ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746743

TRANSFORMER-BASED MULTI-ASPECT MULTI-GRANULARITY NON-NATIVE ENGLISH SPEAKER PRONUNCIATION ASSESSMENT

Yuan Gong[1], *Ziyi Chen*[2], *Iek-Heng Chu*[2], *Peng Chang*[2], *James Glass*[1]

[1]MIT CSAIL, Cambridge, MA 02139, USA [2]PAII Inc., Palo Alto, CA 94306, USA

{yuangong,glass}@mit.edu {chenziyi253,zhuyixing276,changpeng805}@pingan.com.cn

ABSTRACT

Automatic pronunciation assessment is an important technology to help self-directed language learners. While pronunciation quality has multiple aspects including accuracy, fluency, completeness, and prosody, previous efforts typically only model one aspect (e.g., accuracy) at one granularity (e.g., at the phoneme-level). In this work, we explore modeling multi-aspect pronunciation assessment at multiple granularities. Specifically, we train a Goodness Of Pronunciation feature-based Transformer (GOPT) with multi-task learning. Experiments show that GOPT achieves the best results on speechocean762 with a public automatic speech recognition (ASR) acoustic model trained on Librispeech.

Index Terms— Pronunciation assessment, Transformer

1. INTRODUCTION

Computer assisted pronunciation training (CAPT) is an important technology for self-directed language learning [1, 2, 3], which facilitates non-native (L2) speakers to learn foreign spoken (L1) languages. Compared with conventional classes, CAPT is more economical and convenient, and also allows language learners to receive immediate feedback on their pronunciation. Due to its usefulness, CAPT has been extensively studied, with the majority of these efforts focusing on scoring phoneme-level pronunciation quality (e.g., [4, 5, 6, 7, 8, 9, 10]). Overall pronunciation quality includes many other aspects such as word- and utterance-level fluency, prosody, stress, etc., which have been typically modeled separately (e.g., [11, 12, 13, 14, 15, 16]). However, phoneme-, word-, and utterance-level scores of accuracy, fluency, prosody, and stress are potentially correlated, therefore modeling them jointly instead of separately may allow a machine learning model to learn a more comprehensive representation and in turn improve its performance. In reality, it is also desirable to have a *single* model that can assess *multiple* aspects of pronunciation simultaneously.

As a step in this direction, in this paper we propose a new pronunciation assessment model, named GOPT, based on Goodness of Pronunciation (GOP) features and a Transformer self-attention architecture [17]. We use the open-source speechocean762 dataset [18] that contains one phoneme-level, three word-level, and five utterance-level labels including accuracy, prosody, and fluency and apply *multi-aspect multi-grained* supervision for GOPT training. This not only enables GOPT to measure multiple aspects of pronunciation quality, but also boosts its performance for each assessment task. In addition, the Transformer architecture captures the contextual information between phonemes and words of an utterance. As a consequence, GOPT noticeably outperforms previous methods on the speechocean762 benchmark for both phoneme- and utterance-level assessment tasks (there is no previous work reporting word-level scores). To our knowledge, this is the first work studying multi-aspect L2 speaker pronunciation assessment in a multi-granularity fashion.

2. RELATED WORK

As mentioned, CAPT has been extensively studied with a long history. One major focus of this area is automatic mispronunciation detection, where GOP [4] and its variants (e.g., [10, 5, 7, 8]) are dominant methods. To capture the correlation between phonemes and words of an utterance, self-attention based models such as Transformer [17] have been added on top of GOP features for score modeling to improve performance [9, 19]. There are also some non-GOP based methods such as a wav2vec2-based method [20] and a deep feature based method [21] where transfer learning is usually needed due to the limited L2 training material.

Conversely, automatic assessment of other aspects of pronunciation quality are usually modeled independently, e.g., fluency [11, 12], prosody [13, 14], intonation [15, 16]. There are only a few previous efforts on multi-granularity pronunciation assessment [19, 22]. In these works, however, only a single score is considered for each granularity. In addition, the hierarchical architecture in [19] requires a relatively sophisticated training scheme to optimize.

To the best of our knowledge, this paper is the first to simultaneously consider multiple pronunciation quality aspects (accuracy, fluency, prosody, etc) along with multiple granularities (phoneme, word, utterance). In addition, we show that a BERT-style [23] non-hierarchical standard Transformer architecture can perform well on most assessment tasks. Unlike many previous efforts using non-public datasets or acoustic models, in this work, we intentionally use a public acoustic

Code at https://github.com/YuanGongND/gopt.

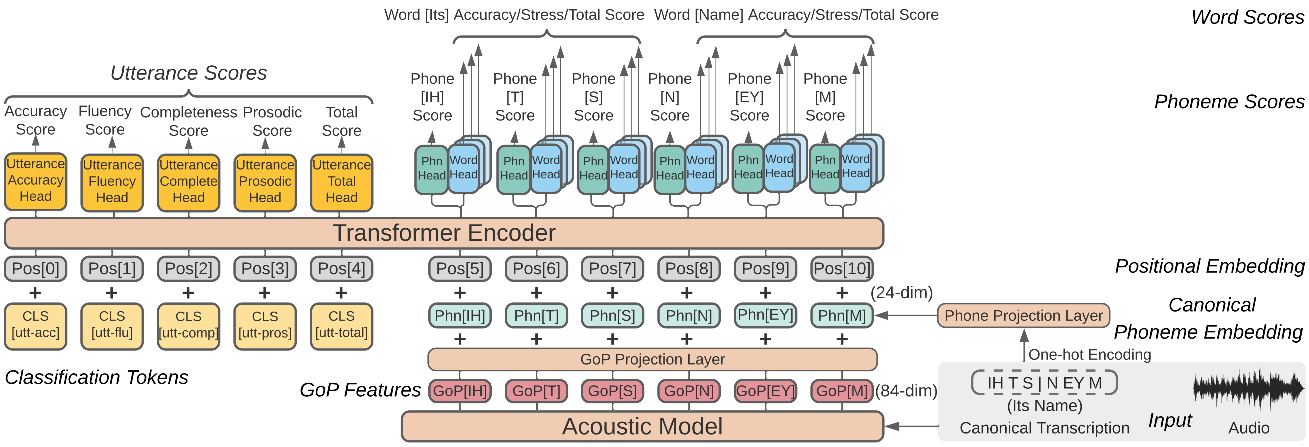

Fig. 1. Illustration of the proposed GOPT architecture with a sample utterance "Its Name", actual utterances used are longer.

model and dataset for our main experiments (which achieves state-of-the-art results) for easy reproduction and comparison.

3. GOODNESS OF PRONUNCIATION TRANSFORMER

3.1. Speechocean762 Dataset

Speechocean762 [18] is a free open-source dataset designed for pronunciation assessment, consisting of a total of 5,000 English utterances collected from 250 non-native speakers. One major advantage of speechocean762 is that it provides rich label information. Specifically, for each utterance, it provides five utterance-level aspect scores: accuracy, fluency, completeness, prosody, and total score (ranging from 0-10). For each word, it provides three word-level aspect scores: accuracy, stress, and total score (ranging from 0-10). It also provides an accuracy score for each phoneme (ranging from 0-2). Each score is annotated by five experts. Thus, it provides a total of 8 labels for different granularities and pronunciation quality aspects. However, the rich annotation has not been fully utilized by previous work. We re-scale utterance and word-level scores to 0-2, making them on the same scale as the phoneme scores. The training set consists of 2,500 utterances, 15,849 words, and 47076 phones; the test set consists of 2,500 utterances, 15,967 words, and 47,369 phones.

3.2. GOPT Architecture Overview

An overview of the GOPT architecture is shown in Figure 1. For the pronunciation assessment task, the canonical transcription is known. We first input the audio and corresponding canonical transcription to the acoustic module to get a sequence of frame-level phonetic posterior-probabilities, which are then force-aligned at the phoneme-level and converted to 84-dimensional goodness of pronunciation (GOP) features (discussed in Section 3.3). The GOP feature is then projected to 24-dimensions with a dense layer. In parallel, we generate a sequence of canonical phoneme embeddings (also at the phoneme-level) by first converting each canonical phoneme to a one-hot encoding and then projecting it to the same 24-dimensions as the projected GOP feature. The reason for using a canonical phoneme embedding is because different phonemes have different characteristics and thus the canonical phoneme provides useful information to the Transformer model [21]. We then add the projected GOP feature, canonical phoneme embedding, and a 24-dimensional trainable positional embedding together and input it to the Transformer encoder. For simplicity, we intentionally follow the original Transformer encoder architecture [17] as close as possible but scale it down to 3 layers and an embedding dimension of 24.

Unlike previous work [19, 21] that use a hierarchical architecture to get utterance level representations, we prepend a set of five trainable `[cls]` tokens to the phoneme-level input sequence in a similar way as BERT [23], each corresponding to one utterance aspect label, and use the output of the Transformer encoder of these `[cls]` aspect tokens as the corresponding utterance-level representations. The reason why this regime works is that the Transformer can learn the correlation between the utterance-level tokens and phoneme-level tokens through the attention mechanism.

During training we apply multi-task learning to the model. Specifically, we use one regression head for each phoneme, word, and utterance label (eight in total). Each regression head is a 24 × 1 dense layer with layer normalization. Utterance-level regression heads are added on top of the output of the Transformer of the corresponding utterance `[cls]` tokens. Phoneme- and word-level regression heads are added on top of the Transformer output of each corresponding phoneme. We propagate the word score to each of its phonemes during training and average the output of phonemes that belong to the word in inference. We use mean squared error (MSE) loss for each assessment task. Since we normalize the scores to the same scale, for simplicity, we first average the losses of each granularity and then sum them up with the same weight, i.e., $\mathcal{L} = \mathcal{L}_{utterance} + \mathcal{L}_{word} + \mathcal{L}_{phoneme}$, where $\mathcal{L}_{utterance}$

Model	Phoneme Score		Word Score (PCC)			Utterance Score (PCC)				
	MSE ↓	PCC ↑	Accuracy ↑	Stress ↑	Total ↑	Accuracy ↑	Completeness ↑	Fluency ↑	Prosodic ↑	Total ↑
RF [18]	0.130	0.440	-	-	-	-	-	-	-	-
SVR [18]	0.160	0.450	-	-		-	-	-	-	-
Lin et.al [21]	-	-	-	-	-	-	-	-	-	0.720
LSTM	0.089	0.591	0.514	**0.294**	0.531	**0.720**	0.076	0.745	0.747	0.741
	±0.000	±0.003	±0.003	±**0.012**	±0.004	±**0.002**	±0.086	±0.002	±0.005	±0.002
GOPT	**0.085**	**0.612**	**0.533**	0.291	**0.549**	0.714	**0.155**	**0.753**	**0.760**	**0.742**
(Librispeech)	±**0.001**	±**0.003**	±**0.004**	±0.030	±**0.002**	±0.004	±**0.039**	±**0.008**	±**0.006**	±**0.005**
GOPT	0.069	0.679	0.588	0.146	0.601	0.727	0.011	0.692	0.694	0.732
(PAII-A)	±0.000	±0.001	±0.004	±0.004	±0.003	±0.004	±0.069	±0.015	±0.009	±0.006

Table 1. Comparing the performance of various pronunciation assessment tasks between GOPT and baseline models. GOPT (PAII-A) depends on a different acoustic model so its results (shown in grey) cannot be directly compared with other models.

and $\mathcal{L}_{word}$ are averaged utterance and word level losses of five utterance-level labels and three word-level labels, respectively; $\mathcal{L}_{phoneme}$ is the phoneme loss. The entire network (except the acoustic model) is trained end-to-end.

3.3. Acoustic Model and GOP Feature

For our main experiment we use a public ASR acoustic model[1] trained with Librispeech [24] 960-hour data. The model is based on the factorized time-delay neural network (TDNN-F) and trained with the Kaldi Librispeech S5 recipe.

Acoustic model trained on both L1 and L2 speech generates better alignment for L2 speech and may output better GOP features [25]. To explore if GOPT works with different acoustic models, we also test with two PAII internal acoustic models PAII-A and PAII-B, both are also TDNN-F models. PAII-A is trained with 452 hours L1 TED-LIUM 3 [26] data and 1,696 hours of L2 data collected from 5,994 non-native speakers; PAII-B is trained with 995 hours of L1 data (from WSJ [27], TED-LIUM 3 [26], and Librispeech [24]) and 6,591 hours of L2 data from 672k non-native speakers.

In this work, we use the log phone posterior (LPP) and log posterior ratio (LPR) defined in [8] as GOP features. Specifically, the LPP of a phone p is defined as follows:

$$LPP(p) \approx \frac{1}{t_e - t_s + 1} \sum_{t=t_s}^{t_e} \log p(p|o_t) \tag{1}$$

$$p(p|o_t) = \sum_{s \in p} p(s|o_t) \tag{2}$$

where t_s and t_e are the start and end frame indexes; o_t is the input observation of the frame t, s is the state belonging to the phone p. LPR of a phone p_j versus p_i is defined as:

$$LPR(p_j|p_i) = \log p(p_j|\mathbf{o}; t_s, t_e) - \log p(p_i|\mathbf{o}; t_s, t_e) \tag{3}$$

The Librispeech acoustic model we use has a total of 42 pure phones, thus the GOP feature of phone p can be defined as a 84-dimensional vector as follows:

$$[LPP(p_1)..., LPP(p_{42}), LPR(p_1|p)..., LPR(p_{42}|p)] \tag{4}$$

4. EXPERIMENTS

For all experiments, we train the model with an Adam optimizer, an initial learning rate of 1e-3, a batch size of 25, and MSE loss for 100 epochs using the official speechocean762 training set, and evaluate on the official test set. The learning rate is cut in half every five epochs after the 20th epoch, and the result of the last epoch is reported. We repeat each experiment five times with different random seeds and report the mean and standard deviation of the results. Since the speechocean762 labels are imbalanced (biased towards high scores), we use the Pearson correlation coefficient (PCC) as the main evaluation metric but also report MSE of the phoneme accuracy score to make a comparison with previous work. Note that while we re-scale the utterance and word level scores, PCCs and phoneme-level MSE are not impacted.

4.1. Main Results

We compare the following six models: 1) Random forest regression (RF) model implemented in the code repository of [18]; 2) Support vector regressor (SVR) based model in [18]; 3) Deep feature and transfer learning-based model presented in [21]; 4) An LSTM based model implemented by us. To make a fair comparison, the LSTM model has the same depth and embedding dimension as the GOPT model and is trained with the same setting. The output of the last token is used as the utterance representation and, as with GOPT, the word score is propagated to its phones; 5) The proposed GOPT model with the Librispeech acoustic model. 6) The proposed GOPT model with the PAII-A acoustic model. It is worth mentioning that models 1-5 are all based on acoustic models trained with the same Librispeech data, and models 1,2,4,5, and 6 use the same GOP features (model 3 does not use GOP features but deep transfer learning). Therefore, we make a fair comparison and the performance difference is not due to the acoustic model and GOP features.

We show the results in Table 1. The key findings are as follows: First, the proposed GOPT model can perform well on most assessment tasks except word stress score and sentence completeness score assessment, demonstrating that it

[1] https://kaldi-asr.org/models/m13

Setting	Phoneme	Word	Utterance
Training Task			
Only Phoneme	0.605±0.002	-	-
Only Word	-	0.536±0.004	-
Only Utterance	-	-	0.736±0.011
Joint*	**0.612±0.003**	**0.549±0.002**	**0.742±0.005**
Canonical Phoneme Embedding			
w/o Phn Embed	0.512=0.006	0.472±0.006	0.719=0.002
w/ Phn Embed*	**0.612±0.003**	**0.549±0.002**	**0.742±0.005**
# Transformer Layer (ASR params not included in #params)			
3* (27K Params)	**0.612±0.003**	**0.549±0.002**	**0.742±0.005**
6 (48K Params)	0.605±0.003	0.543±0.004	0.731±0.003
Embedding Dimension (ASR params not included in #params)			
12 (8K Params)	0.608±0.003	0.544±0.008	0.741±0.011
24* (27K Params)	**0.612±0.003**	**0.549±0.002**	**0.742±0.005**
48 (94K Params)	0.605±0.003	0.545±0.006	0.738±0.004
96 (355K Params)	0.586±0.006	0.530±0.006	0.725±0.004

Table 2. The ablation results, we only show the PCC of phoneme, word, and utterance total scores due to space limitation. * denotes the setting used in the base GOPT model.

is possible to have a single model for multi-aspect and multi-granularity pronunciation assessment. Specifically, the GOPT achieves 0.085 MSE and 0.612 PCC for the phoneme accuracy score assessment, noticeably outperforming the models in [18]; GOPT achieves 0.742 PCC for the utterance-level score assessment, noticeably outperforming the model in [21] which uses more sophisticated features than GOP. We hypothesize that the poor utterance completeness assessment performance is due to the highly imbalanced distribution of the completeness score in the training data. Second, the multi-task learning scheme can be also applied to an LSTM, which achieves similar results for utterance assessment with GOPT. However, the performance of the LSTM for phoneme-level and word-level assessment are worse than the GOPT, demonstrating that the Transformer architecture is better at modeling fine-grained pronunciation units. Third, using the PAII-A acoustic model trained on both L1 and L2 speech can further boost the phoneme and word assessment performance by around 10%, but the utterance-level performance is worse than just using the Librispeech acoustic model. We also evaluate GOPT with PAII-B acoustic model, it leads to similar results with GOPT with PAII-A acoustic model.

4.2. Ablations

We conduct a set of ablation studies to show the performance impact of various factors. We set the GOPT model mentioned in Section 3 with three Transformer layers, embedding dimension of 24, canonical phoneme embedding, and trained with all phoneme, word, and utterance assessment tasks as the base GOPT model, and then change one factor at a time to observe the performance change.

Scoring Model	Acoustic Model					
	Librispeech		PAII-A		PAII-B	
	MSE ↓	PCC ↑	MSE ↓	PCC ↑	MSE ↓	PCC ↑
SVR	0.160	0.450	0.118	0.538	0.115	0.561
GOPT	0.085 ±0.001	0.612 ±0.003	0.069 ±0.000	0.679 ±0.001	0.071 ±0.001	0.662 ±0.001

Table 3. Comparing the phoneme assessment performance between the SVR based [18] model and proposed GOPT model with various acoustic models.

We show the results in Table 2. First, we see that the GOPT trained with multi-task learning achieves better results than any single-task learning model, demonstrating that multi-task learning not only allows the model to conduct multi-aspect and multi-granularity pronunciation assessment simultaneously, but also improves the performance of each individual task. Second, we see that the canonical phoneme embedding is crucial to the performance as the model trained without it performs much worse for all tasks. However, it is worth mentioning that canonical phoneme embedding is not the reason why GOPT outperforms previous methods since canonical phoneme embedding is also used in [21]. In [18], each phoneme has a separate classifier, which serves a similar function as a canonical phoneme embedding. Third, we explore the performance impact of the size of GOPT model, and see that increasing either the width or depth of the network cannot further improve the performance, indicating that a small model is preferred with the relatively small dataset. Further, although the GOP feature is 84-dimensional, we show that an embedding size of 24 is sufficient to represent pronunciation quality with a Transformer.

Finally, in Table 3, we compare the phoneme assessment performance between the SVR [18] model and the proposed GOPT model with various acoustic models. We show that the proposed GOPT consistently leads to a significant performance improvement regardless of the acoustic model, demonstrating that the GOPT is model agnostic and can be used with different acoustic models.

5. CONCLUSION

In this paper, we present the Transformer-based multi-aspect multi-granularity pronunciation assessment model GOPT. We show that with the multi-task learning scheme, a single GOPT model can conduct multiple pronunciation tasks simultaneously, and its performance is better than the same model trained with a single task. Experiments show the GOPT can noticeably outperform previous methods on speechocean762.

6. REFERENCES

[1] Maxine Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, 2009.

[2] Klaus Zechner, Derrick Higgins, et al., "Automatic scoring of non-native spontaneous speech in tests of spoken english," *Speech Communication*, 2009.

[3] Silke M Witt, "Automatic error detection in pronunciation training: Where we are and where we need to go," in *International Symposium on Automatic Detection on Errors in Pronunciation Training*, 2012.

[4] S.M Witt and S.J Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, 2000.

[5] Feng Zhang, Chao Huang, et al., "Automatic mispronunciation detection for mandarin," in *ICASSP*, 2008.

[6] Dean Luo, Yu Qiao, et al., "Analysis and utilization of mllr speaker adaptation technique for learners' pronunciation evaluation," in *Interspeech*, 2009.

[7] Yow-Bang Wang and Lin-Shan Lee, "Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training," in *ICASSP*, 2012.

[8] Wenping Hu, Yao Qian, et al., "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, 2015.

[9] Jiatong Shi, Nan Huo, and Qin Jin, "Context-aware goodness of pronunciation for computer-assisted pronunciation training," in *Interspeech*, 2020.

[10] Joost van Doremalen, C Cucchiarini, and H Strik, "Using non-native error patterns to improve pronunciation verification," in *Interspeech*, 2010.

[11] Catia Cucchiarini, Helmer Strik, and LWJ Boves, "Quantitative assessment of second language learners' fluency: an automatic approach," in *ICSLP*, 1998.

[12] Catia Cucchiarini, Helmer Strik, and Lou Boves, "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology," *The Journal of the Acoustical Society of America*, 2000.

[13] Paul Christopher Bagshaw, *Automatic prosodic analysis for computer aided pronunciation teaching*, Ph.D. thesis, 1994.

[14] Joseph Tepperman and Shrikanth Narayanan, "Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners," in *ICASSP*, 2005.

[15] Juan Pablo Arias, Nestor Becerra Yoma, and Hiram Vivanco, "Automatic intonation assessment for computer aided language learning," *Speech Communication*, 2010.

[16] Kun Li, Xixin Wu, and Helen Meng, "Intonation classification for l2 english speech using multi-distribution deep neural networks," *Computer Speech and Language*, 2017.

[17] Ashish Vaswani, Noam Shazeer, et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[18] Junbo Zhang, Zhiwen Zhang, et al., "speechocean762: An open-source non-native english speech corpus for pronunciation assessment," in *Interspeech*, 2021.

[19] Binghuai Lin, Liyuan Wang, Xiaoli Feng, and Jinsong Zhang, "Automatic scoring at multi-granularity for l2 pronunciation.," in *Interspeech*, 2020.

[20] Xiaoshuo Xu, Yueteng Kang, Songjun Cao, Binghuai Lin, et al., "Explore wav2vec 2.0 for mispronunciation detection," *Interspeech*, 2021.

[21] Binghuai Lin and Liyuan Wang, "Deep feature transfer learning for automatic pronunciation assessment," *Interspeech*, 2021.

[22] Tobias Cincarek, Rainer Gruhn, et al., "Automatic pronunciation scoring of words and sentences independent from the non-native's first language," *Computer Speech and Language*, 2009.

[23] Jacob Devlin, Ming-Wei Chang, et al., "Bert: Pre-training of deep bidirectional transformers for language understanding," *ACL*, 2018.

[24] Vassil Panayotov, Guoguo Chen, et al., "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*, 2015.

[25] Ming Tu, Anna Grabek, et al., "Investigating the role of l1 in automatic pronunciation evaluation of l2 speech," in *Interspeech*, 2018.

[26] François Hernandez, Vincent Nguyen, et al., "Ted-lium 3: twice as much data and corpus repartition for experiments on speaker adaptation," in *SPECOM*, 2018.

[27] Mitchell P Marcus, Mary Ann Marcinkiewicz, et al., "Building a large annotated corpus of english: the penn treebank," *Computational Linguistics*, 1993.