



# An Alignment Method Leveraging Articulatory Features for Mispronunciation Detection and Diagnosis in L2 English

Qi Chen<sup>1</sup>, Binghuai Lin<sup>2</sup>, Yanlu Xie<sup>1</sup>

<sup>1</sup>Beijing Language and Culture University, China

<sup>2</sup>Smart Platform Product Department, Tencent Technology Co., Ltd, China

cifsqq@163.com, binghuailin@tencent.com, xieyanlu@blcu.edu.cn

## Abstract

Mispronunciation Detection and Diagnosis (MD&D) technology is used for detecting mispronunciations and providing feedback. Most MD&D systems are based on phoneme recognition. However, few studies have made use of the predefined reference text which has been provided to second language (L2) learners while practicing pronunciation. In this paper, we propose a novel alignment method based on linguistic knowledge of articulatory manner and places to align the phone sequences of the reference text with L2 learners speech. After getting the alignment results, we concatenate the corresponding phoneme embedding and the acoustic features of each speech frame as input. This method makes reasonable use of the reference text information as extra input. Experimental results show that the model can implicitly learn valid information in the reference text by this method. Meanwhile, it avoids introducing misleading information in the reference text, which will cause false acceptance (FA). Besides, the method incorporates articulatory features, which helps the model recognize phonemes. We evaluate the method on the L2-ARCTIC dataset and it turns out that our approach improves the F1-score over the state-of-the-art system by 4.9% relative.

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

## 1. Introduction

Computer-assisted pronunciation training (CAPT) provides L2 learners with personalized pronunciation feedback so that they can learn a foreign language in the absence of language teachers. Mispronunciation detection and diagnosis (MD&D) is one of the key technologies of the CAPT system. The module is helpful for detecting mispronunciation and offering instructive feedback.

Most of the existing MD&D methods work by recognizing the phone sequence of a speech and comparing it with the canonical phone sequence. Differences between the recognized phones and the canonical phones are identified as mispronunciations. Goodness of Pronunciation (GOP) [1] proposed by Witt provides pronunciation scores for L2 learners by calculating the logarithmic posterior probability of phonemes. Extended Recognition Network (ERN) [2] aims to evaluate the details of pronunciation errors and provide diagnostic feedback on specific errors. Acoustic Phonological Model (APM) proposed in [3] adopts multi-distribution DNN to predict the most likely mispronunciations. Recently, an automatic speech recognition method based on Connectionist Temporal Classification (CTC) [4] has also been introduced to MD&D tasks. Leung et al. proposed an end-to-end MD&D model CNN-RNN-CTC [5], which

does not require forced alignment and achieves better performance than traditional methods. In recent researches, articulatory features are employed to boost MD&D, for the reason that they can describe the pronunciation mechanisms and positions of articulators [6]. A decision tree-based framework proposed in [7] detects mispronunciations caused by using inaccurate speech attributes. Mao proposed several model architectures[8] for exploiting articulatory features in MD&D.

The process of MD&D is that the system first gives the learner a predefined reference text and a pronunciation demonstration, and the learner tries to imitate the standard pronunciation of the reference text. The system performs phone recognition, then the recognized phone sequence is aligned with the canonical phone sequence of the reference text to detect mispronunciations. Since the reference text is known, it is a waste to ignore this prior knowledge in the phone recognition process. Therefore, it can be used as another input to the model in addition to the acoustic features. How to introduce the reference text information reasonably becomes a crucial problem in the MD&D task.

The acoustic-phonemic model (APM) proposed in [3] aligns Mel-frequency cepstral coefficients (MFCC) and canonical phonemes in the reference text by forced alignment. Then, MFCC and 7 canonical phonemes are used as acoustic features and phonemic features respectively. SED-MDD[9] uses an attention-based method to align the phonemes in the reference text with the acoustic features of each speech frame by calculating their attention weights. The aligned acoustic features and the reference text information are then concatenated as input of the model. Through introducing reference text information, these methods improve MD&D performance. However, given that the pronunciation produced by L2 learners often does not match the standard pronunciation, indiscriminate use of reference text information will cause false acceptance (FA), in that model predictions are close to standard pronunciations and fail to detect mispronunciations. This situation will impair the MD&D performance.

In this paper, we propose an alignment method based on articulatory features to align the phone sequences of the reference text with L2 learners' speech. The manner and places of articulation are used as constraints to align the reference text with the speech. After getting the alignment results, we concatenate the corresponding phoneme embedding and the acoustic features of each speech frame as input of the model. In this way, the reference text information is introduced as extra input. With the help of articulatory constraints, our approach utilizes valid information in the reference text while avoiding the introduction of misleading information and causing false acceptance.

## 2. Method

### 2.1. Articulatory Features

The articulatory features can directly measure pronunciation quality and give corrective feedback based on articulatory manner and places, such as the height of the tongue. According to relevant linguistic studies, we use the articulatory feature space proposed in [10] to describe articulatory features from eight dimensions. Each dimension is divided into several blocks. The articulatory feature space is shown in Table 1. Moreover, we use the mapping table in [10] to map phonemes to articulatory features.

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

Since articulatory features contain distinctive information, especially for some phoneme pairs that are easily confusing, the use of articulation features in MD&D helps to recognize phonemes.

### 2.2. Alignment Method Leveraging Articulatory Features

The purpose of alignment is to find the phoneme corresponding to each speech frame. The distinctive information contained in the articulatory features contributes to the process. We obtain the alignment results by calculating the similarities between the articulatory features of speech and the articulatory features of the canonical phone sequence.

To begin with, we use a fine-tuned pre-trained model Wav2vec 2.0[11] as an extractor to extract articulatory features  $(d_1, \dots, d_t)$ .  $d_i$  represents the articulatory features of the  $i_{th}$  speech frame. Then, we map the canonical phone sequence to articulatory features by the mapping table. The articulatory features of the canonical phone sequences are denoted by  $(h_1, \dots, h_n)$ . Inspired by [12], we take a monotonic chunk-wise attention mechanism to perform soft attention over small ‘‘chunks’’ for alignment. The chunk-wise attention mechanism splits the canonical phone sequence into small chunks. Specifically, the location of chunks will be set adaptively. We align each speech frame with the canonical phones in chunks. Chunk size is set to 3  $(h_j, h_{j+1}, h_{j+2})$ . The similarities of  $d_i$  and  $h_j$ ,  $d_i$  and  $h_{j+1}$ ,  $d_i$  and  $h_{j+2}$  in the small chunk are calculated by:

$$x = d_i - h_j \quad (1)$$

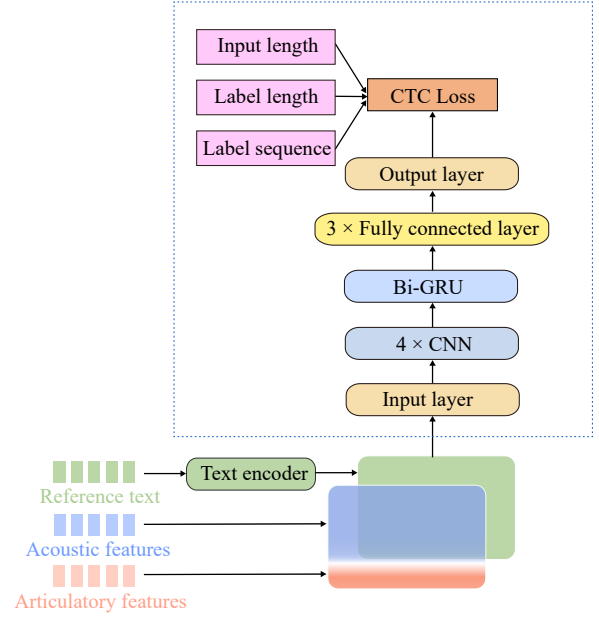


Figure 1: *Two-channel input by combining three different features*

$$similarity = \begin{cases} 1 - \frac{\sum_{i=0}^5 \frac{x_i}{cardinality_i - 1}}{6} \\ 0, \text{ if } x_6 = 1 \\ 0, \text{ if } x_7 = 1 \end{cases} \quad (2)$$

where  $x_i$  denotes the  $i_{th}$  dimension of articulatory features. Figure 2 depicts the proposed alignment method. Details of the monotonic chunk-wise attention are shown in [12].

If any one of the three similarities is greater than the threshold (set to 0.8), we will take a text encoder to extract phone embeddings of the three canonical phones in the chunk. The extracted embeddings are the same size as the concatenation of acoustic features and articulatory features. Then, we have:

$$\alpha_{ij} = \frac{\exp(similarity(d_i, h_j))}{\sum_{t=j}^{j+2} \exp(d_i, h_t)} \quad (3)$$

$$c_i = \sum_{t=j}^{j+2} \alpha_{ij} h_t \quad (4)$$

The phone embeddings are combined with the concatenation of acoustic features and articulatory features into two-channel features as input of the model.

If all the three similarities are smaller than the threshold, the phoneme represented by  $d_i$  will not be considered to exist in the standard phoneme sequence, and the 0-padding embedding will be combined with the concatenation of acoustic features and articulatory features. This operation makes the alignment meet the first two criteria mentioned in the following part. Figure 1 illustrates the process of feature usage.

With the use of unrestricted soft attention, it is possible for the occurrence of speech mismatching with reference text, often with skipping, repeating, or attention collapse (unintelligible gibberish when the model fails to focus on an input token).

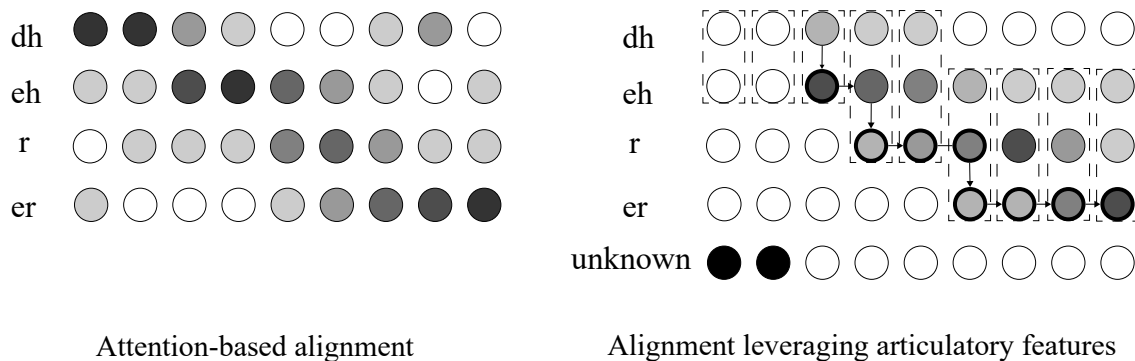


Figure 2: Illustration for different attention mechanisms

In addition, when an insertion error occurs, the inserted speech frames that do not appear in the reference text will still be concatenated with the reference text phoneme. The system may recognize the mispronunciation as a phoneme in the reference text, resulting in false acceptance of the mispronunciation. To address this problem, our proposed alignment method is designed to adhere to the following criteria:

1. In case of the appearance of an insertion error, the frame at the current time step cannot be matched with any phoneme of the reference text. The alignment mechanism should align the frame with an unknown label.
2. The alignment mechanism should detect and skip the missed phoneme in the reference text as a deletion error occurs.
3. The alignment of speech and reference text should be monotonical (the chunk can only keep unmoved or move forward) to avoid repetition.

Figure 2 illustrates the comparison of two different alignment methods being discussed.

### 3. Experiments

#### 3.1. Speech Corpus

We conducted experiments on the TIMIT [13] and L2-ARCTIC [14], both of which are publicly available corpora. TIMIT is a native (L1) English corpus containing 6,300 utterances from 630 speakers. L2-ARCTIC is a non-native English corpus built for MD&D, which encompasses recordings of 24 non-native speakers (12 male and 12 female) whose L1 languages include Hindi, Korean, Mandarin, Spanish, Arabic, and Vietnamese. TIMIT is labeled with 61 phones, while L2-ARCTIC is labeled with 48 phones and a /err/ label which means the pronunciation does not belong to any canonical phones. For training purposes, we mapped TIMIT 61-phone to 39-phone according to the mapping table from [15] and merged it with the L2-arctic phone set.

#### 3.2. Experimental Setup

We use 80-dimensional filter-banks as acoustic features, and the Cepstral Mean and Variance Normalization (CMVN) is used to

reduce the differences of speakers. Moreover, the CNN-RNN-CTC model in [5] is configured as our baseline model in combination with the proposed alignment method for experiments. We take the attention-based method mentioned in [9] as a comparison. All the CNN-RNN-CTC models use the same model hyperparameters unless otherwise specified.

A text encoder consists of two fully connected layers and a bidirectional LSTM[16] is designed to extract robust phone embeddings from reference text. The input to the text encoder is a sequence of phones, where each phone is represented as an embedding. We use two fully connected layers and the Relu activation function [17] to obtain the bottleneck feature of the phone sequence. Then, the processed sequence is passed to a bidirectional LSTM to extract phone embedding.

The model and CTC loss function are implemented using PyTorch. During the training process, the model passes the input labels, label length, input length, and output from the model to the function to calculate the loss. The recognized phoneme sequences are aligned with the labeled phoneme sequences by using the Needleman-Wunsch algorithm [18]. The aligned results are used to calculate the insertion errors, deletion errors, substitution errors, and other metrics of the model.

#### 3.3. Evaluation

In evaluation, we follow the metrics of a previous study [19]. The true acceptance (TA) and true rejection (TR) rates indicate correct pronunciation detection, while the false rejection (FR) and false acceptance (FA) indicate incorrect detection. The false rejection rate (FRR) and false acceptance rate (FAR) are calculated by Equation 5 and Equation 6, respectively.

$$FRR = \frac{FR}{TA + FR} \quad (5)$$

$$FAR = \frac{FA}{FA + TR} \quad (6)$$

Precision, Recall, and F-measure are evaluation criteria widely used to measure the performance of MD&D, with the following equations.

$$Recall = \frac{TR}{TR + FA} \quad (7)$$

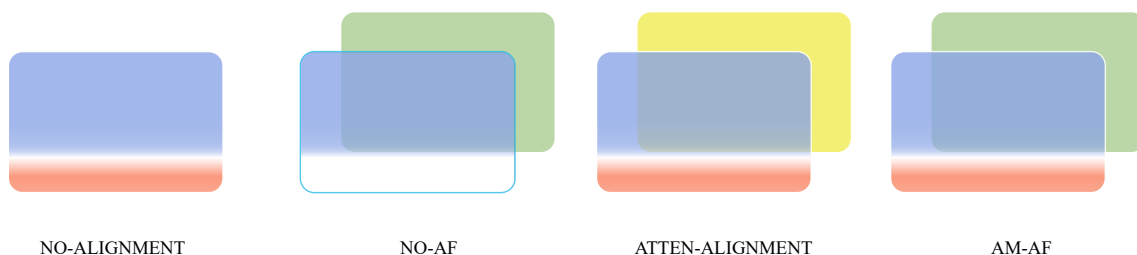


Figure 3: Usage of different features in the ablation study

$$Precision = \frac{TR}{TR + FR} \quad (8)$$

$$F - measure = 2 * \frac{PrecisionRecall}{Precision + Recall} \quad (9)$$

In the practical application of MD&D products, if many correct pronunciations are considered as mispronunciations, the L2 learners will be confused. Therefore, the trade-off between FAR and FRR is an important evaluation metric [24], and we utilize the DCF (arithmetic mean of FAR and FRR) to evaluate the performance of the system. The formula of Detection Cost Function (DCF) is as follows:

$$DCF = C_{FR} * FRR * P_{target} + C_{FA} * FAR * (1 - P_{target}) \quad (10)$$

where  $C_{FR}$  is the cost of false rejection.  $C_{FA}$  is the cost of false acceptance. In this experiment, the former is set to 0.9, and the latter is set to 0.1.  $P_{target}$  is the prior probability, set to 0.9.

### 3.4. Performance of MD&D

The MD&D results are presented in Table 2. With the help of the alignment method, our approach(AM-AF) outperforms the GOP and the approach in [20]. Compared with CTC-ATT, the relative improvement in F1-score is 4.9%.

Table 2: Performance of mispronunciation detection and diagnosis with different approaches.

Models	Recall(%)	Precision(%)	F1-score(%)
GOP	35.42	52.88	42.42
CTC-ATT	46.57	70.28	56.2
AM-AF	65.38	53.41	58.79

Considering that our method uses multiple features, now we investigate which element contributes the most to the performance and whether the alignment method we proposed is effective. Along with the mentioned method, we trained three additional variants, each with a certain feature removed. NO-ALIGNMENT does not include reference text information but

articulatory features. NO-AF does not incorporate articulatory features, but uses the proposed alignment method to include reference text information. ATTEN-ALIGNMENT uses an attention-based method to align reference text with acoustic features, including both reference text information and articulatory features. Figure 3 illustrates the usage of different features in the ablation study. The introduction of the reference text information is significantly beneficial to improve MD&D performance. The proposed alignment method reduces false rejections as well as false acceptances compared to the attention-based alignment method. Our approach also has better performance from the DCF perspective.

Table 3: Ablation study.

Models	FRR(%)	FAR(%)	DCF
NO-ALIGNMENT	28.30	34.71	0.23
NO-AF	14.87	30.62	0.12
ATTEN-ALIGNMENT	21.54	36.92	0.18
AM-AF	12.24	24.62	0.10

## 4. Conclusion

In this paper, we proposed an alignment method based on articulatory features for aligning reference text and speech in the MD&D task. We evaluate our approach on two publicly available corpora, and experiments show that with this approach, the model can implicitly learn valid information in the reference text while avoiding introducing misleading information that causes false acceptance.

## 5. Acknowledgements

This work is supported by National Social Science Foundation of China (18BYY124), Center for Language Education and Cooperation (YHJC21YB-128), Advanced Innovation Center for Language Resource and Intelligence (KYR17005), Science Foundation of Beijing Language and Culture University (the Fundamental Research Funds for the Central Universities) (19PT04), the Science Foundation and Special Program for Key Basic Research fund of Beijing Language and Culture University (21YJ040004). The corresponding author of the paper is Yanlu Xie.

## 6. References

- [1] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, pp. 95–108, 2000.
- [2] 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.
- [3] K. Li, X. Qian, and H. 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] 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*, 2006, pp. 369–376.
- [5] 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.
- [6] P. Ladefoged, "A course in phonetics," 5th ed. Boston, MA: Thomson Wadsworth, 2006.
- [7] W. Li, K. Li, S. M. Siniscalchi, N. F. Chen, and C.-H. Lee, "Detecting mispronunciations of L2 learners and providing corrective feedback using knowledge-guided and data-driven decision trees," in *INTERSPEECH*, 2016.
- [8] S. Mao, Z. Wu, X. Li, R. Li, X. Wu, and H. M. Meng, "Integrating articulatory features into acoustic-phonemic model for mispronunciation detection and diagnosis in L2 English speech," *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2018.
- [9] 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.
- [10] J. Tepperman and S. S. Narayanan, "Using articulatory representations to detect segmental errors in nonnative pronunciation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 8–22, 2008.
- [11] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec2.0: A framework for self-supervised learning of speech representations," vol. 33, 2020.
- [12] C.-C. Chiu and C. Raffel, "Monotonic chunkwise attention," *ArXiv*, vol. abs/1712.05382, 2018.
- [13] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continuous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.
- [14] G. Zhao, S. Sonsaat, A. O. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-arctic: A non-native English speech corpus," *Perception Sensing Instrumentation Lab*, 2018.
- [15] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.
- [16] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *ArXiv*, vol. abs/1508.01991, 2015.
- [17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, pp. 84 – 90, 2012.
- [18] L. Muffikhah and R. Eka, "An improved needleman-wunsch algorithm for pairwise sequence alignment of protein-albumin," *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 10, pp. 83–87, 2018.
- [19] X. Qian, H. M. Meng, and F. K. Soong, "Capturing L2 segmental mispronunciations with joint-sequence models in computer-aided pronunciation training (capt)," *2010 7th International Symposium on Chinese Spoken Language Processing*, pp. 84–88, 2010.
- [20] B.-C. Yan, M.-C. Wu, H.-T. Hung, and B. Chen, "An End-to-End Mispronunciation Detection System for L2 English Speech Leveraging Novel Anti-Phone Modeling," in *Proc. Interspeech 2020*, 2020, pp. 3032–3036. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2020-1616>