applied sciences

MDPI

Article

End-to-End Mispronunciation Detection and Diagnosis Using Transfer Learning

Linkai Peng [1,2,†], Yingming Gao [3,†], Rian Bao [1], Ya Li [3] and Jinsong Zhang [1,*]

[1] School of Information Science, Beijing Language and Culture University, Beijing 100083, China; penglinkai96@gmail.com (L.P.); boroooo@163.com (R.B.)
[2] NetEase Youdao, Beijing 100193, China
[3] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China; yingming.gao@bupt.edu.cn (Y.G.); yli01@bupt.edu.cn (Y.L.)
* Correspondence: jinsong.zhang@blcu.edu.cn; Tel.: +86-010-82303207
† These authors contributed equally to this work.

Abstract: As an indispensable module of computer-aided pronunciation training (CAPT) systems, mispronunciation detection and diagnosis (MDD) techniques have attracted a lot of attention from academia and industry over the past decade. To train robust MDD models, this technique requires massive human-annotated speech recordings which are usually expensive and even hard to acquire. In this study, we propose to use transfer learning to tackle the problem of data scarcity from two aspects. First, from audio modality, we explore the use of the pretrained model wav2vec2.0 for MDD tasks by learning robust general acoustic representation. Second, from text modality, we explore transferring prior texts into MDD by learning associations between acoustic and textual modalities. We propose textual modulation gates that assign more importance to the relevant text information while suppressing irrelevant text information. Moreover, given the transcriptions, we propose an extra contrastive loss to reduce the difference of learning objectives between the phoneme recognition and MDD tasks. Conducting experiments on the L2-Arctic dataset showed that our wav2vec2.0 based models outperformed the conventional methods. The proposed textual modulation gate and contrastive loss further improved the F1-score by more than 2.88% and our best model achieved an F1-score of 61.75%.

Keywords: mispronunciation detection and diagnosis (MDD); computer-aided pronunciation training (CAPT); transfer learning; pretrained model; text modulation gate

Citation: Peng, L.; Gao, Y.; Bao, R; Li, Y.; Zhang J. End-to-End Mispronunciation Detection and Diagnosis Using Transfer Learning. *Appl. Sci.* **2023**, *13*, 6793. https://doi.org/10.3390/app13116793

Academic Editor: Douglas O'Shaughnessy

Received: 3 March 2023
Revised: 16 April 2023
Accepted: 25 April 2023
Published: 2 June 2023

1. Introduction

A computer-aided pronunciation training (CAPT) system facilitates oral learning of language learners by supporting flexible learning patterns and the use of fractional time. As an indispensable module of CAPT systems, mispronunciation detection and diagnosis (MDD) techniques have attracted a lot of attention from academia and industry over the past decade. Similar to the role of oral language teachers, MDD can provide language learners with instant feedback about pronunciation problems, either at segmental or supra-segmental levels, to improve their oral proficiency. Considering the rapidly increasing number of language learners, especially for distant learning and online learning, a high-performance MDD is in great demand to assure the precise diagnosis of pronunciation errors. The present study focuses on phonetic mispronunciations in second-language learning. Moreover, we only consider MDD of constrained speech, that is, the (reference) text/prompt/transcription (In this study, the terms "text", "prompt", and "transcription" refer to the textual content to be learned to pronounce by the speakers and they will be interchangeable) to be uttered by speakers is known to the system.

Previous MDD studies can be roughly grouped into two categories, both of which have fully made use of the transcriptions. The first category is based on confidence measures.

With the help of forced alignment techniques in the automatic speech recognition (ASR) area, researchers usually first align the acoustic frame/phone level segments with given texts. Then they calculate posterior probability as an indicator of whether the pronunciation is correct or not. The most representative methods are goodness of pronunciation (GOP) and its variants that use a posterior probability ratio to evaluate pronunciation and detect errors [1–3]. Although the approaches based on confidence measures can detect incorrect pronunciations, they lack the ability to provide specific diagnostic information, i.e., how the target speech is mispronounced. The second category is based on extended search lattices of speech recognition, among which extended recognition network (ERN) [4] is the most representative approach. ERN analyzes the text first and then incorporates a finite number of phonetic error patterns into the decoding network based on handcrafted or data-driven rules. The recognized phoneme sequence of learners' speech will be compared with the canonical phonetic transcription to derive insertion, deletion, or substitution errors. However, ERN cannot guarantee that all mispronunciations are covered. The MDD performance will degrade when unseen mispronunciations occur. Moreover, building the above multistage systems is complicated and laborious; errors of each stage will accumulate to the final result.

Inspired by the End-to-End (E2E) ASR framework, connectionist temporal classification (CTC)-based methods [5] are recently proposed for MDD and achieve encouraging performances [6,7]. The E2E approaches avoid manually designing phonetic rules and the complex modeling process. In spite of effectiveness, training those E2E MDD models requires large-scale supervised speech corpora covering enough mispronunciation samples. It is worth noting that MDD is a data-scarce task. Acquiring non-native speech data with reliable corresponding annotations by experienced experts is very costly and time-consuming. Moreover, it is more difficult and even impractical to develop such datasets with a large size.

Transfer learning, as a developing sub-field of deep learning, has the potential to alleviate the problem of data scarcity. By pretraining a model from relevant tasks and then using the pretrained model as a feature extraction module for specific downstream tasks, transfer learning has been successfully applied to multiple research domains [8–10]. As a useful technique of transfer learning, self-supervised pretraining (SSP) attempts to first create pseudo-labels from unlabeled data and then learn powerful context representation in a supervised fashion [11,12]. As the most powerful SSP model, wav2vec2.0 has also been successfully shown to be effective in the field of MDD. In [13,14], the researchers investigated MDD by using the well-trained wav2vec2.0 as a model building block and fine-tuning it for MDD tasks.

Instead of using wav2vec2.0 as the backbone of MDD, some studies use wav2vec2.0 as the feature extractor of their systems [14–16]. Although wav2vec2.0 has been applied to MDD tasks, its feasibility under different configurations, especially for ultra-low resource MDD, still remains to be explored.

In addition to the way of transferring knowledge learned from related tasks to downstream tasks, transfer learning also enables us to learn better representations across modalities, e.g., transferring knowledge from a modality with rich or clean data to that with scarce or noisy data [17]. In the field of MDD, there is a line of studies attempting to leverage the prior linguistic information extracted from the given transcriptions for MDD [18–21]. The first category of such studies is to convert the phoneme sequence of given text into textual embedding via a text encoder. Feng et al. [19] feeds phoneme sequences into a sentence encoder and then combines them with audio features via attention. Frame-level cross-entropy loss is calculated with the help of a manually labeled phoneme boundary. Ye et al. [22] proposed to utilize three kinds of information for MDD: acoustic, phonetic, and linguistic (APL) representations for MDD. Among them, the phonetic representations are extracted with pretrained ASR models, and the linguistic representations are extracted from canonical phoneme sequences by a linguistic encoder. To obtain better alignment between the acoustic and linguistic embeddings,

Chen et al. [15] leveraged articulatory features that are derived from the reference text. The second category conducts data augmentation based on the given text to alleviate the data scarcity problem. Facing the imbalance problem between correct and incorrect pronunciation samples in the training data, Fu et al. [20] designed three data augmentation techniques based on the given transcriptions by randomly replacing the phone labels while keeping the corresponding acoustic signals unchanged. Zhang et al. [23] proposed an L2-GEN method that generates mispronounced L2 phoneme sequences that mimic L2 speech and then synthesizes new L2 speech for data augmentation. Despite the effective network design, most previous studies directly incorporated textual features into speech representation via a naive attention mechanism. We contend that textual features contribute very differently when they are assigned to attend different acoustic features. For correct pronunciation, transcription can guide the model to step towards text–audio joint representation for better inference. However, it is difficult to align prior phonemes with acoustic features when mispronunciation occurs and hence limits the potential performance improvement.

Most of the above-mentioned MDD studies are based on the E2E ASR framework. The main idea is to train a mature phoneme recognition model with L1/L2/L1-L2 hybrid datasets, and then obtain results comparing the recognized phoneme sequence with the reference phoneme sequence. Despite the modeling simplicity, the E2E ASR-based methods have a side effect. Because of the inconsistent optimization goals between ASR and MDD tasks, the best model optimized with respect to the recognition error does not necessarily result in the best MDD performance [24]. This is because MDD usually uses an F1-score as the performance metric that considers both the correct and incorrect pronunciations. Most previous models are optimized with a sole phoneme recognition objective directly or implicitly constrained by extra error states towards the correct diagnosis. Such a single recognition loss tries to predict each phoneme equally. To some extent, we hope the system can detect more mispronunciations with little/no sacrifice of performance on correct pronunciations. To tackle the inconsistency problem of optimization goals, some researchers attempt to incorporate the ASR task with auxiliary tasks, e.g., by classifying phone states (i.e., whether the phone is correctly pronounced) [18], classifying accents of learners' L1 speech [25], and estimating word-level mispronunciation probabilities [26]. Due to the difference of learning objectives between phoneme recognition and MDD tasks, previous methods fail to focus on mispronunciations explicitly and, thus, are less effective in detection and diagnosis.

Based on the above introduction and analysis, we can find that training MDD systems faces the problem of data scarcity. This study is motivated by that fact that some previous studies tackled this problem to some extent by using SSP models and incorporating prior text and that it is still worth investigating how to make full use of them. Aiming to further improve the performance of MDD, we propose an E2E mispronunciation MDD method using transfer learning that leverages the pretrained acoustic model and the linguistic knowledge from text. Specifically, the main contributions of this study are as follows:

(1) We explore the use of the pretrained model wav2vec2.0 for MDD tasks, especially under different configurations for low resource MDD;
(2) We propose an effective text–audio gate control module to effectively leverage the linguistic information from text modality. It can enforce the model to align textual information to the most related acoustic regions while ignoring irrelevant parts automatically;
(3) To further unleash the power of prior text, we refine the loss to bridge the learning objective gap between phoneme recognition and MDD by explicitly discriminating the probability of reference and annotation sequences. Conducting experiments on the L2-Arctic dataset confirm the effectiveness of our proposals.

The remainder of this paper is structured as follows. Section 2 introduces the related work wav2vec2.0. Section 3 introduces the proposed methodology for MDD, including the pretrained model based MDD framework, the textual modulation gate, and the contrastive

learning strategy. Section 4 shows the experiments and results. Section 5 concludes this paper and gives outlook for future work.

2. Related Work

Transfer learning [27,28] and domain adaptation [29,30] are both techniques used in machine learning. Transfer learning refers to the process of taking knowledge learned from one task (source task) and applying it to a different but related task (target task), whereas domain adaptation is a specific type of transfer learning where the goal is to adapt a model trained on one domain to perform well on a different but related domain. They are related and sometimes used interchangeably. The present study leverages transferred acoustic knowledge from pretrained models with audio modality and linguistic knowledge from prior text for the MDD task. Moreover, it is also related to multimodal learning [31], which involves training a model on data from multiple modalities to jointly learn representations that capture the relationships between the modalities. In other words, the model learns to integrate information from different modalities to make predictions. To make the title concise and the terminology reflect all related concepts, we use "transfer learning" throughout this article.

2.1. Technical Basis

In speech-related domains, several SSP models have been designed, such as APC [32], wav2vec [33], and wav2vec2.0 [12]. They are used as the building blocks of models for many tasks, e.g., ultra-low resource ASR [34] and emotion recognition [35].

This study uses the pretrained model wav2vec2.0 [12] to obtain the acoustic representations. The architecture of wav2vec2.0 is shown in the left part of Figure 1. It is composed of three modules: a CNN-based acoustic encoder, a transformer-based context network, and a vector quantizer. The encoder module built with seven blocks of temporal convolution layers encodes the raw audio sample $\mathcal{X}$ into latent speech representation $\mathcal{Z}$. The encoder module compresses about a 25 ms region of 16 kHZ audio every 20 ms. After that, the encoded representation $\mathcal{Z}$ will be used to provide inputs for two branches. For the branch of the transformer-based context network, a certain proportion of consecutive time steps of latent representation Z are masked and used to learn the contextual representations $\mathcal{C}$. For the branch of the vector quantizer, vector quantization (VQ) [36] is used to discretize the unmasked Z in continuous space to a finite number of entries $\mathcal{Q}$, which is detailed in the upper-right part of Figure 1. During pretraining, a contrastive predict coding (CPC) [37] criterion is adopted. The outputs of the transformer-based context network and quantized outputs of vector quantizer are used to compute the contrastive loss so that the pretrained model can distinguish the latent representation from a series of distractors sampled from other masked time steps. In the meantime, VQ is expected to learn the underlying speech units that are reported to correspond to phonemes. Alternatively, learning general acoustic representations can be achieved using other pretrained models, e.g., HuBERT [38]. In this study, we chose the wav2vec2.0 model for reasons of its popularity and ease of implementation.

As for transferring linguistic knowledge from prior text for E2E MDD, the common way is to convert sentence transcription to textual embeddings and then incorporate them with attention operation. Specifically, the input of the model is the phoneme sequence converted from prior text. Suppose an N-length phoneme sequence of the input sentence $S = [s_1, ..., s_i, ..., s_N]$ where s_i is the phoneme at the i-th position in the prior phoneme sequence, the audio sample $\mathcal{X}$ and phoneme sequence S can be converted to acoustic and textual representations by an audio encoder and text encoder, respectively, using the following equations:

$$h^Q = \texttt{Audio\_encoder}(\mathcal{X}) \tag{1}$$

$$h^K, h^V = \texttt{Text\_encoder}(S) \tag{2}$$

where h^Q, h^K, and h^V are the query, key, and value, respectively, for the attention operations. Next, a fixed-length context vector can be obtained using the following equation:

$$c = \texttt{Attention}(h^Q, h^K, h^V). \tag{3}$$

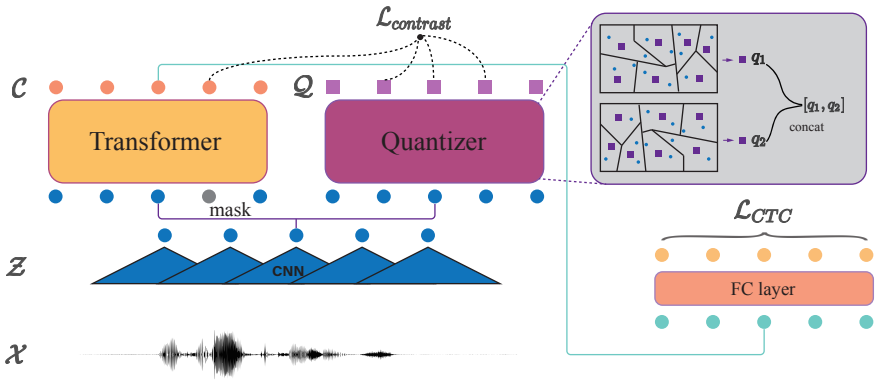

Figure 1. Left: The structure of the pretrained model wav2vec2.0 framework and its corresponding criterion in the pretraining stage. Top-right: The details of vector quantization module. Bottom-right: The use of acoustic representations learned from wav2vec2.0 model for MDD task by adding fully connected (FC) layers.

2.2. Previous Methods for MDD

There exist a couple of studies that transferred acoustic knowledge from audio modality for MDD. In our previous work [39], we created an MDD model by stacking only a fully connected layer on the top of pretrained wav2vec2.0. The experimental results showed that the proposed MDD model trained only with one-third of the data achieved the comparable performance to the conventional methods, suggesting the effectiveness of wav2vec2.0 for learning robust features. In the same period, Wu et al. [13] used a similar way of using wav2vec2.0 for MDD and also showed its effectiveness on another corpus. Instead of using wav2vec2.0 for directly extracting MDD features, Zhang et al. [14] proposed to use wav2vec2.0 and a K-means clustering algorithm to convert the original continuous speech into audio vectors and then into discrete acoustic units. By modifying this unit sequence, they augmented the acoustic data and finally improved the MDD performance.

As for transferring linguistic knowledge from text modality, the researchers usually first aligned the audio with phoneme sequence, and then used the joint representation as input for the MDD module. Fu et al. [20] proposed to concatenate the acoustic representations and the attention results (calculated by Equation (3)) for MDD. Different from the study [19] that relied on manually labeled phoneme boundary, this study used an E2E framework. Moreover, instead of extracting linguistic features as input, Zhang et al. [25] proposed an ASR and alignment unified transformer-based MDD framework where the prior target text was used as the condition for the decoder input. Contrastive learning, a kind of technique that maximizes the intra-class similarity and minimizes the inter-class similarity, has been used extensively over the years in various applications [40,41]. In the filed of SSP, contrastive learning can be used to learn useful representations in a self-supervised manner. Wickstrøm et al. [42] proposed a contrastive learning framework that enabled transfer learning clinical time series by exploiting a data augmentation scheme in which new samples were generated by mixing two data samples with a mixing component. For the task of Chinese spell-checking, Lin et al. [43] proposed reverse contrastive learning which explicitly forced the model to minimize the distance in language representation space between similar sample pairs. In the context of MDD, we can anchor the transcription in order to generate the dissimilarity/similarity.

3. Methodology

In this section, we first present our basic MDD framework by fine-tuning the wav2vec2.0 model. Next, we introduce the enhanced MDD framework that incorporates the textual information of prior text via the text–audio gate control module. Finally, we introduce the proposed contrastive learning to fill the learning objective gap between phoneme recognition and MDD tasks.

3.1. Fine-Tuning Wav2vec2.0 for MDD

Once the wav2vec2.0 model is well pretrained, it can be used as a feature extractor for the downstream task MDD. It is worth mentioning that in [12], only using a fully connected (FC) layer stacked on the wav2vec2.0 encoder obtains the state-of-the-art phone recognition results on the TIMIT corpus. Following this structure, we also investigate the effectiveness of the pretrained model wav2vec2.0 by adding an FC layer on top of it for the MDD task. The basic MDD framework is shown in the bottom-right part of Figure 1, which is in line with that of our previous work [39]. The acoustic representation output from the wav2vec2.0 is fed into the FC layer. During fine-tuning the pretrained model, quantization is disabled and the CTC loss is adopted. The aim of fine-tuning the pretrained wav2vec2.0 is to further update the model parameters for the MDD task in a supervised fashion. The MDD model predicts phone sequence from the raw speech waveform. In this context, the MDD task can be regarded as a variant of the phone recognition task.

3.2. Enhanced MDD by Textual Information

3.2.1. The Framework of Enhanced MDD

Although we have included a powerful pretrained acoustic model into our MDD system, there is still possibly much room for improvement by combining with reference texts. A common approach is to convert the reference text into phonemes and transform each phoneme into a high-dimensional linguistic feature vector via a text encoder. Fu et al. [20] implemented such a MDD framework based on CNN and Bi-LSTM neural networks. As shown in the left part of Figure 2, their framework takes audio signals and phoneme sequences as inputs. The audio encoder built with CNN and Bi-LSTM and the text encoder built with Bi-LSTM convert the audio signal and phoneme sequence to acoustic and linguistic representations, respectively. Given a canonical phoneme sequence with length N, the linguistic representation $\mathbf{H}^{\text{text}} = [\mathbf{h}_1^{\text{text}}, \mathbf{h}_2^{\text{text}}, ..., \mathbf{h}_N^{\text{text}}]$ can be derived by the text encoder. After that, the model improves the aligned representation learning by performing attention for feature aggregation. The acoustic and linguistic representations are concatenated and fed into a linear layer for MDD. This model is referred to as BaselineConventional. We replace the audio encoder and text encoder with the pretrained wav2vec2.0 model and transformer respectively while keeping the same fusion strategy, which is shown in the right part of Figure 2. Moreover, the concatenation of acoustic and textual representations is further fed into another transformer layer. This model is referred to as BaselineConcatenate.

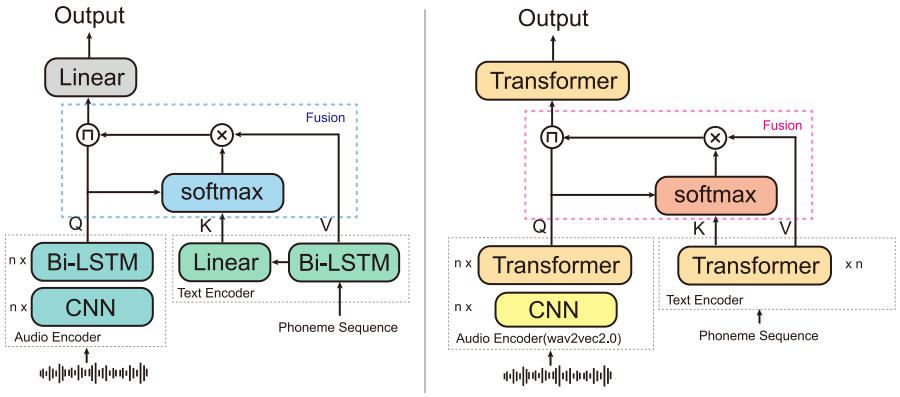

Figure 2. An overview of baselines: (**left**) The BaselineConventional model from [20]. It takes phoneme sequences and the fbank feature as inputs and improves the aligned representation learning by performing attention for feature aggregation. (**right**) Our BaselineConcatenate model where we replace the audio encoder and text encoder of the BaselineConventional model with the more powerful wav2vec2.0 model and transformer while keeping the same fusion strategy. It should be noted that only the audio branch is pretrained.

3.2.2. Textual Modulation Gate

The MDD model shown in the Figure 2 takes the reference text (i.e., the canonical transcription) as the input. However, some phonemes in the canonical transcription of the reference text will be "polluted" when the speakers mispronounce them. In this case, the "polluted" reference text is thus not paired with associated audio features. Therefore, we propose a *textual modulation gate* based on attention fusion. On the textual side, we run an information monitor to filter out texts whose prior knowledge is strong enough to deteriorate the performance. To this end, we design and compare four different textual modulation gates. The first is shown in the bottom-left corner of Figure 3. For $\mathbf{H}^{\text{text}}$ and $\mathbf{H}^{\text{audio}}$, we have:

$$\boldsymbol{\alpha}_{t,n} = \text{sigmoid}(\text{score}(\mathbf{h}_n^{\text{text}}, \mathbf{h}_t^{\text{audio}})) \tag{4}$$

$$\text{score}(\mathbf{h}_n^{\text{text}}, \mathbf{h}_t^{\text{audio}}) = \mathbf{h}_n^{\text{audio}}(\mathbf{h}_t^{\text{text}})^{\mathbf{T}} \tag{5}$$

$$\mathbf{c}_t = \sum_{n=1}^{N} \boldsymbol{\alpha}_{t,n}\mathbf{h}_n^{\text{text}} \tag{6}$$

$$\mathbf{g} = \text{sigmoid}(W \cdot \mathbf{h}_t^{\text{audio}} + U \cdot \mathbf{c}_t + b) \tag{7}$$

$$\mathbf{y}_t = \mathbf{h}_t^{\text{audio}} + \mathbf{g} \odot \mathbf{c}_t \tag{8}$$

where $\odot$ is an element-wise product. We compute attention weights, between textual embedding $\mathbf{h}_n^{\text{text}} \in \mathbf{H}^{\text{text}}$ and acoustic embedding $\mathbf{h}_t^{\text{audio}} \in \mathbf{H}^{\text{audio}}$, which is used for re-weighting the textual representation. Then we choose the implementation of linear projection, summation, and sigmoid activation sequentially to generate the textual gate before feeding them into the transformer layer for CTC prediction. We refer to the formula above as *TextGate*. Furthermore, we further explore three variants of gate modulating. Two of them only differ in the activation function from *TextGate*: one with the softmax function and another with the tanh function (see the *TextGate*σ and *TextGate*ϕ subplots in Figure 3). The third variant uses one gate to control on the textual branch and another to control on the acoustic branch, which is referred to as *DoubleGate*. It should be noted that, except for the BaselineConcatenate model that uses the "concatenation" operation, the rest models use the "add" operation. This design is for the convenience of easily extending the model by stacking more textual modulation gating layers.

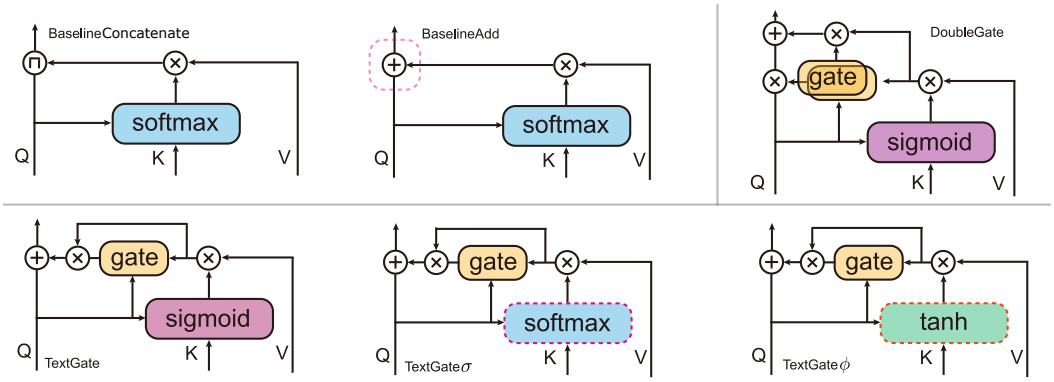

Figure 3. Variants of textual-gate modulating: (BaselineConcatenate) is identical to the right model in Figure 2. (BaselineAdd) uses another popular feature fusion operation "add" instead. (DoubleGate) not only performs control on textual branch, but also monitors acoustic branch with another gate. (TextGate), (TextGateσ) and (TextGateϕ) attempt to look into different activation functions.

3.3. Contrastive Learning

Phoneme recognition aims to infer phonemes from the acoustic signals correctly as much as possible, irrespective of whether we should pay more attention to mispronunciations. During the model training with regard to phoneme recognition, the model optimizer seeks to update parameters and hence improve the recognition performance by detecting more canonicals (i.e., correct pronunciations) in proportion. With the given prior texts, we propose an objective based on contrastive learning to reduce the difference of learning objectives between the phoneme recognition and MDD tasks.

Although we cannot directly construct negative pairs and positive pairs as usual to define the similarity, we introduce a supervised contrastive loss derived from CTC [5]. Addressing the variable length (T) input frames, $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T]$, conditionally independent probability of label sequence:

$$p(\boldsymbol{\pi}|\boldsymbol{X}) = \prod_{t=1}^{T} y_{\pi_t}^t = y_{\pi_1}^1 y_{\pi_2}^2 ... y_{\pi_t}^t ... y_{\pi_T}^T, \forall \boldsymbol{\pi} \in \Phi(L). \tag{9}$$

$$p(\boldsymbol{\pi}^e|\boldsymbol{X}) = \prod_{t=1}^{T} y_{\pi_t^e}^t = y_{\pi_1^e}^1 y_{\pi_2^e}^2 ... y_{\pi_t^e}^t ... y_{\pi_T^e}^T, \forall \boldsymbol{\pi}^e \in \Phi(L^e). \tag{10}$$

where $y_{\pi_t}^t$ denotes the softmax output of label π_t at time t, and $\Phi(\cdot)$ is a map function which can generate all possible intermediate label representations from unmodified label sequence. A modified label sequence $\boldsymbol{\pi}^e$ is made by inserting, deleting, and substituting phones. Suppose there is only one substitution mispronunciation occurring at position t, for the π-π^e pair, $y_{\pi_k}^k = y_{\pi_k^e}^k, \forall k \neq t$. Then we can define the dissimilarity for modified annotation and sequence.

$$D_{\text{contrast}}^{\pi,\pi^e,X} = \ln p(\pi^e|X) - \ln p(\pi|X) = \ln y_{\pi_t^e}^t - \ln y_{\pi_t}^t. \tag{11}$$

We incorporate margin into the dissimilarity and sum up all possible negative pairs. Then our contrastive loss can be expressed as:

$$\mathcal{L}_{\text{contrast}}^{(L,L^e,\boldsymbol{X})} \triangleq \sum_{\substack{\pi,\pi^e \in \\ B^{-1}(L,L^e)}} \max(\ln p(\pi^e|\boldsymbol{X}) - \ln p(\pi|\boldsymbol{X}) + m, 0). \tag{12}$$

In order to train the network, we incorporate the additional contrastive loss $\mathcal{L}_{contrast}$:

$$\mathcal{L} = \mathcal{L}_{CTC} + \mathcal{L}_{\text{contrast}} \tag{13}$$

where $\mathcal{L}_{CTC}$ is a loss for phoneme recognition (CTC computes the probability of all possible intermediate sequences via dynamic programming. Similarly, we implemented the loss as followed approximately. m is set to 16 empirically):

$$\mathcal{L}_{\text{contrast}}^{(L,L^e,\boldsymbol{X})} = \max(\ln p(L^e|\boldsymbol{X}) - \ln p(L|\boldsymbol{X}) + m, 0). \tag{14}$$

4. Experiments and Results

4.1. Speech Corpora

To validate our proposals, we conducted experiments on the publicly available corpus L2-arctic [44] that is commonly used in previous MDD studies. The L2-arctic corpus is an English speech dataset that is designed for research in voice conversion, accent conversion, and mispronunciation detection. It is composed of non-native utterances with mispronunciation. There are 24 non-native speakers (12 males and 12 females) and six mother tongues (L1): Hindi, Korean, Spanish, Arabic, Vietnamese, and Chinese. Moreover, most previous studies incorporated English native data for model training so that the ASR-based MDD models can accurately identify the correctly pronounced parts of the utterances. This is motivated by the fact that the L2-arctic corpus is not only composed of L2 data but also with a small size. Therefore, we also use the TIMIT [45] corpus, containing 6300 utterances produced by 630 speakers, as an additional corpus. Here, the original training subset of TIMIT corpus is used. In order to merge the speech samples of the L2-arctic and TIMIT datasets for training, we re-sample the audio files of the L2-arctic corpus to 16 kHZ with the open-source tool SoX [46]. For the TIMIT corpus, we map its phone set with 61 phonemes to that with 39 units according to the mapping table used in [47] and finally merge it into L2-arctic phone set.

4.2. Experimental Setup

For the well-pretrained wav2vec2.0 models, we used those from fairseq toolkit [48]: wav2vec2.0-BASE, wav2vec2.0-LV60, and wav2vec2.0-XLSR. They use the same model architecture but different numbers of layers (i.e., different numbers of parameters). Accordingly, they are trained with different amounts of data for pretraining. The BASE and LV60 models use 960 h of Librispeech [49] and 53,200 h of LibriVox for pretraining, respectively. For the XLSR model, 56,000 h of speech data, consisting of 53 languages, is used for pretraining.

In this paper, the experiments are conducted in two stages. In the first stage, we compare the feasibility of the pretrained model wav2vec2.0 with conventional methods. We also examine the influence of the amount and type of training data to fine-tune the pretrained wav2vec2.0 on MDD performance. In the second stage, using the best configuration found in the first stage as the basis, we mainly focus on comparing the influence of different textual modulation gates on the MDD performance.

4.2.1. Examination of Pretrained Model and Data Configuration

We build the MDD model by stacking one FC layer on the pretrained wav2vec2.0 model, in which the parameters of the FC layer is randomly initialized. After that, we fine-tune the MDD model using training data of the L2-arctic and TIMIT datasets. To better understand the influence of the amount and type of training samples during fine-tuning, we train MDD models with the following four kinds of training data configurations (Source code and configuration files are available at https://github.com/vocaliodmiku/wav2vec2mdd).

- **Default** For this configuration, we merely use the training data from the L2-arctic corpus. The way of data partition for training and testing is consistent with [19] where the data of six speakers (NJS, TLV, TNI, TXHC, YKWK, ZHAA) are held as the test subset while the rest data of other 18 speakers are merged to build the training subset. Moreover, a development set is created by by randomly selecting 20% sentences from each speaker of the training subset. There is no overlap between the training and developing set.

- **−33%** This configuration is designed to explore the feasibility of the pretrained model for ultra-low resource MDD. To this end, we reduce 33% of training data for each language by randomly excluding six speakers from the *Default* training set.
- **−66%** Similar to the *−33%* configuration, we further reduce 33% of training data for each language by randomly excluding another six speakers from the *−33%* training set. In this case, only one speaker is kept for each of the six languages of non-native speakers.
- **+TIMIT** This configuration is to explore the effectiveness of incorporating the native English data. Here, the original training subset of TIMIT corpus is merged into the *Default* training set.

Table 1 summarizes related duration statistics of these data configurations. We can regard the data configurations of *−33%* and *−66%* as the low-resource and ultra-low-resource MDD.

Table 1. The total duration statistics (in hours) of different data configurations.

	Train	Dev	Test
Default	2.50	0.28	0.88
−33%	1.49	0.37	0.88
−66%	0.73	0.19	0.88
+TIMIT	6.07	0.28	0.88

4.2.2. Examination of Textual Modulation Gate

At this stage, we conducted experiments to examine the effect of incorporating the proposed textual modulation gates on the MDD performance (Source code and configuration files are available at https://github.com/vocaliodmiku/wav2vec2mdd-Text). It should be noted that, at this stage, we use all available training data, i.e., the TIMIT data configuration. However, to save computational time, we adopt the simplest pretrained model wav2vec2.0-BASE as the basic audio encoder. Under this basic acoustic model condition, we compare the MDD performances of a series of textual modulation gates. The dimension of attention and gating mechanism is set to 768. After finding the best configuration for leveraging the prior text by textual modulation gate and contrastive learning, we replace the simplest pretrained model wav2vec2.0-BASE with the most powerful pretrained model wav2vec2.0-XLSR.

The experiments are conducted with PyTorch. All models are trained 142 epochs using the Adam optimizer with an initial learning rate of 5×10^{-5} on one RTX3060 GPU. The audio encoder is frozen in the first 10,000 steps. The other model training configurations are set to default values or consistent with the setting in [20].

4.3. Performance Evaluation

We follow the evaluation metrics of previous studies [50]. For the E2E model, the MDD results can be obtained by comparing the recognized phoneme sequence and canonical phoneme sequence of the reference text sequence after alignment. The mispronunciations are detected when inconsistency between the two phoneme sequences occurs. Accordingly, there are four types of MDD results: true acceptance (TA), false rejection (FR), true rejection, and false acceptance (FA). For correctly pronounced phones, TA means the recognized phoneme sequence is consistent with the canonical phoneme sequence of the reference text whereas FR means inconsistency. For mispronounced phones, TR indicates the mispronunciation has been detected whereas false accept (FA) fails to do it. Furthermore, TR can be divided into correct diagnosis and diagnosis error. In addition, other metrics such as recall (TR/(FA + TR)), precision (TR/(FR + TR)), and the F-1 score ($2 \times$ ((precision $\times$ recall)/(precision + recall))) can be derived using the above statistics.

4.4. Experimental Results

4.4.1. Comparison with Different Amounts and Types of Training Data

Table 2 lists the results of our implemented wav2vec2.0 based MDD models. Comparing the first three models shows that the model LV60 outperforms the model BASE. We speculate these improvements of mispronunciation detection benefit from the robust acoustic representation learned from large amounts of unlabeled data. According to the language transfer theory, second language learners tend to transfer the L1 phonetic phenomenon to second language (L2) learning [51], and some researchers argue that using cross-language training corpus can boost MDD performance [52]. This is supported by the results of the current study. As can be seen from Table 2, the multilingual pretrained model XLSR performs better than the monolingual pretrained model LV60 (59.37% vs. 58.75%).

Table 2. MDD performances of different pretrained models and training data configurations.

Models	Data	Correct Pronunciations		Mispronunciations			F1
		True Accept	False Rejection	False Accept	True Rejection		
					Corroct Diag.	Diag. Error	
wav2vec2.0-BASE	-	94.12%	5.88%	49.53%	65.86%	34.14%	54.28%
wav2vec2-LV60	-	94.01%	5.99%	43.37%	68.08%	31.91%	58.75%
wav2vec2-XLSR	-	94.57%	5.43%	43.95%	65.75%	34.25%	**59.37%**
wav2vec2-XLSR	−33%	94.11%	5.89%	41.23%	69.13%	30.87%	59.27%
wav2vec2-XLSR	−66%	93.35%	6.65%	46.06%	64.67%	35.33%	55.52%
wav2vec2.0-XLSR	+TIMIT	94.30%	5.70%	41.80%	70.72%	29.28%	60.44%

To examine the MDD performance on ultra-low resource conditions, we compared the four MDD models using the same pretrained model XLSR but different training data configurations. As we can see from the last four rows of Table 2, reducing 33% of training data slightly degrades the MDD performance compared with the *default* training data configuration. When another 33% of the training data are reduced (i.e., the *−66%* data configuration), the results deteriorate more. However, considering the relatively small size of the training data (here, only one speaker's data for each language is kept), achieving a F1-score of 55.52% is still acceptable. In Figure 4, the upper and lower subplots depict the confusion matrices for the MDD models trained with the *default* and *−66%* data configurations. The model trained with the *−66%* training data can still retain the ability to detect mispronunciations while identifying correct pronunciations. Despite the difference of overall performances, the MDD models trained with different amounts of training data share similar patterns. This suggests that the acoustic representations extracted from the pretrained wav2vec2.0 model have robustness to unseen data for the MDD task even when the annotated data are low-resource.

Finally, we explore the use English native data (i.e., the L1 data) to facilitate the MDD performance. By introducing the additional training data the TIMIT corpus, the performance of the model trained with the *+TIMIT* data configuration further improves by more than 1% in terms of F1-score (59.37% vs. 60.44%). On the other hand, compared with the *−33%* data configuration, the extra 33% data of the *default* data configuration does not increase the performance too much (59.27% vs. 59.37%). This suggests that the acoustic information conveyed by the correct pronunciations of the L1 data may break the bottleneck of only using L2 data.

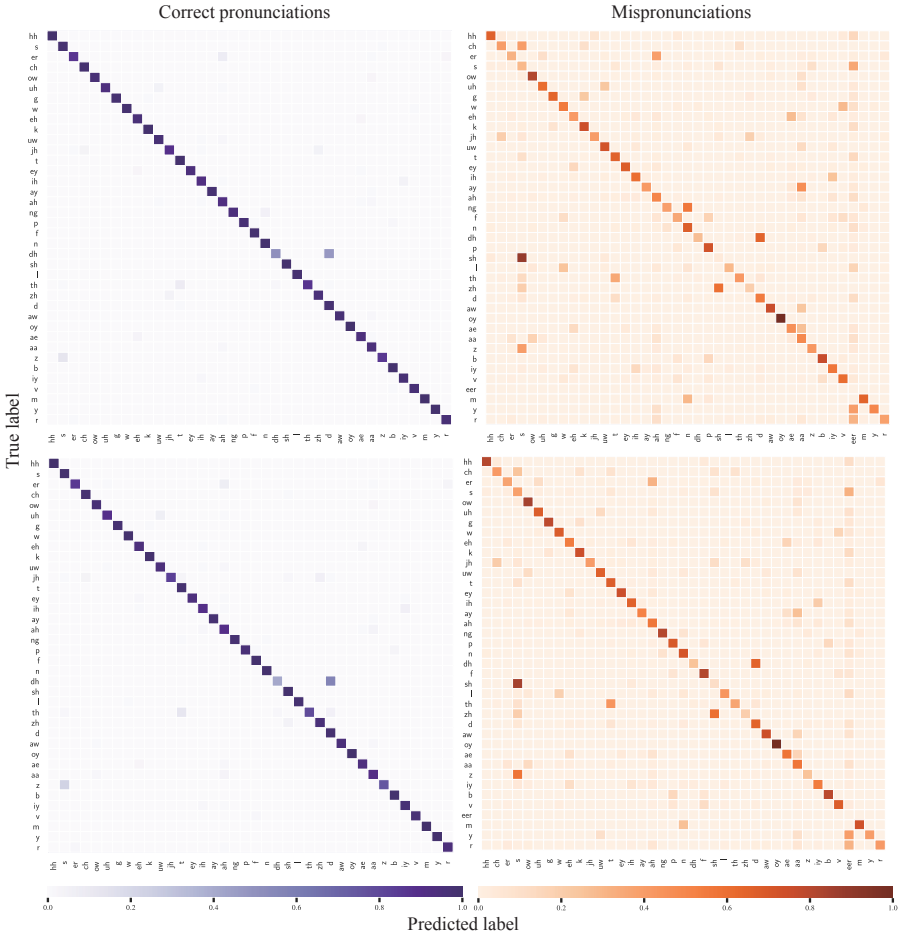

Figure 4. Confusion matrices of the XLSR-based MDD models trained with the *default* (the upper two subplots) and *−66%* (the lower two subplots) data configurations.

4.4.2. Comparison with Conventional Methods

Table 3 lists the MDD results of the proposed methods and other conventional methods. Our pretrained model based methods surpass the well-known GOP based method [53] and CTC based methods [7,20]. As for the training data, in CTC-ATT [7], the researchers also included a small portion of the Librispeech corpus, in addition to the L2-arctic and TIMIT datasets, to build a three-stage model that is complicated. Compared to their work, our XLSR based model trained with less data increase the F1-score by 4.44% (56.02% vs. 60.44%). Even without using the TIMIT training data, our XLSR based model can still achieve an F1-score of 59.37%. As for the model architecture, our model is similar to that in [20] whereas our model replaced their Bi-LSTM with transformer based architecture and made full use of acoustic knowledge transferred from wav2vec2.0 models. A close observation shows that the benefit of using pretrained models mainly comes from the improvements of the precision metric while the recall metric does not improve too much. This means that our proposed models made fewer false-rejection errors while detecting slightly more real mispronunciations.

Table 3. The comparison of MDD performances between our proposed methods with other conventional methods.

Models	Precision (%)	Recall (%)	F1 (%)
GOP [53]	35.42	52.88	42.42
CTC-ATT [7]	46.57	70.28	56.02
CNN-RNN-CTC+VC [20]	56.04	56.12	56.08
XLSR	63.12	56.05	59.37
XLSR(+TIMIT)	62.86	58.20	**60.44**

4.4.3. Using Textual Modulation Gate and Contrastive Learning

Table 4 lists the MDD results of using different textual modulation gates as well as the baseline models. It should be noted that, as mentioned in Section 4.2.2, we used the simplest pretrained model wav2vec2.0-BASE as the audio encoder for most of the MDD systems, and we only used the most powerful wav2vec2.0-XLSR for the best textual modulation gate, i.e., the models *TextGateXLSR* and *TextGateXLSRContrast*.

Table 4. Comprehensive performance comparison between different textual modulation gates.

Models	Correct Pronunciations		Mispronunciations			F1
	True Accept	False Rejection	False Accept	True Rejection		
				Corroct Diag.	Diag. Error	
BaselineConventional	92.65%	7.35%	43.88%	74.96%	25.04%	56.08%
BaselineConcatenate	93.68%	6.31%	42.87%	68.52%	31.48%	58.87%
BaselineAdd	94.15%	5.85%	45.36%	63.21%	36.79%	57.51%
DoubleGate	94.59%	5.41%	44.00%	68.68%	31.32%	59.34%
TextGate	94.50%	5.50%	42.52%	68.26%	31.74%	60.27%
TextGateσ	94.29%	5.71%	41.89%	69.86%	30.14%	60.34%
TextGateϕ	94.53%	5.47%	46.33%	64.22%	35.78%	57.48%
TextGateσ *R* *	95.07%	4.93%	47.66%	63.62%	36.38%	57.47%
TextGateXLSR	94.94%	5.06%	43.72%	68.79%	31.21%	60.23%
TextGateContrast	93.72%	6.28%	40.43%	69.77%	30.23%	60.32%
TextGateXLSRContrast	93.81%	6.19%	38.62%	71.08%	28.92%	**61.75%**

* Reversed version of TextGate (maybe we can call it AudioGate); g is used to multiple h_t^Q instead of c_t.

In the stage of examining the benefit of leveraging the prior text, we used three baseline models. The BaselineConventional and BaselineConcatenate models correspond to left and right parts of Figure 2, respectively. The BaselineAdd model only differs in the fusion style from the BaselineConcatenate model ("add" versus "concatenate"). Under fair architecture and training settings, our baselines based on the pretrained wav2vec2.0 model surpass the BaselineConventional [20], suggesting the audio encoder wav2vec2.0 in our baselines provides more powerful representations. Moreover, the "concatenate" operation for information fusion is better than the "add" operation.

As expected, the proposed approaches *DoubleGate* and *TextGate* outperform the BaselineConcatenate/BaselineAdd method by +0.5%/1.8% and +1.4%/2.7%, respectively. When leveraging linguistic knowledge, most previous studies directly incorporated textual features into speech representation via a naive attention operation. They did not consider the fact that textual features contribute very differently when they are assigned to attend different acoustic features. The text–audio joint representation should be carefully handled especially when mispronunciation occurs. The gating mechanism can be used to modulate the information flowed from text modality to audio modality. The textual modulation gate proposed in this study successfully plays the role of validating information that comes from texts. Figure 5 shows the attention weights output by the *TextGate* and *BaselineConcatenate* model. Since the textual modulation gate can take responsibility for turning on/off textual information flow, attention patterns look neat and natural, whereas audio–text correlation maps for the model without a gate would be chaotic.

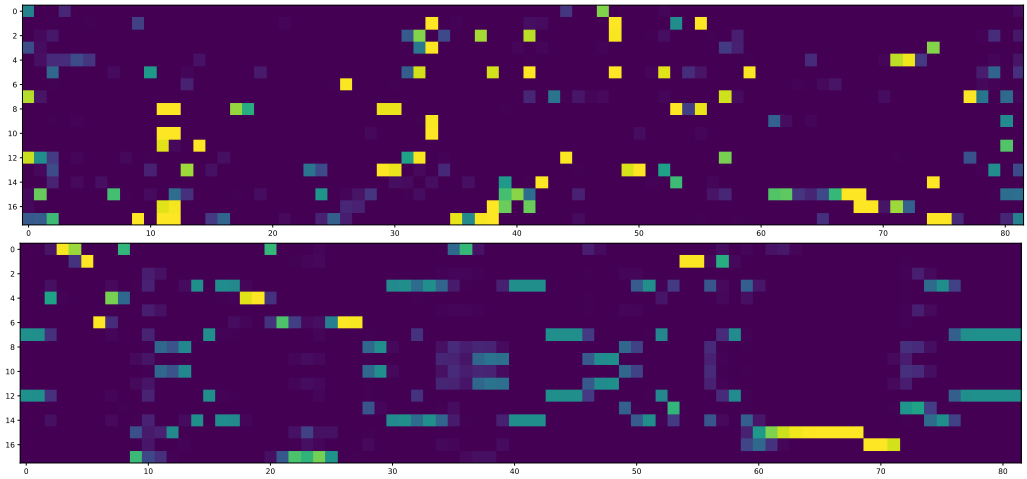

Figure 5. Learned correlation maps of BaselineConcatenate (**upper**) and TextGate (**lower**).

Within the TextGate series, the difference in performance between using sigmoid and softmax activation functions is marginal. However, using tanh activation function obtains a much lower performance, and we also notice that the "AudioGate", performing control on the audio branch, reports poor performance. These results suggest that even incorporating the extra reference text, the fusion framework needs to be carefully designed.

Based on our experimental results, we also found that better phoneme recognition model implementations cannot always report better results in the context of MDD, which is in line with the result reported in [24]. Figure 6 lists phone error rate (PER) and F1-score of different MDD models. As the PER decreases, the F1-score trend is hard to conclude. The failure is due to the mismatch of learning objectives between MDD and phoneme recognition. Therefore, we further explored the benefit of introducing contrastive learning. Moreover, we tried to utilize a more powerful acoustic model (XLSR) for further improvement. We integrated the proposed *TextGate* into *XLSR* model in which the contrastive loss was used. *TextGateXLSRContrast* obtains a performance gain of 1.5% F1-score compared to the model *TextGateContrast* and achieves the best performance with an F1-score of 61.75% (see the bottom two rows of Table 4), suggesting the effectiveness of our models by introducing textual modulation gates with a contrastive loss. Moreover, *TextGateXLSRContrast* reports the lowest false accept rate, which corresponds to our discussion. Given the prior text, the model learns more discriminative features about the reference and annotation phonemes which are hard to distinguish, naturally making the prior transcripts more informative.

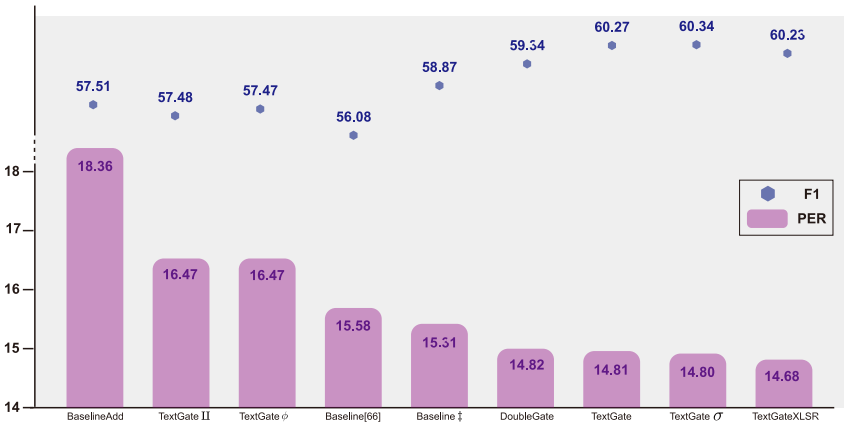

Figure 6. Performances of MDD models with different textual modulation gates in terms of F1-score and phone error rate (PER).

5. Conclusions and Future Work

Training MDD models for CAPT faces the problem of data scarcity. In this study, we proposed to use transfer learning to tackle this problem. First, we explored the effectiveness of the pretrained acoustic model wav2vec2.0 for MDD tasks, especially for ultra-low resource MDD. Moreover, we proposed textual modulation gates that assign more importance to the relevant text information while suppressing irrelevant text information for MDD. Furthermore, given the transcriptions, we proposed an extra contrastive loss to reduce the difference of learning objectives between the phoneme recognition and MDD tasks. By conducting a series of experiments, we have verified the wav2vec2.0 model pretrained with unlabeled data in a self-supervised fashion can provide robust general acoustic representations for the MDD task. Even when the annotated L2 training data are low-resource, the model can still achieve promising performance. Moreover, the proposed textual modulation gate that explicitly incorporates prior transcription in model training effectively learns a better refined text–audio representation for MDD. Our best wav2vec2.0 based model improves the baseline models by +4.32% in absolute F1-score on L2-arctic dataset. By using the best textual modulation gate and contrastive learning, our best model further improve F1-score over the corresponding baseline methods by +2.88%, achieving the highest F1-score of 61.75%. In future work, we will investigate extracting more information from transcriptions, such as transferring phonetic knowledge to constrain the text–audio attention matrix and optimize the learning object toward MDD.

Author Contributions: Methodology, L.P., Y.G.; writing—original draft preparation, L.P., Y.G. and R.B.; writing—review and editing, L.P., Y.G. and Y.L.; resources, Y.G. and Y.L.; supervision, J.Z.; project administration, J.Z.; funding acquisition, Y.L. and J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Fundamental Research Funds for the Central Universities (Grant Number: 2023RC13), the National Natural Science Foundation of China (NSFC) (Grant Number: 62271083), the Advanced Innovation Center for Language Resource and Intelligence (Grant Number: KYR17005), and the Wutong Innovation Platform of Beijing Language and Culture University (Grant Number: 19PT04).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The publicly available L2-arctic corpus used in this study can be found here: L2-ARCTIC-V2.0.

Acknowledgments: We would like to express our sincere gratitude to Kaiqi Fu, Binghuai Lin, and Dengfeng Ke for their valuable comments and suggestions during our preparation for the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Witt, S.M.; Young, S.J. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Commun.* **2000**, *30*, 95–108. [CrossRef]
2. Hu, W.; Qian, Y.; Soong, F.K. A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL). In Proceedings of the Interspeech, Lyon, France, 25–29 August 2013; pp. 1886–1890.
3. Zheng, J.; Huang, C.; Chu, M.; Soong, F.K.; Ye, W.P. Generalized segment posterior probability for automatic Mandarin pronunciation evaluation. In Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing—ICASSP'07, Honolulu, HI, USA, 15–20 April 2007; Volume 4, pp. 201–204.
4. Harrison, A.M.; Lo, W.K.; Qian, X.j.; Meng, H. Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training. In Proceedings of the International Workshop on Speech and Language Technology in Education, Warwickshire, UK, 3–5 September 2009.
5. Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 369–376.

6. Leung, W.K.; Liu, X.; Meng, H. CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 8132–8136.
7. Yan, B.C.; Wu, M.C.; Hung, H.T.; Chen, B. An End-to-End Mispronunciation Detection System for L2 English Speech Leveraging Novel Anti-Phone Modeling. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; pp. 3032–3036.
8. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
9. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018impring.pdf (accessed on 5 April 2023).
10. Erhan, D.; Courville, A.; Bengio, Y.; Vincent, P. Why does unsupervised pre-training help deep learning? In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics JMLR Workshop and Conference Proceedings, Sardinia, Italy, 13–15 May 2010; pp. 201–208.
11. Doersch, C.; Zisserman, A. Multi-task self-supervised visual learning. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2051–2060.
12. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12449–12460.
13. Wu, M.; Li, K.; Leung, W.K.; Meng, H. Transformer Based End-to-End Mispronunciation Detection and Diagnosis. In Proceedings of the Interspeech, Brno, Czech Republic, 27 May–1 April 2021; pp. 3954–3958.
14. Zhang, Z.; Wang, Y.; Yang, J. End-to-end Mispronunciation Detection with Simulated Error Distance. In Proceedings of the Interspeech 2022, Incheon, Republich of Korea, 18–22 September 2022; pp. 4327–4331.
15. Chen, Q.; Lin, B.; Xie, Y. An Alignment Method Leveraging Articulatory Features for Mispronunciation Detection and Diagnosis in L2 English. In Proceedings of the Interspeech 2022, Incheon, Republich of Korea, 18–22 September 2022; pp. 4342–4346.
16. Lin, B.; Wang, L. Phoneme Mispronunciation Detection By Jointly Learning To Align. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–25 May 2022; pp. 6822–6826.
17. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 423–443. [CrossRef] [PubMed]
18. Zheng, N.; Deng, L.; Huang, W.; Yeung, Y.T.; Xu, B.; Guo, Y.; Wang, Y.; Jiang, X.; Liu, Q. CCA-MDD: A Coupled Cross-Attention based Framework for Streaming Mispronunciation detection and diagnosis. *arXiv* **2021**, arXiv:2111.08191.
19. Feng, Y.; Fu, G.; Chen, Q.; Chen, K. SED-MDD: Towards sentence dependent end-to-end mispronunciation detection and diagnosis. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 4–8 May 2020; pp. 3492–3496.
20. Fu, K.; Lin, J.; Ke, D.; Xie, Y.; Zhang, J.; Lin, B. A Full Text-Dependent End to End Mispronunciation Detection and Diagnosis with Easy Data Augmentation Techniques. *arXiv* **2021**, arXiv:2104.08428.
21. Jiang, S.W.F.; Yan, B.C.; Lo, T.H.; Chao, F.A.; Chen, B. Towards Robust Mispronunciation Detection and Diagnosis for L2 English Learners with Accent-Modulating Methods. *arXiv* **2021**, arXiv:2108.11627.
22. Ye, W.; Mao, S.; Soong, F.; Wu, W.; Xia, Y.; Tien, J.; Wu, Z. An Approach to Mispronunciation Detection and Diagnosis with Acoustic, Phonetic and Linguistic (APL) Embeddings. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 6827–6831.
23. Zhang, D.; Ganesan, A.; Campbell, S.; Korzekwa, D. L2-GEN: A Neural Phoneme Paraphrasing Approach to L2 Speech Synthesis for Mispronunciation Diagnosis. In Proceedings of the Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Republic of Korea, 18–22 September 2022; Ko, H., Hansen, J.H.L., Eds.; ISCA: Lyon, France, 2022; pp. 4317–4321.
24. Zhang, L.; Zhao, Z.; Ma, C.; Shan, L.; Sun, H.; Jiang, L.; Deng, S.; Gao, C. End-to-end automatic pronunciation error detection based on improved hybrid ctc/attention architecture. *Sensors* **2020**, *20*, 1809. [CrossRef]
25. Zhang, Z.; Wang, Y.; Yang, J. Text-conditioned transformer for automatic pronunciation error detection. *Speech Commun.* **2021**, *130*, 55–63. [CrossRef]
26. Korzekwa, D.; Lorenzo-Trueba, J.; Drugman, T.; Calamaro, S.; Kostek, B. Weakly-supervised word-level pronunciation error detection in non-native English speech. *arXiv* **2021**, arXiv:2106.03494.
27. West, J.; Ventura, D.; Warnick, S. *Spring Research Presentation: A Theoretical Foundation for Inductive Transfer*; Brigham Young University, College of Physical and Mathematical Sciences: Provo, UT, USA, 2007; Volume 1.
28. Lin, Y.P.; Jung, T.P. Improving EEG-based emotion classification using conditional transfer learning. *Front. Hum. Neurosci.* **2017**, *11*, 334. [CrossRef] [PubMed]
29. Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; Vaughan, J.W. A theory of learning from different domains. *Mach. Learn.* **2010**, *79*, 151–175. [CrossRef]
30. Redko, I.; Morvant, E.; Habrard, A.; Sebban, M.; Bennani, Y. *Advances in Domain Adaptation Theory*; Elsevier: Amsterdam, The Netherlands, 2019.
31. Ramachandram, D.; Taylor, G.W. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Process. Mag.* **2017**, *34*, 96–108. [CrossRef]

32. Chung, Y.A.; Hsu, W.N.; Tang, H.; Glass, J.R. An Unsupervised Autoregressive Model for Speech Representation Learning. In Proceedings of the INTERSPEECH, Graz, Austria, 15–19 September 2019; pp. 146–150.
33. Schneider, S.; Baevski, A.; Collobert, R.; Auli, M. wav2vec: Unsupervised Pre-Training for Speech Recognition. In Proceedings of the INTERSPEECH, Graz, Austria, 15–19 September 2019; pp. 3465–3469.
34. Yi, C.; Wang, J.; Cheng, N.; Zhou, S.; Xu, B. Applying wav2vec2. 0 to Speech Recognition in various low-resource languages. *arXiv* **2020**, arXiv:2012.12121.
35. Sharma, M. Multi-lingual multi-task speech emotion recognition using wav2vec 2.0. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 4–10 June 2022; pp. 6907–6911.
36. Van den Oord, A.; Vinyals, O.; Kavukcuoglu, K. Neural Discrete Representation Learning. In Proceedings of the NIPS, Long Beach, CA, USA, 4–9 December 2017.
37. Oord, A.v.d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748x.
38. Hsu, W.N.; Bolte, B.; Tsai, Y.H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3451–3460. [CrossRef]
39. Peng, L.; Fu, K.; Lin, B.; Ke, D.; Zhang, J. A Study on Fine-Tuning wav2vec2. 0 Model for the Task of Mispronunciation Detection and Diagnosis. In Proceedings of the Interspeech, Brno, Czech Republic, 27 May–1 April 2021; pp. 4448–4452.
40. Jaiswal, A.; Babu, A.R.; Zadeh, M.Z.; Banerjee, D.; Makedon, F. A survey on contrastive self-supervised learning. *Technologies* **2020**, *9*, 2. [CrossRef]
41. Le-Khac, P.H.; Healy, G.; Smeaton, A.F. Contrastive representation learning: A framework and review. *IEEE Access* **2020**, *8*, 193907–193934. [CrossRef]
42. Wickstrøm, K.; Kampffmeyer, M.; Mikalsen, K.Ø.; Jenssen, R. Mixing up contrastive learning: Self-supervised representation learning for time series. *Pattern Recognit. Lett.* **2022**, *155*, 54–61. [CrossRef]
43. Lin, N.; Fu, S.; Lin, X.; Jiang, S.; Yang, A. A Chinese Spelling Check Framework Based on Reverse Contrastive Learning. *arXiv* **2022**, arXiv:2210.13823.
44. Zhao, G.; Sonsaat, S.; Silpachai, A.O.; Lucic, I.; Chukharev-Hudilainen, E.; Levis, J.; Gutierrez-Osuna, R. L2-ARCTIC: A non-native English speech corpus. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 2783–2787.
45. Garofolo, J.S.; Lamel, L.F.; Fisher, W.M.; Fiscus, J.G.; Pallett, D.S. *DARPA TIMIT Acoustic-Phonetic Continous Speech Corpus CD-ROM. NIST Speech Disc 1-1.1*; NASA STI/Recon Technical Report n; NIST Publications: Washington, DC, USA, 1993; Volume 93, p. 27403.
46. SoX. Audio Manipulation Tool. Available online: http://sox.sourceforge.net/ (accessed on 15 March 2021).
47. Lee, K.F.; Hon, H.W. Speaker-independent phone recognition using hidden Markov models. *IEEE Trans. Acoust. Speech Signal Process.* **1989**, *37*, 1641–1648. [CrossRef]
48. Ott, M.; Edunov, S.; Baevski, A.; Fan, A.; Gross, S.; Ng, N.; Grangier, D.; Auli, M. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In Proceedings of the NAACL-HLT (Demonstrations), Minneapolis, MN, USA, 2–7 June 2019.
49. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An asr corpus based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; pp. 5206–5210.
50. Li, K.; Qian, X.; Meng, H. Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *25*, 193–207. [CrossRef]
51. Chang, C.B.; Mishler, A. Evidence for language transfer leading to a perceptual advantage for non-native listeners. *J. Acoust. Soc. Am.* **2012**, *132*, 2700–2710. [CrossRef] [PubMed]
52. Duan, R.; Kawahara, T.; Dantsuji, M.; Nanjo, H. Cross-lingual transfer learning of non-native acoustic modeling for pronunciation error detection and diagnosis. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *28*, 391–401. [CrossRef]
53. Yan, B.C.; Chen, B. End-to-End Mispronunciation Detection and Diagnosis From Raw Waveforms. *arXiv* **2021**, arXiv:2103.03023.