely on extra phoneme alignment information, and the accuracy of alignment need to be optimized separately [14].

Instead of separate optimization, we seek for a joint alignment optimization scheme to facilitate mispronunciation detection. Many works have been studied for phoneme alignment. The common approach for forced alignment is based on a generative model of the speech signal using Hidden

These authors annotated with [1] contributed equally to this work.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

Markov Models (HMM). HMM for each acoustic phone is trained, and the sequence of frames in an utterance is aligned to the phone sequence by finding the sequence of hidden states that maximizes the utterance likelihood [15]. However, as phoneme boundaries are not represented in the model, the training procedure does not explicitly optimize for boundary accuracy [16]. Some work proposed a discriminative method for phoneme alignment and treated it as a problem of sequence labelling for phoneme classes and boundaries based on LSTM [17]. Here, we also treat phoneme alignment as a sequential labelling problem such that it can be optimized jointly with mispronunciation detection.

In this paper, we propose a method for phoneme mispronunciation detection with joint optimization of alignment instead of separate ones. First, speech is encoded into acoustic representations, and canonical phonemes are encoded into phoneme embeddings based on an acoustic and a phoneme encoder, respectively. Second, the acoustic representation of each frame is fused with phoneme embeddings based on the attention mechanism, which has been widely used in machine translation [18] and speech recognition [19]. Last, acoustic features together with weighted phoneme representations are fed into a CNN-based layer to better explore local context [20], and then phoneme mispronunciation and phoneme alignment are optimized jointly based on a multi-task learning (MTL) framework. We conduct experiments based on a public dataset L2-ARCTIC [21] and achieves the state-of-the-art (SOTA) performance with an F1-score of 63.04%. We further explore the relationship between the accuracy of alignment and phoneme mispronunciation detection, and find that the improvement of alignment can facilitate mispronunciation detection.

## 2. PROPOSED METHOD

The proposed network is composed of an acoustic encoder and a phoneme encoder as shown in Figure 1. The acoustic encoder takes speech signals as input and outputs acoustic representations. The phoneme encoder utilizes different numerical representations for different phoneme types in the phoneme set. An attention mechanism is utilized to combine these two representations. An additional CNN layer is implemented for the combined representations for better capture of local context. The detection task of phoneme mispronunciation and alignment task of phoneme classification are jointly optimized based on an MTL framework.

### 2.1. Feature representation learning

The acoustic representations for $i$th frame of the utterance, denoted as $h^i_{\text{speech}}$, are extracted from a pre-trained acoustic encoder called wav2vec 2.0 [6]. It is a framework for self-supervised learning of representations from raw audio data. It is composed of a multi-layer convolutional feature encoder and optimized by contrastive loss based on large amounts of unlabeled data. Taking into account that each phone type in the phone set shows distinct characteristics [21], we employ different numerical representations for different phonemes. The feature representations for $j$th phoneme are denoted as $h^j_{\text{phone}}$.

Based on the aforementioned acoustic and phoneme representations, we fuse them based on the multi-head attention (MHA) mechanism following the previous work [18]. MHA models the relationship between queries, keys, and values. The attention weight can be defined as Eq. (1) and Eq. (2):

$$\text{AttentionScore}(Q, K) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}}) \tag{1}$$

$$\text{Attention}(Q, K, V) = \text{AttentionScore}(Q, K) * V \tag{2}$$

where $Q$ denotes the queries and $K$ denotes the keys with the dimensionality of $d_k$. $V$ represents the values with the dimensionality of $d_v$.

In this paper, we take representations of speech $H_{\text{speech}}$ as queries and representations of phonemes $H_{\text{phone}}$ as keys and values. We obtain weighted sum of representations of these relevant phonemes based on Eq. (2). Then, the fused representation for $i$th frame is obtained by concatenating acoustic and summed phoneme representations as shown in Eq. (3).

$$h^i_{\text{fusion}} = [\text{Attention}(h^i_{\text{speech}}, H_{\text{phone}}), h^i_{\text{speech}}] \tag{3}$$

### 2.2. Training and inference of phoneme mispronunciation

The fused frame representations contains the weighted sum of phoneme representations that explores the global context influence through MHA. We can further explore the local context around one frame by adding an extra CNN-based layer on top of the attention layer for better frame classification. Two separate fully connected (FC) layers following the CNN layer are utilized for mispronunciation and phoneme classification, respectively. Note that here we treat phoneme alignment as sequential labelling of phoneme classes. The cross-entropy classification loss between the predicted results and true labels is defined as shown in Eq. (4) and Eq. (5), where $m$ is the total number of frames, $c_{\text{phn}}$ is the number of phoneme classes, and $c_{\text{mis}}$ is the number of mispronunciation classes. The network is optimized based on an MTL framework. The total loss is defined as Eq. (6), where $\lambda$ is a hyper-parameter to balance the loss among tasks of phoneme and mispronunciation classification.

$$L_{\text{phone}} = -\sum_{i=1}^{m}\sum_{j=1}^{c_{\text{phn}}} {y_{\text{phone}}}_i^j \times \log({p_{\text{phone}}}_i^j) \tag{4}$$

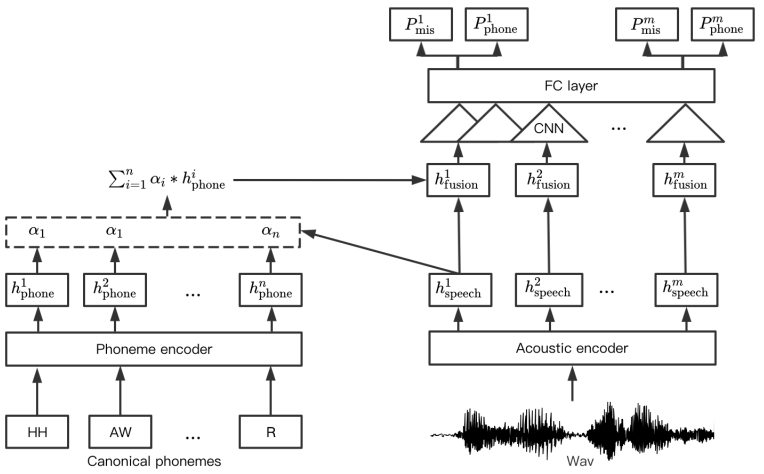

**Fig. 1**. Network structure of mispronunciation detection

$$L_{\text{mis}} = -\sum_{i=1}^{m}\sum_{j=1}^{c_{\text{mis}}} {y_{\text{mis}}}_i^j \times \log({p_{\text{mis}}}_i^j) \tag{5}$$

$$L_{\text{total}} = \lambda L_{\text{phone}} + (1-\lambda) L_{\text{mis}} \tag{6}$$

Once the network is trained, we can perform the inference which results in probability sequences of phoneme and mispronunciation predictions at the frame level. We average the mispronunciation probabilities for the adjacent frames in the sequence in case of the same predicted phoneme. This will result in a sequence of predicted phonemes as well as their corresponding averaged mispronunciation probabilities. Finally, we align the predicted phoneme sequence with the canonical one. The mismatched canonical phonemes will be treated as mispronounced, or even if matched, those with averaged mispronunciation probabilities higher than a predefined threshold will be treated as mispronounced as well.

## 3. EXPERIMENTS

### 3.1. Experimental setup

The acoustic representations extracted from wav2vector 2.0 has the dimensionality of 384. There are 40 phonemes (1 for silence and the rest for 39 phonemes) in total, where each of them is encoded into an embedding with a dimensionality of 32. Acoustic representations are transformed to the dimensionality of 32 to match that of phoneme embeddings. The frame-level acoustic representations are aligned to representations of canonical phoneme sequences based on MHA with one head. The input channel, output channel, and kernel size of the CNN layer are $64$, $64$, and $9$, respectively. The two FC layers for mispronunciation and phoneme classification have the same dimensionality of $64 * 2$. $\lambda$ applied in MTL is set to be $0.5$.

Before joint optimization of mispronunciation and phoneme classification, we use TIMIT [22] corpus for pre-training the phoneme classification part of the proposed network for better initialization. The TIMIT corpus is a native corpus containing a total of 6300 sentences, with 10 sentences spoken by each of 630 speakers. It includes time-aligned orthographic, phonetic, and word transcriptions for each utterance. We pre-train the network based on the TIMIT corpus and achieve an F1-score of 93.13% with a tolerance of 20ms.

The L2-ARCTIC [21] corpus is then utilized for joint optimization as well as evaluation of the performance of mispronunciation detection. It is a speech corpus of non-native English intended for mispronunciation detection. It contains different types of mispronunciation errors such as substitutions, deletions, and additions. We use the common train and test split as previous work: six speakers (NJS, TLV, TNI, TXHC, YKWK, ZHAA) are selected as test sets and the rest are train sets. Note that for consistency, we map the 61-phone setting of the TIMIT corpus to the 39-phone setting of the L2-ARCTIC corpus.

We evaluate the proposed method using Precision (PR), Recall (RE), F1-score metrics. Given canonical phonemes, true acceptation (TA) and true rejection (TR) indicate that phonemes with correct/incorrect pronunciation are classified correctly, while false acceptation (FA) and false rejection (FR) indicate wrong classification results. Based on the definitions, we these metrics to evaluate the performance.

### 3.2. Comparative study

We compare the results of our proposed method with those of two feature-based and two ASR-based methods. Two feature-based methods relying on extra external alignment include GOP-based (GOP) [23] and Wav2vec-based (W2V) method [14], and two ASR-based methods include CTC-Attention-based (CTC-ATTN) method [5] and Wav2vec-CTC-based (W2V-CTC) method [24]. Results are shown in Table 1.

**Table 1**. Comparison of ours and previous work

| Methods | Model | PR(%) | RE(%) | F1(%) |
|---|---|---|---|---|
| Feature | GOP[23] | 46.31 | 53.82 | 49.78 |
| | W2V [14] | 58.0 | 64.3 | 61.0 |
| ASR | CTC-ATTN[5] | 46.57 | 70.28 | 56.02 |
| | W2V-CTC [24] | 62.86 | 58.2 | 60.44 |
| Joint | Proposed | 77.12 | 53.31 | **63.04** |

The results demonstrate the proposed method outperforms previous work by achieving the SOTA performance in F1-score. The 2% advantage in F1-score when compared to feature-based methods can be attributed to the joint optimization of phoneme alignment and mispronunciation detection, and the 3% gap between the ASR-based methods and ours could be due to the inconsistent goals between ASR training and mispronunciation detection.

### 3.3. Ablation studies

To validate the effectiveness of the proposed method, we carry out some ablation studies. First, we demonstrate the impact of network initialization based on the TIMIT corpus. Second, we illustrate the necessity of important components including the CNN and attention layer by removing one of them. Note that removing the attention layer indicates training without extra canonical phoneme information. Results are shown in Table 2, which clearly demonstrate the importance of pre-training as well as the CNN and attention layers.

**Table 2**. Comparison with different settings

| Model | PR(%) | RE(%) | F1(%) |
|---|---|---|---|
| Proposed(No pre-training) | 76.32 | 51.13 | 61.23 |
| Proposed(No CNN) | 59.14 | 61.30 | 60.2 |
| Proposed(No attention) | 71.23 | 51.27 | 59.62 |
| Proposed | 77.12 | 53.31 | **63.04** |

### 3.4. Relationship between phoneme alignment and mispronunciation detection

To further validate the effectiveness of joint optimization, we show the mispronunciation detection (MD) performance under phoneme alignment with different levels of accuracy. We experiment with golden alignments, i.e. human-labelled alignments, as the upper bound of alignment accuracy. Specifically, we assign phoneme embeddings for each frame based on the golden alignments and perform the classification calculation. Results are shown in the Figure 2.

From the figure, we can see the MD performance improves as phoneme alignment accuracy increases. Unsurprisingly, it achieves the best performance with golden alignments. The performance is also bounded by factors such as better feature representations other than the alignment accuracy.

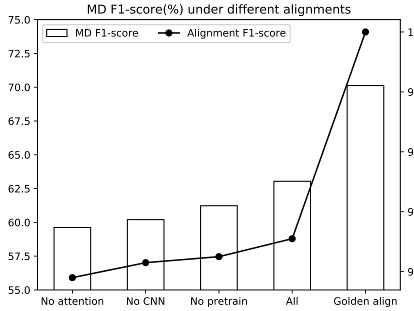

**Fig. 2**. MD under different alignment accuracy

## 4. CONCLUSION

In this paper, we propose a method for phoneme mispronunciation detection by jointly learning to align. An attention mechanism is implemented for fusing acoustic and phoneme representations, which are derived from a pre-trained acoustic encoder and a phoneme encoder, respectively. We conduct experiments base on the L2-ARCTIC corpus and achieve the SOTA performance with an F1-score of $63.04\%$. By investigating the relationship between the performance of mispronunciation detection and the accuracy of phoneme alignment, we demonstrate the importance of good alignment and the advantage of joint optimization. In the future, we will investigate other factors that influence mispronunciation detection performance.

## 5. REFERENCES

[1] A Neri, C Cucchiarini, and H Strik, "Feedback in computer assisted pronunciation training: When technology meets pedagogy," in *Proceedings of the 10th International CALL Conference on" CALL professionals and the future of CALL research", University of Antwerp, Belgium*. Antwerpen: Universiteit Antwerpen, 2002, pp. 179–188.

[2] Wai-Kit Lo, Shuang Zhang, and Helen Meng, "Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system," in *Eleventh annual conference of the international speech communication association*, 2010.

[3] Kun Li, Xiaojun Qian, and Helen Meng, "Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks," *IEEE/ACM*

*Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 193–207, 2016.

[4] Wai-Kim Leung, Xunying Liu, and Helen 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.

[5] Bi-Cheng Yan, Meng-Che Wu, Hsiao-Tsung Hung, and Berlin Chen, "An end-to-end mispronunciation detection system for l2 english speech leveraging novel antiphone modeling," *arXiv preprint arXiv:2005.11950*, 2020.

[6] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[7] Minglin Wu, Kun Li, Wai-Kim Leung, and Helen Meng, "Transformer based end-to-end mispronunciation detection and diagnosis," *Proc. Interspeech 2021*, pp. 3954–3958, 2021.

[8] Long Zhang, Ziping Zhao, Chunmei Ma, Linlin Shan, Huazhi Sun, Lifen Jiang, Shiwen Deng, and Chang Gao, "End-to-end automatic pronunciation error detection based on improved hybrid ctc/attention architecture," *Sensors*, vol. 20, no. 7, pp. 1809, 2020.

[9] Daniel Korzekwa, Jaime Lorenzo-Trueba, Thomas Drugman, Shira Calamaro, and Bozena Kostek, "Weakly-supervised word-level pronunciation error detection in non-native english speech," *arXiv preprint arXiv:2106.03494*, 2021.

[10] Zhan Zhang, Yuehai Wang, and Jianyi Yang, "Text-conditioned transformer for automatic pronunciation error detection," *Speech Communication*, vol. 130, pp. 55–63, 2021.

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

[12] Wei Li, Nancy F Chen, Sabato Marco Siniscalchi, and Chin-Hui Lee, "Improving mispronunciation detection for non-native learners with multisource information and lstm-based deep models," in *INTERSPEECH*, 2017, pp. 2759–2763.

[13] Ann Lee et al., *Language-independent methods for computer-assisted pronunciation training*, Ph.D. thesis, Massachusetts Institute of Technology, 2016.

[14] Xiaoshuo Xu, Yueteng Kang, Songjun Cao, Binghuai Lin, and Long Ma, "Explore wav2vec 2.0 for mispronunciation detection," *Proc. Interspeech 2021*, pp. 4428–4432, 2021.

[15] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi.," in *Interspeech*, 2017, vol. 2017, pp. 498–502.

[16] Andreas Stolcke, Neville Ryant, Vikramjit Mitra, Jiahong Yuan, Wen Wang, and Mark Liberman, "Highly accurate phonetic segmentation using boundary correction models and system fusion," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 5552–5556.

[17] Felix Kreuk, Yaniv Sheena, Joseph Keshet, and Yossi Adi, "Phoneme boundary detection using learnable segmental features," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8089–8093.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[19] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[20] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi, "Understanding of a convolutional neural network," in *2017 International Conference on Engineering and Technology (ICET)*. Ieee, 2017, pp. 1–6.

[21] Guanlong Zhao, Sinem Sonsaat, Alif O Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna, "L2-ARCTIC: A non-native english speech corpus," *Perception Sensing Instrumentation Lab*, 2018.

[22] John S Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium, 1993*, 1993.

[23] Wenping Hu, Yao Qian, and Frank K Soong, "A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL)," in *Interspeech*, 2013, pp. 1886–1890.

[24] 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," *Proc. Interspeech 2021*, pp. 4448–4452, 2021.