A Novel End-to-End CAPT System for L2 Children Learners

Anonymous ACL submission

Abstract

Recently, Conformer-based model shows promising results in automatic speech recognition (ASR) task. There still is a dearth of research on Conformer based model for computer-assisted pronunciation learning (CAPT) system. In this paper, a Conformer-based CAPT system is introduced to provide the mispronunciation detection and diagnosis. We apply the Conformer as the main pronunciation error detection model in phoneme level since superior phoneme recognition performance. Then, the features, including the Log Phone Posterior (LPP), the Log Posterior Ratio (LPR) and some other features, extracted from the Conformer decoder, are trained by a XGBoost model to predict phoneme and sentence level scores labeled by experts. Both results on open datasets and our internal Chinese children data demonstrate that the Conformer-based system, which has smaller model size and detailed diagnosis, achieves better performance compared with neutral network (NN)-based system.

1 Introduction

The computer-aided pronunciation training (CAPT) application plays an important role in the computer-aided language learning (CALL) task especially for the second language (L2) learners. The mispronunciation detection and diagnosis (MD&D) module is the core of CAPT system, providing individualized pronunciation evaluation and associated phone-level diagnosis.

A classic method for MD&D module is pronunciation scoring using the goodness of pronunciation (GOP) (Witt, 1999; Witt and Young, 2000) which is based on the confidence measures of the acoustic model (AM). GOP employs the radio between the likelihood of canonical and the most likely pronounced phones computed by the alignments on given text and the voice of learner. The hybrid deep neural network-hidden markov models (DNN-HMM) architecture is the mostly used in the MD&D system (Hu et al., 2015; Zhang et al., 2021) with more accurate measures using some discriminative training algorithms. Some GOP variations (Sudhakara et al., 2019; Wana et al., 2020) are also introduced in the recent years showing good correlates with human/expert labels.

However, the GOP based methods can only provide the pronunciation scores without insertion errors in the pronunciation. It is hard to detect inserted phonemes/words and give more diagnostic details for the language learners. To deal with these shortcomings, some end-to-end (E2E) based methods (Feng et al., 2020; Yan et al., 2020; Fu et al., 2021; Jiang et al., 2021) are introduced employing a free-phone recognition system with connectionist temporal classification (CTC) (Paterlini-Brechot and Benali, 2007; Leung et al., 2019) or hybrid CTC- Attention model (Watanabe et al., 2017; Yan et al., 2020). With the development of automatic speech recognition (ASR), Conformer (Gulati et al., 2020) based E2E models achieves state-of-the-art (SOTA) performance by combining the convolution neural networks (CNN) and Transformers, to model both local and global dependencies of an audio sequence in a parameter-efficient way. Nevertheless, there is still a lack of CAPT system designed by Conformer model to deal with both MD and diagnosis. So it is necessary to compute a score based on the recognized results to meet the requirements of the exercises and the examinations in the language education.

In this paper, we propose a Conformer-based CAPT system focused on MD&D. Firstly, a Conformer model is trained in the phonetic level. Then, with the force-alignment and the posterior probability given by the trained model, features such as the Log Phone Posterior (LPP), the Log Posterior Ratio (LPR) (Hu et al., 2015) and some other features are extracted. Finally, these features are trained with a XGBoost model to predict phone-level human labeled scores. Compared with (Zhang et al., 2021), the proposed system shows promising performance on the speechocean762[1] data set. A group of experiments also show better performance on Chinese children's speech MD&D task than well tuned online Chain model based engine with smaller model size.

The remainder of this paper is organized as follows: Section 2 introduces the proposed Conformer based CAPT system. Section 3 presents the experiments and results. The conclusion is drawn in the Section 4.

2 Proposed CAPT Framework

Predominant approaches to CAPT primarily performs MD&D by extending ASR technologies, especially through post processing recognition scores. The proposed CAPT framework is shown in Figure. 1 which consists of acoustic model, feature extraction module and feedback module.

2.1 Acoustic Model

For the CAPT, acoustic models (e.g. DNN-HMM) are applied to infer acoustic features. Conformer (Gulati et al., 2020) based model draw immense interest recently and became the dominated model due to its ability to pay attention on both local and global dependencies of the utterance. The encoder is stacked by several blocks which consist of a positionwise feed-forward module, a multihead self-attention module, a convolution module, and another feed-forward module in the end. Only encoder architecture is used in our model to predict phoneme result and evaluate pronunciation for the latency-accuracy trade-off.

In order to achieve phoneme level MD task, the Conformer based encoder is employed to train phonetic level targets. It is necessary to convert transcriptions from word level to phoneme level. The method for obtaining phonetic representation of an input utterance is mainly based on the pronunciation dictionary. There are some notices in the conversion. If a word is not in the vocabulary, the out of vocabulary (OOV), the phonetic sequence of this word will be converted to the '<unk>' unit which represents unknown phones. It is also inevitable that there are some words with multiple pronunciation. In the conversion, the pronunciation will be decided by the part of speech which is implemented by HMM model (Hajic et al., 2009).

[1]https://www.openslr.org/101/

2.2 Feature Extraction

In our Conformer-based system, we employ CTC-based end-to-end force alignment approach to compute phoneme level features. The log phone posterior ratio between the canonical phone and the one with the highest score is used to approximate GOP score as shown,

$$\mathrm{GOP}(p) \approx \log \frac{P(p|\mathbf{o};t_s,t_e)}{\max_{q\in Q} P(q|\mathbf{o};t_s,t_e)} \quad (1)$$

where t_s and t_e are the start and end frame indexes, respectively; Q is the whole phone set; $P(p)$ is the prior of phone p; o is the acoustic aligned segment.

For the phoneme level scoring, LPP and LPR features are extracted by the output and the force alignments. The LPP of phone p is defined as

$$\mathrm{LPP}(p) \approx \sum_{t=t_s}^{t_e} \log P(s_t|\mathbf{o_t}) \Big/ \mathrm{NF}(p), \quad (2)$$

where $\mathrm{NF}(p)$ is the number of frames in the phoneme p; $\mathbf{o}_t$ is the augmented input observations of the frame t; s_t is the senone label of the frame t generated by force alignment with the given canonical phone p; The LPR between phone p_i and p_j is defined as:

$$\begin{aligned}\mathrm{LPR}(p_j|p_i) &= \log P(p_j|\mathbf{o};t_s,t_e) \\ &\quad - \log P(p_i|\mathbf{o};t_s,t_e).\end{aligned} \quad (3)$$

For the sentence level scoring, some features are extracted from the LPP features. There are two feature groups coming from GOP related value and phoneme level recognized results by CTC decoder. The mean values and variances of GOP numerator, GOP denominator and GOP values are computed with the force aligned phoneme sequence. In addition to this, the recognized details including phoneme error rate (PER), the insertion PER, the deletion PER and the substitution PER comparing with the referred phoneme are used for scoring.

2.3 Feedback Module

For phoneme and sentence level scoring, a tree boosting system named XGBoost (Chen and Guestrin, 2016) is trained with the expert labeled score. The phoneme scores are used to detect the mispronunciation. The sentence score is used to represent the overall pronunciation evaluation.

For the phoneme level pronunciation diagnosis, the CTC decoder is applied CTC prefix beam search on the CTC output of the model, which can

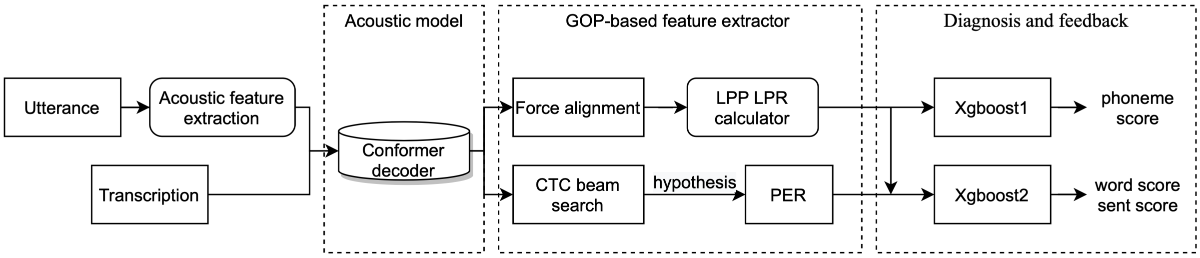

Figure 1: Framework of the proposed CAPT system

give the 1-best results. This phoneme level results are used for the diagnosis based on the reference phoneme sequence.

3 Experiment

3.1 Databases

With reproducibility in mind, open datasets are adopted. Librispeech (Panayotov et al., 2015) is used to train the native AM. A new open-source non-native english speech corpus named "speechocean762" (Zhang et al., 2021), recorded by L2 language learners, is used for pronunciation assessment. To demonstrate the performance on non-native children speech, our internal dataset Yiqi Voice and Yiqi Evaluation are also included. The Yiqi Voice contains 3,000 hours Chinese children's speech of designated transcription and conversation which are used to train AM. The Yiqi Evaluation contains about 100,000 utterances uniformly ranging from 0 to 8 in score labeled by experts which is used for MD&D. All of our internal data are anonymized and eligible to be used for research purposes.

3.2 Implementation Details

In the baseline model, 40 high-resolution mel-frequency cepstral coefficients and 100 dimensional i-vector features are extracted. We use 5 layer time-delay neural networks (TDNN) with 1280 dimensions. The AM output is used for the forced alignment and the computation to obtain the GOP values and the GOP-based features. As regards the evaluation model, support vector machines (SVM) classifiers are built for each phone with the GOP-based features and the corresponding manual scores. Furthermore, chain model (Povey et al., 2016) is applied to compare with Conformer model. A 4-gram language model is trained for the word and phoneme level decoding.

In this paper, Wenet[2] is chosen to train byte pair encoding (BPE) and phoneme level Conformer models. 80 dimensional log-mel filterbank computed are extracted for the model input. In front of the encoder, two convolution sub-sampling layers are used with 4 times sub-sampling in total. We use 12 Conformer blocks in which have 4 multi-head attention with 64 output dimensions. For the XGBoost training, the learning rate is 0.1 and the weights of L1 regularization term is 0.001. The maximum tree depth is 7.

3.3 Recognition Performance

To evaluate the pronunciation, it is necessary to predict the phoneme level score together with the recognized phoneme level results for MD&D.

	Model Arch.	test-clean	test-other
WER	HMM-TDNN	4.3	10.62
	Conformer Attention Rescore	3.12	8.55
PER	HMM-TDNN	9.55	19.86
	Conformer Attention Rescore	1.84	5.48
	Conformer CTC Beam Search	1.87	5.61

Table 1: Comparison of different models in word and phoneme level.

	Model Arch.	Yiqi Voice
WER	Chain model	9.59
	Conformer	8.6

Table 2: Comparisons on Yiqi Voice dataset.

Table 1 and Table 2 show the comparisons of different models on word level and phoneme level. It can be seen that Conformer shows much better word error rate (WER) and PER than HMM-TDNN

[2] https://github.com/mobvoi/wenet

on Librispeech and Yiqi Voice. The PER performance is adequate to detect phoneme level pronunciation error and diagnose the phoneme details. In the Conformer decoding, the CTC beam search decoding method apply CTC prefix beam search on the CTC output and the attention rescore method rescores the n-best candidates applied by the CTC beam search introduced in (Yao et al., 2021). It can be shown that the attention rescore performs better than CTC beam search method slightly. To balance the latency-accuracy trade-off, only the encoder part of AED is applied in this paper.

3.4 Mispronunciation Detection and Scoring Performance

The peroformace of the proposed system is compared with Kaldi recipe on the "speechocean762" dataset and our online CAPT system on the Yiqi Evaluation dataset.

3.4.1 Results on the Speechocean762

For evaluating the system's performance, the scoring results are shown in Table 3 comparing the baseline which can be found in Kaldi recipe[3] (Zhang et al., 2021). It can be seen that our proposed system outperforms the baseline from 0.45 to 0.5 on Pearson correlation coefficient (PCC) and from 0.16 to 0.11 on mean squared error (MSE). Further more, the performance on sentence level on the "speechocean762" test sets shows 0.66 on PCC metrics.

	Level	MSE	PCC
HMM-TDNN	Phoneme	0.16	0.45
Proposed		0.11	0.50
Proposed	Sentence	1.399	0.654

Table 3: Comparisons of the evaluation performance between Kaldi recipe and our proposed system.

3.4.2 Results on the Yiqi Evaluation

To show the advantage of Conformer model in CAPT system, the performance of the proposed system is compared with our online well optimized system which consists of chain model tuned on Yiqi Voice, GOP module and XGBoost scoring model. The engine is implemented using fixed-point, with server-based multi-thread parallel method, for hundreds of millions MD&D requests per day. As shown in the Table 4, the Conformer performs better than the well optimized system with smaller model size.

Model Arch.	Model Size	MSE	PCC
Chain model	33M	0.944	0.90
Proposed	19M	0.802	0.91

Table 4: Comparisons on Yiqi Evaluation dataset

3.5 Diagnosis Advantage

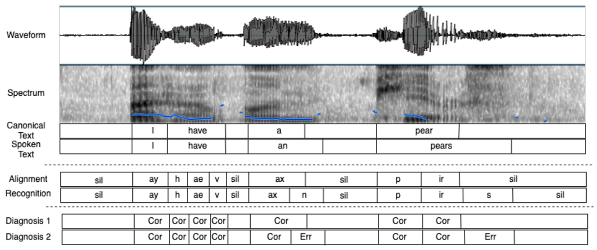

Figure 2: An example to detect and diagnose the insertion errors.

In the traditional alignment based MD&D system, some likely error patterns can be covered making use of the context-dependent phonological rules in the diagnosis. There are still some errors can not be diagnosed if they are not designed in the alignment graph. For example shown in Figure 2, when the reference is "I have a pear", the student pronounces "I have an pears". This kind of insertion error can not be detected and diagnosed based on the alignment methods (Diagnosis 1). In the proposed recognition based diagnosis (Diagnosis 2), the insertion errors like phoneme '/n/' and '/s/' can be detected and used to show the diagnosis for the learners.

4 Conclusion

In this paper, we proposed a Conformer/XGBoost based CAPT system providing MD&D. In the system, a phoneme level Conformer is trained with standard English corpus in this paper. After that, LPP, LPR, and some other features are extracted considering the reference phoneme sequence. Finally, we employ XGBoost to predict the pronunciation scores. The experiment results show that our proposed system outperforms the baseline system and well tuned online Chain model based system with smaller model size. In future, we will introduce works for the abundant dimension evaluations including scores of phoneme, word and sentence level, stress, fluency, prosody and so on.

[3] https://github.com/kaldi-asr/kaldi/tree/master/egs/gop_speechocean762

References

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Yiqing Feng, Guanyu Fu, Qingcai Chen, and Kai Chen. 2020. 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)*, pages 3492–3496. IEEE.

Kaiqi Fu, Jones Lin, Dengfeng Ke, Yanlu Xie, Jinsong Zhang, and Binghuai Lin. 2021. A full text-dependent end to end mispronunciation detection and diagnosis with easy data augmentation techniques. *arXiv preprint arXiv:2104.08428*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Jan Hajic, Jan Raab, Miroslav Spousta, et al. 2009. Semi-supervised training for the averaged perceptron pos tagger. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 763–771.

Wenping Hu, Yao Qian, Frank K Soong, and Yong Wang. 2015. Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. *Speech Communication*, 67:154–166.

Shao-Wei Fan Jiang, Bi-Cheng Yan, Tien-Hong Lo, Fu-An Chao, and Berlin Chen. 2021. Towards robust mispronunciation detection and diagnosis for l2 english learners with accent-modulating methods. *arXiv preprint arXiv:2108.11627*.

Wai-Kim Leung, Xunying Liu, and Helen Meng. 2019. 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)*, pages 8132–8136. IEEE.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Patrizia Paterlini-Brechot and Naoual Linda Benali. 2007. Circulating tumor cells (ctc) detection: clinical impact and future directions. *Cancer letters*, 253(2):180–204.

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. 2016. Purely sequence-trained neural networks for asr based on lattice-free mmi. In *Interspeech*, pages 2751–2755.

Sweekar Sudhakara, Manoj Kumar Ramanathi, Chiranjeevi Yarra, and Prasanta Kumar Ghosh. 2019. An improved goodness of pronunciation (gop) measure for pronunciation evaluation with dnn-hmm system considering hmm transition probabilities. In *INTERSPEECH*, pages 954–958.

Zhenyu Wana, John HL Hansen, and Yanlu Xie. 2020. A multi-view approach for mandarin non-native mispronunciation verification. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8079–8083. IEEE.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. 2017. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

Silke M Witt and Steve J Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, 30(2-3):95–108.

Silke Maren Witt. 1999. Use of speech recognition in computer-assisted language learning.

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

Zhuoyuan Yao, Di Wu, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei. 2021. Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. *arXiv e-prints*, pages arXiv–2102.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021. speechocean762: An open-source non-native english speech corpus for pronunciation assessment. *arXiv preprint arXiv:2104.01378*.