A Wav2Vec2 and Multi-Head Attention Based Framework for Pronunciation Error Detection in Tibetan Mandarin Learners

Zhenye Gan [1, a], Ming Wang [1, b], and Le Wei [1, c]

[1] College of Physics and Electronic Engineering,Northwest Normal University, Harvard University, Cambridge, Lanzhou 730070, China.

[a] ganzy@nwnu.edu.cn, [b] wm00084@163.com, [c] wlljy@nwnu.edu.cn

Abstract. Aiming at the systematic pronunciation bias problem of Tibetan native speakers when learning Mandarin Chinese, this paper proposes an end-to-end detection method based on the self-supervised learning Wav2Vec 2.0 model fusing the multi-head self-attention mechanism (MHA) with a CTC decoder. The model is fine-tuned to adapt to the pronunciation characteristics of Tibetan, and the MHA is utilized to enhance the ability of capturing long-distance dependent features. Experiments on a self-constructed Mandarin pronunciation bias dataset of Tibetan students show that the proposed model significantly outperforms the traditional ASR model and the baseline Wav2Vec2-CTC system in terms of detection accuracy (DAR) and F1 scores, which validates its effectiveness in low-resource speech learning scenarios.

Keywords: Tibetan Mandarin Learners, MDD,self-supervised learning.

1. Introduction

In this work, we address the task of Chinese pronunciation bias detection for Tibetan native speakers by proposing an end-to-end framework based on Wav2Vec 2.0. Due to the phonological differences between Tibetan and Mandarin and the limitations of traditional handcrafted features [1], we build a phoneme-level annotated corpus and introduce a detection model enhanced with Multihead Attention Mechanism (MHA). The system fine-tunes a pre-trained Wav2Vec 2.0 model with CTC loss [13], allowing better temporal alignment and feature extraction [2]. Experiments show that this approach performs well in low-resource scenarios and effectively captures fine-grained pronunciation deviations. The main contributions include: (1) constructing and annotating a Mandarin dataset of Tibetan speakers to address low-resource constraints; (2) proposing an MHA-enhanced Wav2Vec2-CTC system; and (3) evaluating performance using multi-level metrics to validate its effectiveness.

2. Related Work

In recent years, automatic speech recognition (ASR) has made significant progress, but it still faces challenges in dealing with accents, dialects, and non-native speaker pronunciation [3]. Improving the system's robustness to multilingual, accented, and noisy speech, and developing personalized adaptive technologies have become research priorities [4][5]. Traditional manual feature extraction methods, such as MFCC, have limited representation capabilities and rely on large amounts of labeled data, making them difficult to adapt to low-resource scenarios. Current mainstream models widely adopt LSTM, CNN, and Transformer self-attention mechanisms, combined with CTC loss to achieve non-aligned sequence modeling [6].

Self-supervised learning (SSL) has driven the development of low-resource speech recognition. Wav2Vec 2.0 achieves excellent performance in low-resource tasks through unsupervised pre-training plus fine-tuning with a small amount of labeled data [7], but it still faces transfer and adaptation issues in detecting minority pronunciation biases [8]. Therefore, end-to-end modeling based on Wav2Vec 2 and the CTC structure has emerged as an effective solution [9]. Compared to traditional ASR systems that rely on multi-module architectures and forced alignment (e.g., HMM-GMM) [10], CNN/RNN models do not require manual alignment, thereby enhancing the efficiency and accuracy of MDD [11].

MDD aims to detect pronunciation errors in non-native speakers, and end-to-end methods can provide efficient feedback. Existing research, such as the CNN-RNN-CTC architecture proposed by Leung et al., combines CTC with temporal modeling to improve robustness [12]; CTC has also become the mainstream training scheme for MDD due to its adaptability to non-aligned data [6]. Additionally, the CTC/Attention hybrid model accelerates convergence through multi-task learning, improving detection performance [14]. Zhang et al. further combined fundamental frequency features to enhance Mandarin Chinese evaluation performance [15]. Faced with the challenges of labeling low-resource languages like Tibetan, researchers have introduced information fusion strategies to reduce data dependency [16]. In summary, models combining end-to-end and SSL can effectively reduce MDD annotation costs and demonstrate stronger generalization capabilities in cross-language scenarios.

3. Method

3.1 MDD System

In this study, an ASR-based pronunciation bias detection framework is constructed. After the user's speech input, the system converts it into a sequence of phonemes and compares it with a standard template to recognize pronunciation deviations, and finally outputs a diagnostic report (MDD result) with error types, locations, and suggestions for improvement. The system provides accurate pronunciation diagnosis and feedback, and is suitable for scenarios such as speech teaching and language assessment,as shown in Fig. 1.

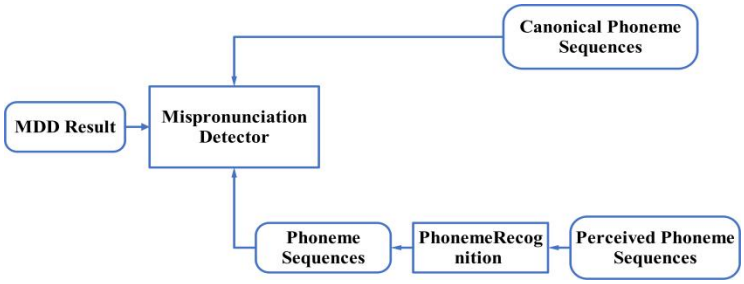

Fig. 1 Architecture of asr-based MDD system

3.2 Wav2Vec 2.0 pre-training and fine tuning

Wav2Vec2 is a self-supervised speech model that learns contextual representations from raw speech through contrastive learning [8]. Its architecture consists of a CNN encoder, a Transformer network, and a quantization module (e.g,Fig.2). The model encodes raw speech into latent representations Z = {zt}, uses the Transformer to learn contextual representations C = {ct} from partially masked segments, and quantizes the unmasked portions using a codebook to form positive-negative sample pairs, which are used to compute the pre-training loss Lpre. After pre-training, the model removes the quantization module, adds a linear layer at the Transformer output to map to the phoneme label space, and fine-tunes using CTC loss to achieve end-to-end phoneme recognition [1]. To enhance robustness, the system introduces speech rate perturbations of 95%, 100%, and 105% as augmentation strategies, effectively improving the model's adaptability to speech rate changes. The overall system can directly predict phoneme sequences from raw speech, achieving a structurally simple and accurate end-to-end MDD (Figure 2(b)).

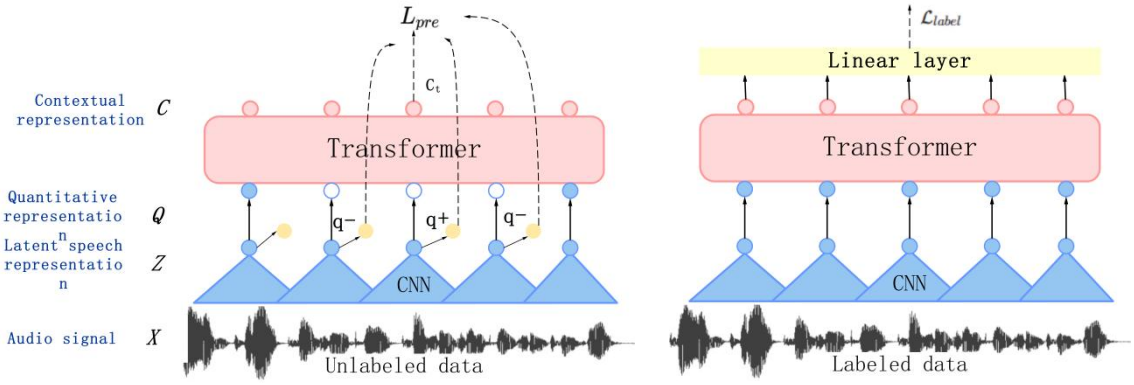

Fig. 2 (a) Wav2vec 2.0 pre-trained model and (b) fine-tuned model

3.3 Multi-head Self-Attention

This study innovatively introduces a multi-head attention mechanism (MHA) between the Wav2Vec2 feature extractor and the CTC decoding layer, which breaks through the limitations of traditional feed-forward neural networks. The design transforms 768-dimensional speech features into more discriminative representations by paying parallel attention to the multidimensional contextual information of the input sequences: on the one hand, it efficiently models the long-distance dependencies in the speech signals, and on the other hand, it accurately captures the synergistic features of articulatory biases. Experiments show that this architecture is particularly suitable for the task of non-native speaker pronunciation recognition, and significantly improves the model's ability to parse the Chinese pronunciation features of Tibetan native speakers compared to the baselinesystem(e.g, Fig. 3).

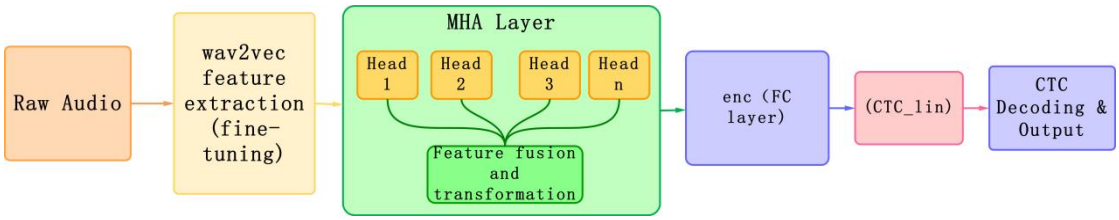

Fig. 3 Architecture diagram based on self-supervised learning and MHA

4. Experiments And Results

4.1 Datasets

This study was conducted based on a self-constructed Mandarin pronunciation deviation (MDD) dataset for Tibetan native speakers, focusing on typical pronunciation errors caused by Tibetan dialect interference. In order to improve the detection effect, Aishell1 and THCHS30 standard Mandarin corpus were integrated to construct a comparison training set containing bias and standard pronunciation. The data covers 16 kHz Mandarin speech recorded in a quiet environment by Tibetan learners aged 18-28, and the types of bias such as phoneme substitution, addition, deletion, and intonation are labeled using the Praat tool to cover multiple types of life scenarios, which provides high-quality low-resource data support for the MDD study (e.g., Tables 1 and 2).

Table 1 Standard Mandarin Dateset

Datasets	Spkrs	Numbers
Aishell1	180	84000
Thchs30	30	10000

Table 2 Mispronunciations dataset

Elements	Numbers
Sentences	2000
Phone	45,780

4.2 Experimental Settings

The experiments were fine-tuned and evaluated on the Mandarin dataset and the Tibetan native speaker Mandarin dataset. The hardware is a server equipped with NVIDIA RTX A6000 graphics card (64GB) and the software is based on PyTorch 1.11 (CUDA 12.4, Python 3.8) framework. The input audio sampling rate is 16kHz and the training batch size is set to 32 to ensure stability and efficiency.

4.3 Evaluation indicators

The evaluation metrics for MDD tasks are primarily used to assess system performance and encompass four categories: True Acceptance (TA), False Rejection (FR), True Rejection (TR), and False Acceptance (FA). In this experiment, the FAR, FRR, DAR, and PER metrics are used to evaluate model performance. Precision (P), recall (R), and F1 scores can be calculated using TR, FR, and FA: P = TR/(FR + TR); FR = TR/(FA + TR); F1 = 2pr/(p + r). The F1 score measures the MDD model's overall false reading detection capability.

4.4 Result

In this paper, we introduce the SSL pre-trained model as a feature extractor and verify its effectiveness in the MDD task. Compared with the traditional methods, the SSL model significantly reduces the dependence on a large amount of labeled data and reduces the training cost, while the FAR significantly decreases and the DAR is close to the optimum. The experimental results show that the improved model outperforms the baseline across the board in terms of F1, Recall, and FAR, with a 1.6% improvement in F1 and a 0.05% decrease in PER.F1 and Diagnostic Accuracy reflect the MDD performance, while PER measures the phoneme recognition accuracy. Despite the significant improvement in F1, the limited improvement in PER suggests that there is still room for optimization of phoneme-level error correction in Wav2Vec2-CTC (MHA)(e.g., Tables 3 and 4).

Table 3 The MDD performance of the proposed method is compared with traditional methods in the laboratory

Model	FAR	FRR	DAR	Recall
GOP-based				
DNN-HMM-FBank	42.66%	44.31%	48.95%	57.34%
ASR model：				
CNN-GRU-CTC	23.13%	7.15%	87.75%	76.87%
DFSMN-CTC	20.32%	6.51%	88.73%	79.68%
SSL model（ours）：				
Wav2vec2-CTC（baseline）	21.52%	11.58%	87.45%	78.47%
Wav2vec2-CTC(MHA)	21.16%	10.85%	88.15%	78.84%

Table 4 Performance comparison between Wav2vec2-CTC (baseline) and Wav2vec2-CTC (MHA)

Model	FAR	FRR	DAR	Recall	F1	PER
Wav2vec2-CTC（baseline）	21.52%	11.58%	87.45%	78.47%	54.95%	10.86%
Wav2vec2-CTC(MHA)	21.16%	10.85%	88.15%	78.84%	56.45%	10.79%

4.5 Ablation Experimen

The experiment compared the performance of 2/4/8-head attention mechanisms and found that 4-head attention (4d) had the best overall performance in pronunciation error detection: key indicators such as FAR (21.16%), FRR (10.85%), DAR (88.15%), Recall (78.84%), and F1 (56.45%) were all superior(e.g., Tables 5).

Table 5 The impact of different attention heads on the model

Heads	FAR	FRR	DAR	Recall	P	F1	PER
2	21.49%	11.03%	87.99%	78.51%	43.45%	55.94%	10.86%
4	21.16%	**10.85%**	**88.15%**	**78.84%**	**43.96%**	**56.45%**	10.79%
8	**21.60%**	11.06%	87.92%	78.40%	43.36%	55.83%	**9.81%**

5. Summary

This study validated the effectiveness of Wav2Vec 2.0 combined with a multi-head attention mechanism in detecting Mandarin pronunciation errors among Tibetan native speakers. Experimental results show that the model outperforms the baseline in key metrics such as FAR, FRR, DAR, and F1. Additionally, the multi-head attention mechanism effectively enhances the model's ability to capture temporal features, enabling it to demonstrate significant advantages in detecting pronunciation errors related to initial consonants, vowels, and tones. Through ablation experiments, it was validated that using a 4-head attention mechanism achieves optimal model performance. Future research will focus on three directions: (1) expanding the dataset and adopting multi-label strategies to enhance generalization capabilities; (2) exploring multi-task joint learning for pronunciation detection, correction, and sentiment analysis; (3) developing a real-time feedback-based personalized training system. These improvements will drive the development of a more comprehensive pronunciation teaching platform, promoting ethnic language exchange and reforms in Chinese language education.This approach could also be extended to other minority languages, supporting broader second language pronunciation learning.

References

[1] Lu K H, Chen K Y. A context-aware knowledge transferring strategy for CTC-based ASR[C]//2022 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2023: 60-67..

[2] Cordonnier J B, Loukas A, Jaggi M. Multi-head attention: Collaborate instead of concatenate[J]. arXiv preprint arXiv:2006.16362, 2020.

[3] Meng W, Yolwas N. A study of speech recognition for Kazakh based on unsupervised pre-training[J]. Sensors, 2023, 23(2): 870.

[4] Hsu J Y, Chen Y J, Lee H. Meta learning for end-to-end low-resource speech recognition[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 7844-7848.

[5] Zhu, Cuicui, et al. "Pronunciation error detection model based on feature fusion." Speech Communication 156 (2024): 103009.

[6] Lee J, Watanabe S. Intermediate loss regularization for ctc-based speech recognition[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6224-6228.

[7] L. Peng, K. Fu, B. Lin, D. Ke, and J. Zhan, “A Study on Fine-Tuning wav2vec2.0 Model for the Task of Mispronunciation De-tection and Diagnosis,”in Proc.Interspeech 2021, 2021, pp.4448–4452.

[8] Chen Q, Lin B, Xie Y. An Alignment Method Leveraging Articulatory Features for Mispronunciation Detection and Diagnosis in L2 English[C]//INTERSPEECH. 2022: 4342-4346.

[9] Peng, Linkai, et al. "End-to-End Mispronunciation Detection and Diagnosis Using Transfer Learning." Applied Sciences 13.11 (2023): 6793.

[10] Su D, Wu X, Xu L. GMM-HMM acoustic model training by a two level procedure with Gaussian components determined by automatic model selection[C]//2010 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2010: 4890-4893.

[11] Zhou, S. H. (2021). Research on Detection of Mandarin Pronunciation Errors for Tibetan Students Based on CNN [Master's thesis, Northwest Normal University].

[12] Leung W K, Liu X, Meng H. CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 8132-8136.

[13] Zhu C, Wumaier A, Wei D, et al. Pronunciation error detection model based on feature fusion[J]. Speech Communication, 2024, 156: 103009.

[14] Watanabe S, Hori T, Kim S, et al. Hybrid CTC/attention architecture for end-to-end speech recognition[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(8): 1240-1253.

[15] Zhang, Long, et al. "End-to-end automatic pronunciation error detection based on improved hybrid ctc/attention architecture." Sensors 20.7 (2020): 1809.

[16] Akhtar, Shamila, et al. "Improving mispronunciation detection of arabic words for non-native learners using deep convolutional neural network features." Electronics 9.6 (2020): 963.