INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

A noise robust method for word-level pronunciation assessment

Binghuai Lin[⋆] *, Liyuan Wang*[⋆]

[⋆]Smart Platform Product Department, Tencent Technology Co., Ltd, China
{binghuailin, sumerlywang}@tencent.com

Abstract

The common approach for pronunciation evaluation is based on Goodness of pronunciation (GOP). It has been found that GOP may perform worse under noise conditions. Traditional methods compensate pronunciation features to improve the performance of pronunciation assessment in noise situations. This paper proposed a noise robust model for word-level pronunciation assessment based on a domain adversarial training (DAT) method. We treat the pronunciation assessment in the clean and noise situations as the source and target domains. The network is optimized by incorporating both the pronunciation assessment and noise domain discrimination. The domain labels are generated from unsupervised methods to adapt to various noise situations. We evaluate the model performance based on English words recorded by Chinese English learners and labeled by three experts. Experimental results show on average the proposed model outperforms the baseline by 3% in Pearson correlation coefficients (PCC) and 4% in accuracy under different noise conditions.

Index Terms: noise robust, pronunciation assessment, domain adversarial, unsupervised method

1. Introduction

For English-as-second language (ESL) learners, many Computer-Assisted Pronunciation Training (CAPT) systems are designed to conduct automatic pronunciation assessment [1].

Features used for automatic pronunciation assessment are usually extracted from an automatic speech recognizer. The hidden Markov model (HMM) likelihood, posterior probability, and pronunciation duration features were proposed for pronunciation assessment in [2]. A variation of the posterior probability ratio called the Goodness of Pronunciation (GOP) [3] was proposed for pronunciation evaluation and error detection, and it has prevailed in the related studies ever since [4, 5, 6]. GOP was further optimized based on the deep neural network (DNN) to improve the accuracy of phoneme mispronunciation detection [7].

Compared with traditional speech scoring methods by human raters, automatic pronunciation assessment has proven to be effective and efficient. However, it's still of great challenge for feature extraction utilizing automatic speech recognition (ASR) in the noise environment. It has been found that the ASR-based speech scoring system is less robust than human raters when facing low-quality audios, e.g., audios recorded by low-quality microphones or with different kinds of noises [8].

In this paper, we propose a noise robust model for automatic word-level pronunciation assessment. The network consists of a word pronunciation scoring model and a noise domain discriminator model. To adapt to various noise situations, we generate noise labels by an unsupervised method with the noise-related input features. As GOP values at different word ratings are influenced by noise differently, we utilize a separate unsupervised noise label generator for each word rating. We conduct experiments under different noise conditions to demonstrate the robustness of the model under different noise conditions, e.g., babble noise and white noise at different values of SNR. We also demonstrate the noise robustness under real noise situations. The remainder of the paper is organized as follows. In section 2, we will introduce the related work. In section 3, we will demonstrate our proposed model. In section 4, the corpus used in the proposed model is introduced, and the hyperparameters of the proposed network are explained. We will show the experimental results of the proposed model and discuss the results in section 5. We will draw the conclusion and future suggestions in section 6.

2. Related work

Many approaches have been proposed to improve noise robustness for ASR or pronunciation assessment. To make ASR robust to the full range of the real-word noise and other acoustic distorting conditions, a wide range of noise robust techniques were analyzed and categorized using five different criteria [9]. Some method proposed a model that maps both clean inputs and their noisy counterparts onto the same point in the representation space and proved to be robust [10]. For CAPT systems, GOP plays an important part in pronunciation assessment and mispronunciation detection. Some studies have been conducted to make GOP more robust to noise. Researchers applied a noise compensation technique, namely Stereo-based Piecewise Linear Compensation for Environments (SPLICE), and took the compensated feature sequences as input to the GOP-based assessment system. The experimental results show the improvements in performance in a noisy classroom environment [11]. However, the performance of SPLICE will deteriorate for input with unknown noise types. The study proposed TGOP (teacher utterance-based GOP) and GL (GOP-like) scores, namely two variants of GOP, and revealed that the correlation coefficient between GOP scores and teacher's ratings had been improved under all noise conditions [12]. It usually needs teachers' utterances corresponding to the learners' sentences.

DAT learns features by taking into account domain invariance, the label predictor, and the domain classifier [13]. It has been widely used in many applications. A DNN-based method for noise robust speech recognition by domain adversarial training was proposed [14]. An unsupervised feature adaptation using adversarial multi-task (MTL) training for automatic evaluation of children's speech has been surveyed [15]. DAT was also utilized to tackle the noise type mismatch problem between the training and testing conditions in speech enhancement [16]. Different from the traditional DAT methods, we generate domain labels by an unsupervised method without human annotations to adapt to various noise situations.

⋆ equal contribution

http://dx.doi.org/10.21437/Interspeech.2021-1005

3. Proposed model

We propose a noise robust method for second language (L2) learners' word-level pronunciation assessment. The baseline model utilizes a neural network with GOP scores and additional feature representations called phoneme embedding as input features. The whole network is optimized by the DAT method, which combines the word score classification task as the label predictor and the noise classification task as the domain discriminator. These two tasks share common phoneme features and are optimized simultaneously. As there are many kinds of unknown noise, and the annotation is time-consuming, we obtain the noise labels automatically by an unsupervised model called K-means clustering [17] based on the extracted noise features depending on different word ratings. The network is shown in Figure 1.

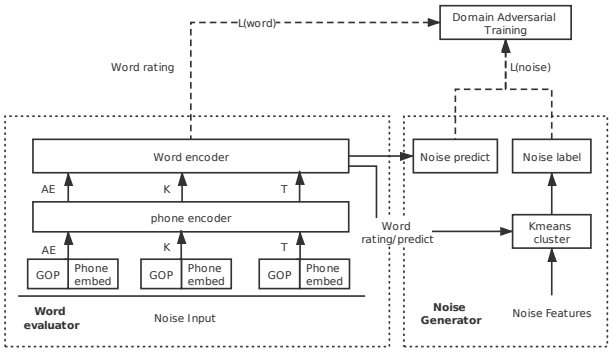

Figure 1: *Noise robust pronunciation assessment model*

3.1. Baseline system for word rating

Our baseline model is based on a neural network. GOP score is used as the input feature of our proposed network. GOP score is defined as Eq. (1) [3]:

$$\mathrm{GOP}(p) = \frac{|\log(P(p|o^p)|}{\mathrm{NF}(p))} = \frac{|\log(\frac{P(o^p|p)P(p)}{\sum_{q\in Q} P(o^q|q)P(q)})|}{\mathrm{NF}(p)} \quad (1)$$

where $P(p|o^p)$ is the posterior probability of phoneme p given pronunciation o and Q represents all possible phonemes corresponding to pronunciation o. $\mathrm{NF}(p)$ is the pronunciation frames of phoneme p and $P(p)$ is prior probability of phoneme p. To calculate GOP, the phonemes of utterances are forced-aligned first by a Kaldi-based ASR system [18].

To convey acoustic information of individual phonemes, we employ a distinct numerical representation for each phoneme which is called phoneme embedding [19]. Combined with GOP scores as input features, we apply a non-linear transformation (i.e., phone encoder) on phoneme features with a sigmoid activation function. Next, we obtain the word feature representations through a non-linear transformation based on the averaged phoneme features (i.e., word encoder). Each word can be classified into five classes indicating degrees of pronunciation proficiency of non-native speakers through a fully connected network (FC) based on the word features. The noise prediction results are generated from an FC layer as well.

The baseline model is optimized with the classification loss of word pronunciation, which is defined as the cross-entropy between predicted classification results and word ratings labeled by human raters.

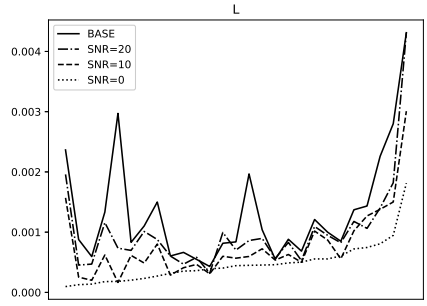

Figure 2: *GOP scores of phoneme 'L' at different SNRs*

3.2. Noise label generator

As there are various noises in real situations, we employ an unsupervised learning model with noise-related features as input to generate noise domain labels.

We extract some audio features for noise classification. To calculate the noise features for the learners' pronunciation, we align phonemes in a word to obtain the beginning and ending times of each phoneme using a Kaldi-based ASR system as mentioned in section 3.1. We compute the power features, which is the transmission rate of energy, for each 10ms frame with a 10ms window shift. Based on the power and the duration of each phoneme in the word, we compute multiple features for noise classification. As SNR is defined as the ratio of signal power to the noise power, it has been regarded as an important feature for noise classification. The final features extracted for noise classification are shown in Table 1.

Table 1: *Features for noise classification*

AudioAvgPower/ AudioPeakPower	Mean/Max value of power in the audio
SpeechAvgPower/ SpeechPeakPower	Mean/Max value of power in the voiced segment of audio
SilAvgPower/ SilPeakPower	Mean/Max value of power in the unvoiced segment of audio
AvgSnr/ PeakSnr	Mean/Max value of SNR in the audio

To take into account the influence of noise on pronunciation assessment, we investigate the relationship between noise features and GOP scores. We show GOP scores of some phoneme at different SNRs by adding babble noise to clean audios manually in Figure 2. The x-axis of the figure shows the sample indexes, and the y-axis represents GOP scores of one phoneme at different SNRs. From Figure 2, it is clear that GOP scores of phoneme 'L' decrease with decreasing SNR values.

As each word rating has GOP values in different ranges and may be influenced by different kinds of noise to varying degrees, we create multiple noise label generators depending on the word ratings. The purpose of the noise label generator is to label each word utterance under different word ratings automatically without human annotators. This is done through the K-means clustering method based on the aforementioned features. Specifically, we cluster noise features under word ratings

from 1 to 5, respectively, and obtain five corresponding noise label generators. A specific noise generator is chosen according to word rating labels during training or prediction from the pronunciation assessment model during inference.

3.3. Training based on DAT

The whole network is trained based on DAT, consisting of a label predictor for word rating and a domain classifier for noise classification.

Noise classification and pronunciation assessment share common features and are optimized simultaneously. The training loss is defined as Eq. (2):

$$L_{\text{total}} = L_{\text{word}}(\theta_F, \theta_y) - \lambda \times L_{\text{noise}}(\theta_F, \theta_D) \quad (2)$$

where θ_F is the feature representations in the phone encoder and word encoder, and θ_y and θ_D are the parameters of the word rating predictor and the domain classifier. λ is the weight to balance two tasks. The noise classification loss is defined as the cross-entropy between predicted noise class and labels generated from the unsupervised noise generators.

The DAT loss is optimized as Eq. (3) and Eq. (4):

$$\theta_F, \theta_y = \arg\min_{\theta_F, \theta_y} L_{\text{total}} \quad (3)$$

$$\theta_D = \arg\max_{\theta_D} L_{\text{total}} \quad (4)$$

4. Experimental setup

4.1. Corpus description

The corpus consists of $7,500$ English words read by 720 Chinese speakers with ages evenly distributed from 16 to 20 years in a clean environment. Each speaker records around 10 words. The phoneme set used here is based on the Carnegie Mellon University (CMU) pronouncing dictionary composed of 39 different phonemes [20]. Each word is rated by three experts on a scale of 1-5 with 1 representing hardly understandable pronunciation and 5 representing native-like pronunciation. Final word ratings are achieved by a majority vote. The inter-rater labeling consistency is evaluated at Kappa, which is calculated by averaging any two raters based on 1000 words randomly chosen, and the final value is 0.63 with the 95% confidence interval $(0.628, 0.632)$ and p-value less than 0.1%, indicating a fairly good quality of labeling. The final distribution of word ratings is shown on the left side of Figure 3. We split $7,500$ English words into training and testing data. The training data consist of $6,300$ English words, and the testing data consist of $1,200$ English words.

To test the robustness of the proposed model, we add different kinds of noise at SNRs ranging from 0dB to 20dB to the clean training data. We add babble noise to simulate the situation when the learners' voice is corrupted by surrounding students' voice and white noise to simulate the environment of low recording quality. These are the most common types of noise observed in CAPT systems [12]. The babble noise is taken from the Microsoft Scalable Noisy Speech Dataset [21]. To test the performance of the proposed method in real noisy environments, we collect 1000 audios under different noise conditions such as babble and humming noise at less than 10dB SNR conditions from one online pronunciation learning application, whose labels are annotated by the same method. The final distribution of word ratings is shown on the right side of Figure 3.

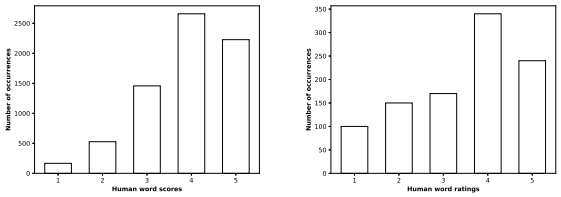

Figure 3: *Distribution of labels under clean and real noisy conditions*

4.2. Hyperparameters

The model input includes GOP scores and phoneme embedding. The phoneme embedding is composed of 39 different phonemes with a feature dimensionality of 15 based on the CMU pronouncing dictionary. The total number of the feature dimensionality is 16, including 15 for phoneme embedding and 1 for phoneme GOP. The number of parameters for the non-linear transformation of phoneme features is 16×20, which converts phoneme features of the dimensionality of 16 to the output dimensionality of 20. Based on averaged phoneme representations of the word, word representations are achieved by another non-linear transformation, and the number of parameters is 20×10. Two FC layers for word pronunciation assessment classification and noise classification are implemented with the number of parameters of 10×5 and $10 \times 5 \times N(noise\ clusters)$, indicating transforming features with the dimensionality of 10 to the word scoring class of 5 and features with the dimensionality of 10 to the number of noise clusters at each word rating. As noise labels come from an unsupervised noise label generator, the number of noise clusters can be tuned to achieve the best performance. The network performance with different numbers of noise clusters will be shown in section 5.

5. Results

We evaluate the performance of the word pronunciation assessment by PCC and accuracy. First, we compare the baseline model with the proposed model with different experimental settings in the clean situation. Then, we analyze the performance of these models under different noise conditions. Finally, we conduct some ablation studies to prove the validation of the proposed model.

5.1. Results of comparison

The baseline model is introduced in section 3.1. To take into account the noise influence directly, we conduct experiments combining GOP and noise features as input of the baseline model as well. Three experiments are conducted, including the baseline model (Base), the baseline model combined with noise features as input (Base + NoiseF), and the proposed model with cluster number of 2, i.e., the baseline model combined with DAT (Noise DAT). The results based on the clean testing data are shown in Table 2, indicating that with the DAT method, the performance of the word pronunciation assessment can be slightly improved in the clean situation.

To test the noise robustness of the proposed model, we also compare results with the performance of the previous method SPLICE [11], which compensates for the degradation of GOP scores under noise situations. Results of the babble noise and white noise of different SNRs are shown in Table 3 and Table

Table 2: *Performance of baselines and the proposed model under clean conditions*

Model	PCC	ACC
Base	72.1	85.3
Base+NoiseF	73.3	85.1
Noise DAT	**74.2**	**86.4**
SPLICE [11]	73.1	85.5

Table 3: *Performance under conditions with babble noise*

		0dB	10dB	20dB
Base	PCC	53.1	62.6	67.9
	ACC	58.7	71.7	82.5
Base+NoiseF	PCC	54.2	63.1	69.7
	ACC	59.1	70.1	81.5
Noise DAT	PCC	**55.7**	**65.0**	**70.6**
	ACC	**62.5**	73.9	**84.3**
SPLICE [11]	PCC	54.6	61.8	67.3
	ACC	60.8	**74.5**	83.9

Table 4: *Performance under conditions with white noise*

		0dB	10dB	20dB
Base	PCC	44.8	51.1	62.7
	ACC	45.5	61.7	76.3
Base+NoiseF	PCC	46.4	55.5	64.7
	ACC	44.5	63.0	75.9
Noise DAT	PCC	**47.8**	**56.5**	**66.7**
	ACC	**51.7**	65.0	**78.7**
SPLICE [11]	PCC	45.6	53.1	64.5
	ACC	50.9	**66.0**	77.5

4. From the results, we can see all three models degrade nearly 5% at least with different kinds of noise addition compared with the baseline shown in Table 3 and Table 4, indicating a significant negative impact of noise on the pronunciation assessment model. The proposed model consistently demonstrates good performance under different noise conditions. From comparison results with the SPLICE method, we can see SPLICE has comparable results under known noise conditions.

To test the robustness of the proposed model in real noise situations, we conduct the experiments based on the data with low SNRs as well. The results are shown in Table 5. From the results, we can see the experiments with Base+NoiseF and SPLICE appear to be less robust in real noise situations, which may result from different noise distributions between the testing data and the training data. Our proposed model proves to be more robust in the unknown real noisy environment.

5.2. Ablation study

In our experiments, we use individual noise generators for each word ratings. We experiment with three different settings in the babble noise environment to validate our setting: (1) our proposed model with noise labels depending on word ratings (Noise DAT+Word rating); (2) our proposed model with noise labels regardless of word ratings (Noise DAT+No word rating); (3) our proposed model with true noise labels, which are determined by SNR and noise type (Noise DAT+Human label). Results are shown in Table 6. From the results, we can see the setting (Noise DAT+No word rating) can hardly achieve performance improvement over the baseline. With true labels, the performance is better than the baseline but inferior to our proposed setting.

Table 5: *Performance under conditions with real noise*

Model	PCC	ACC
Base	70.4	62.1
Base+NoiseF	69.2	50.1
Noise DAT	**71.4**	**65.3**
SPLICE [11]	70.8	60.1

Table 6: *Comparison of methods of generating noise labels in 10dB babble noise*

Model	PCC	ACC
Base	62.6	71.7
Noise DAT+Word rating	**65.0**	**73.9**
Noise DAT+No word rating	62.4	72.5
Noise DAT+Human label	63.1	73.5

Table 7: *Different clusters of Noise DAT (10dB babble noise)*

Clusters	PCC	ACC
2	**65.0**	**73.9**
3	63.2	72.8
4	62.8	72.3

Table 8: *Silhouette values of K-means with different K values*

	Word ratings				
Clusters	1	2	3	4	5
2	**0.32**	**0.35**	**0.32**	**0.33**	**0.65**
3	0.30	0.29	0.24	0.28	0.42
4	0.25	0.25	0.25	0.24	0.4

As noise labels generated from the noise generator can be varied by setting different clustering numbers, we conduct several experiments with 2, 3, and 4 clusters in the situation with 10dB babble noise. Results are shown in Table 7, indicating that the word pronunciation assessment task performs best with a cluster number of 2. The proposed model demonstrates inconsistent performance with different numbers of noise clusters. To analyze the results, we compute silhouette value [15] for the clustering, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. Results are shown in Table 8, indicating good clustering with the clustering number of 2 and the biggest silhouette value.

6. Conclusion

In this paper, we propose a noise robust method for word-level pronunciation assessment. As GOP scores are easily influenced by noise, we employ a domain adversarial training method, which is composed of a label generator for word pronunciation evaluation and a domain discriminator for noise domain discrimination. For adaptation to various noise situations, labels of noise classification are generated from an unsupervised model based on the word ratings. Experimental results show the proposed model performs well in different noisy environments. In the future, we will investigate different noise-related factors to further improve the robustness of automatic pronunciation assessment.

7. References

[1] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

[2] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic pronunciation scoring for language instruction," in *1997 IEEE international conference on acoustics, speech, and signal processing*, vol. 2. IEEE, 1997, pp. 1471–1474.

[3] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[4] S. Kanters, C. Cucchiarini, and H. Strik, "The goodness of pronunciation algorithm: a detailed performance study," 2009.

[5] H. Ryu, H. Hong, S. Kim, and M. Chung, "Automatic pronunciation assessment of korean spoken by l2 learners using best feature set selection," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2016, pp. 1–6.

[6] V. Laborde, T. Pellegrini, L. Fontan, J. Mauclair, H. Sahraoui, and J. Farinas, "Pronunciation assessment of japanese learners of french with GOP scores and phonetic information," 2016.

[7] W. Hu, Y. Qian, and F. K. Soong, "A new DNN-based high quality pronunciation evaluation for computer-aided language learning (call)." in *Interspeech*, 2013, pp. 1886–1890.

[8] L. Chen, "Audio quality issue for automatic speech assessment," in *International Workshop on Speech and Language Technology in Education*, 2009.

[9] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[10] D. Liang, Z. Huang, and Z. C. Lipton, "Learning noise-invariant representations for robust speech recognition," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 56–63.

[11] Y. Luan, M. Suzuki, Y. Yamauchi, N. Minematsu, S. Kato, and K. Hirose, "Performance improvement of automatic pronunciation assessment in a noisy classroom," in *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 428–431.

[12] S. Sudhakara, M. K. Ramanathi, C. Yarra, A. Das, and P. K. Ghosh, "Noise robust goodness of pronunciation measures using teacher's utterance." in *SLaTE*, 2019, pp. 69–73.

[13] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[14] Y. Shinohara, "Adversarial multi-task learning of deep neural networks for robust speech recognition," in *Interspeech*. San Francisco, CA, USA, 2016, pp. 2369–2372.

[15] R. Duan and N. F. Chen, "Unsupervised feature adaptation using adversarial multi-task training for automatic evaluation of childrens speech," *Proc. Interspeech 2020*, pp. 3037–3041, 2020.

[16] C.-F. Liao, Y. Tsao, H.-Y. Lee, and H.-M. Wang, "Noise adaptive speech enhancement using domain adversarial training," *arXiv preprint arXiv:1807.07501*, 2018.

[17] E. W. Forgy, "Cluster analysis of multivariate data: efficiency versus interpretability of classifications," *biometrics*, vol. 21, pp. 768–769, 1965.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[19] X. Li, Z. Wu, H. M. Meng, J. Jia, X. Lou, and L. Cai, "Phoneme embedding and its application to speech driven talking avatar synthesis." in *INTERSPEECH*, 2016, pp. 1472–1476.

[20] R. L. Weide, "The CMU pronouncing dictionary," *URL: http://www.speech.cs.cmu.edu/cgibin/cmudict*, 1998.

[21] C. K. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, "A scalable noisy speech dataset and online subjective test framework," *Proc. Interspeech 2019*, pp. 1816–1820, 2019.