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

A neural network-based noise compensation method for pronunciation assessment

Binghuai Lin[⋆] , *Liyuan Wang*[⋆]

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

Abstract

Automatic pronunciation assessment plays an important role in computer-assisted pronunciation training (CAPT). Goodness of pronunciation (GOP) based on automatic speech recognition (ASR) has been commonly used in pronunciation assessment. It has been found that GOP normally shows deteriorating performance under noisy conditions. Traditional noise compensation methods, which compensate distorted GOP under noisy situations based on the Gaussian mixture model (GMM) or other simple mapping functions, ignore contextual influence and phonemic attributes of the utterance. This usually leads to a lack of robustness with changed conditions. In this paper, we adopt a bidirectional long short-term (BLSTM) network combining phonemic attributes to conduct the compensation for distorted GOP under noisy conditions. We evaluate the model performance based on English words recorded by Chinese learners in clean and noisy situations. Experimental results show the proposed model outperforms the traditional baselines in Pearson correlation coefficient (PCC) and accuracy for pronunciation assessment under various noisy conditions.

Index Terms: GOP compensation, pronunciation assessment, BLSTM, contextual influence, phonemic attributes

1. Introduction

For English-as-second language (ESL) learners, real-time and automatic pronunciation feedback is of great importance to improve English proficiency. The typical approach to tackle this challenge is by the CAPT system with automatic pronunciation assessment as the first and essential step to fulfill the goal [1].

Features used for automatic pronunciation assessment are usually extracted from an ASR model. 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]. Some recent work proposed improved methods for the GOP computation considering factors such as HMM state transition probabilities [8][9].

Compared with traditional speech scoring by human raters, automatic pronunciation assessment has proven to be effective and efficient. However, it is still of great challenge for feature extraction utilizing ASR in the noisy 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 [1].

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 [10]. Some method proposed a model that maps both clean inputs and their noisy counterparts onto the same points in the representation space and proved to be robust [11]. For CAPT systems, GOP plays an important role 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 showed the improvement of performance in a noisy classroom environment [12]. However, the performance of SPLICE will deteriorate for input with unknown noise types. The study proposed teacher utterance-based GOP (TGOP) and GOP like (GL) scores, namely two variants of GOP, and revealed that the correlation coefficient between GOP scores and teachers' ratings had been improved under all noise conditions [13]. It usually depends heavily on the teachers' utterances corresponding to the learners' sentences. These previous studies learn simple mapping functions between the deteriorated GOP and raw GOP of phonemes and may not handle complex changes under noisy conditions.

It has been found the pronunciation of a word or subword unit such as a phone depends heavily on the context, and it is important to model the context-dependent variations [14]. Previous work for GOP-based phoneme mispronunciation classification adopted independent thresholds for each phoneme. It shows an example that fricatives tend to have lower log likelihoods than vowels, suggesting that a higher threshold should be used for vowels [3]. These findings inspired us to take the contextual influence and phonemic attributes into consideration for phoneme GOP compensation.

This paper proposes a neural network (NN)-based noise compensation model for GOP under noisy conditions. Unlike traditional methods based on GMM or simple mapping functions, we adopt a BLSTM network to model the contextual influence in an utterance. As GOP scores of different phonemes vary differently, we combine the phonemic attributes of the utterance for noisy GOP compensation. Experimental results in automatic pronunciation assessment show the noise robustness of the proposed network under different noisy conditions and its superiority to the baselines in PCC and accuracy.

[⋆] equal contribution

http://dx.doi.org/10.21437/Interspeech.2021-843

2. Related work

SPLICE makes the non-linear transformation from distorted feature y to its corresponding clean feature x by probabilistic summation of piecewise linear transformation based on GMM [15][16]. An estimate $\hat{x}$ for x is obtained as Eq. (1):

$$\hat{x} = \sum_k P(k|y) A_k y \tag{1}$$

where A_k is the linear transformation matrix, and k is the index of the GMM component based on noise features such as signal-to-noise ratio (SNR). $P(k|y)$ is calculated by GMM based on the distorted features. The transformation A_k is estimated based on the weighted minimum mean square error criterion defined in Eq. (2):

$$A_k = \underset{A_k}{\operatorname{argmin}} \sum_i P(k|y)(x_i - A_k y)^2 \tag{2}$$

As SPLICE doesn't consider the contextual and phonemic influence of an utterance under noisy conditions, an improved model based on the neural network is proposed in this paper.

3. Proposed model

The NN-based model for GOP compensation under noisy conditions is shown in Figure 1. The network takes distorted phoneme GOP scores as input and outputs compensated GOP features. Other noise features and phonemic attributes are treated as input. A BLSTM network models the contextual influence of phonemes.

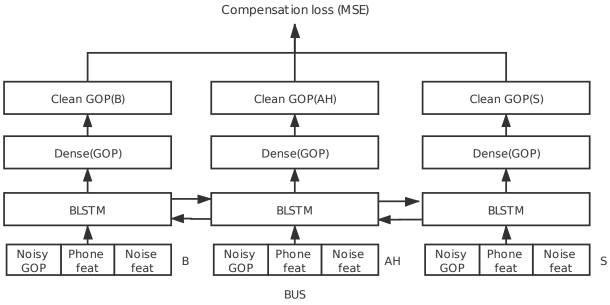

Figure 1: *NN-based model for GOP compensation*

3.1. Features for NN-based GOP compensation method

We focus on the compensation of distorted GOP scores. GOP score is defined as Eq. (3) [3]:

$$\mathrm{GOP}(p) = \frac{|\log(P(p|o^p)|}{\mathrm{NF}(p))} = \frac{|\log(\frac{P(o^p|p)P(p)}{\sum_{q\in Q} P(o^q|q)P(q)})|}{\mathrm{NF}(p)} \tag{3}$$

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 the prior probability of phoneme p. The GOP scores are calculated based on a DNN-HMM-based ASR system.

Besides GOP scores, some other features for compensation are considered as well. First, features for noise representations are included as input. Signal-to-noise ratio (SNR), which is defined as the ratio of the signal power to the noise power, has been regarded as an important feature for quantifying the degrees of noise. Noise at different SNRs has different impacts on GOP scores as shown in Figure 2, where the x-axis represents the sample indexes, and the y-axis shows the GOP values. From the figure, it is clear that GOP scores of phoneme 'AA' and 'L' decrease with decreasing SNR values. We compute the power features, i.e., the transmission rate of energy, for each 10ms frame with a 10ms shift. Based on the power and the duration of each phoneme in the word, we derive multiple noise features as shown in Table 1. Second, as different phonemes may be influenced by noise at various degrees, we take into account some phonemic attributes. As the phoneme pronounces differently depending on its positions in the word [17], we denote phoneme positions by 'B', 'I', 'E', 'S', which represent the beginning, middle, ending positions in a word as well as single-phoneme words. It has been found vowels and consonants behave differently when influenced by noise as well [18]. we use 'C' and 'V' to represent phoneme classes of consonants and vowels separately. We also employ independent numerical representations for each phoneme. We encode these phoneme properties into numerical vectors, which are called positional, phonemic class, and phoneme embedding [19]. We combine these feature representations with phoneme GOPs as the input of the network.

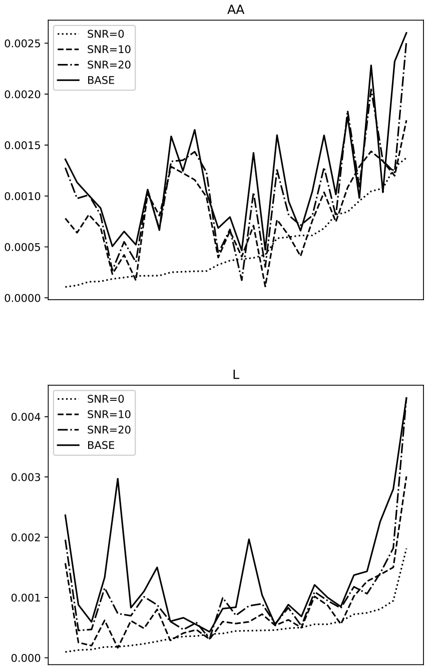

Figure 2: *GOP scores of phoneme 'AA' and 'L' at different SNRs*

3.2. Network optimization

The network takes deteriorated GOP scores as input and outputs compensated GOP ones. The training data for the network is composed of paired deteriorated and raw GOP scores of phonemes. Different kinds of noise (e.g., bubble noise and white noise) at different SNRs are added to clean audios. Then, the GOP scores are computed for both clean and noisy audios

Table 1: *Different noise features*

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

based on Eq. (3). The training loss is defined as Eq. (4), where n is the number of utterances and m is the number of phonemes in one utterance. p_i^j and y_i^j are the jth predicted GOP and the raw GOP scores under clean conditions of the phonemes in ith utterance.

$$L_{\text{GOP}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}(p_i^j - y_i^j)^2 \qquad (4)$$

3.3. Application of the proposed model to pronunciation assessment

We apply the BLSTM-based GOP compensation model to word-level pronunciation assessment. The whole procedure is shown in Figure 3. The GOP compensation model takes distorted GOP scores as input and outputs the compensated GOP ones. The noise robust GOP scores are then fed into the pronunciation assessment model as input features to obtain final pronunciation scores.

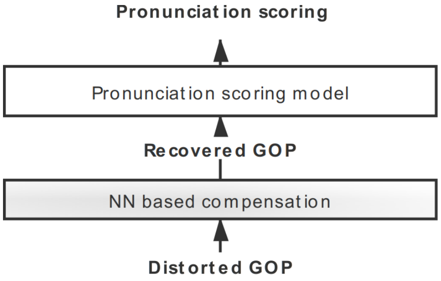

Figure 3: *Procedure of the compensation*

4. Experiment

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 the 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 in Figure 4. 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 [13]. 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 with low SNRs from one online pronunciation learning application, whose labels are annotated by the same method.

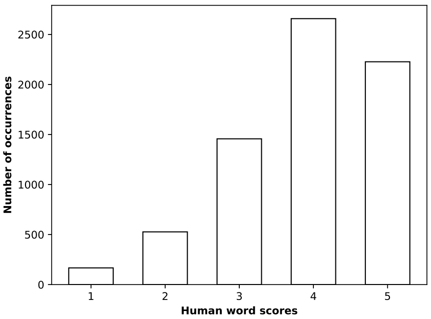

Figure 4: *Distribution of human labels of clean audios*

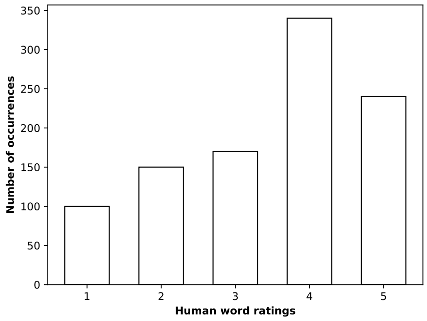

Figure 5: *Distribution of human labels of audios under real noisy situations*

4.2. Experimental setup

The dimensionality of input for the NN-based GOP compensation module is 23, with 1 for GOP, 8 for noise features, 4 for phoneme positions, 8 for phoneme embedding, and 2 for phonemic class. The total types of phoneme positions and phonemic classes are 4 and 2 as mentioned in section 3. The total number of phoneme embedding parameters is 8×39, indicating a total of 39 different phonemes. The dimensionality of BLSTM parameters is 23×10, and the dimensionality of the GOP dense layer is 10×1.

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

Model	PCC	ACC
No compensation	72.1	85.3
SPLICE [15]	73.3	85.1
TGOP [13]	71.5	84.1
Ours	73.1	85.6

In our paper, the scoring model is a hierarchical network combining scoring at multi-granularity, which has a similar network structure as previous work [22]. We make a comparison with compensation methods such as SPLICE [15] and TGOP [13]. In our paper, input features for GMM in SPLICE are noise features defined in Table 1. The number of GMM components is 7, which is carefully determined based on the cluster analysis. The features for non-linear transformation in SPLICE are the average GOP scores of each phoneme in an utterance. The TGOP-based method utilizes aligned boundaries obtained from the forced-alignment considering phonemes in the corresponding teacher's utterance for the learner's GOP computation, where GOP scores are calculated based on the improved method with the best performance [8]. In this paper, we treat the corresponding native references utterance as the teacher's utterances.

4.3. Results

The baseline results are taken from the model for automatic pronunciation assessment introduced in Section 3.3 without the addition of the NN-based compensation module (No compensation). Comparison with other compensation methods such as SPLICE and TGOP is presented here as well. Comparison results under clean conditions are shown in table 2. From the results, we can see all of them achieve comparable performance under clean conditions. Results with distortion of babble noise and white noise are shown in table 3 and 4. From the results, we can see the performance of the pronunciation scoring model deteriorates dramatically with the addition of noise. By compensating for distorted GOP based on SPLICE, TGOP and the proposed network, the performance of the pronunciation network recovers to some extent. The SPLICE and TGOP methods achieve nearly the same performance. The proposed network shows superiority to the other two methods in two of the three SNRs. It is difficult to fully recover the performance due to the deteriorating performance of ASR with more deleted, substituted, and inserted words.

Besides testing the performance with a single type of noise, we also experiment with mixed noise consisting of equivalent white and babble noise at different SNRs. Results are shown in table 5. From the results, we can see the proposed compensation network can still improve the performance and outperform other methods in both PCC and accuracy.

To test the robustness of the proposed model under real noisy situations, we conduct the experiments based on the data with low SNRs of less than 10 as well. The results are shown in Table 6. We can see SPLICE is less robust in real noisy situations, which may result from different noise distributions between the testing data and the training data. Our proposed model proves to be most robust in the unknown real noisy environment.

Table 3: *Performance under babble noise*

		0dB	10dB	20dB
No compensation	PCC	53.1	62.6	67.9
	ACC	58.7	71.7	82.5
SPLICE [15]	PCC	54.6	61.8	67.3
	ACC	60.8	**74.5**	83.9
TGOP [13]	PCC	53.5	62.1	66.8
	ACC	58.4	72.1	83.4
Ours	PCC	**59.3**	**64.2**	**69.1**
	ACC	**68.0**	74.2	**83.8**

Table 4: *Performance under white noise*

		0dB	10dB	20dB
No compensation	PCC	44.8	51.1	62.7
	ACC	45.5	61.7	76.3
SPLICE [15]	PCC	45.6	53.1	64.5
	ACC	50.9	**66.0**	77.5
TGOP [13]	PCC	43.3	49.3	61.3
	ACC	43.3	61.2	76.1
Ours	PCC	**46.2**	**54.4**	**65.1**
	ACC	**53.8**	65.1	**78.1**

Table 5: *Performance under mixed noise*

		0dB	10dB	20dB
No compensation	PCC	47.0	55.9	61.9
	ACC	51.2	65.2	81.9
SPLICE [15]	PCC	46.5	54.1	61.3
	ACC	58.0	**69.0**	83.0
TGOP [13]	PCC	47.5	**56.1**	62.2
	ACC	54.5	67.3	83.2
Ours	PCC	**51.7**	55.3	**63.3**
	ACC	**58.2**	67.6	**83.3**

Table 6: *Performance under real noise*

	PCC	ACC
No compensation	70.4	62.1
SPLICE [15]	70.8	60.1
TGOP [13]	71.5	63.2
Ours	**72.1**	**66.5**

5. Conclusion

This paper proposes an NN-based network for the compensation of distorted GOP under noisy conditions. To consider the contextual and phonemic influence of an utterance, we combine phonemic features as input and utilize a BLSTM network. Experimental results in different synthetic and real noisy situations demonstrate the improved performance of the proposed network. In this work, the proposed model focuses on the word-level pronunciation assessment. We will investigate the application of the compensation network to automatic pronunciation assessment of both words and sentences in the future.

6. References

[1] L. Chen, "Audio quality issue for automatic speech assessment," in *International Workshop on Speech and Language Technology in Education*, 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*. IEEE, 1997, vol. 2, 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] S. Sudhakara, M. K. Ramanathi, C. Yarra, and P. K. Ghosh, "An improved goodness of pronunciation (GoP) measure for pronunciation evaluation with DNN-HMM system considering HMM transition probabilities," in *INTERSPEECH*, 2019, pp. 954–958.

[9] J. Shi, N. Huo, and Q. Jin, "Context-aware goodness of pronunciation for computer-assisted pronunciation training," *arXiv preprint arXiv:2008.08647*, 2020.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] L. R. Bahl, P. V. deSouza, P. Gopalakrishnan, D. Nahamoo, and M. Picheny, "Context dependent modeling of phones in continuous speech using decision trees," in *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*, 1991.

[15] J. Droppo, L. Deng, and A. Acero, "Evaluation of splice on the aurora 2 and 3 tasks," in *Seventh International Conference on Spoken Language Processing*, 2002.

[16] T. Kai, M. Suzuki, K. Chijiiwa, N. Minematsu, and K. Hirose, "Combination of splice and feature normalization for noise robust speech recognition," *Journal of Signal Processing*, vol. 16, no. 4, pp. 323–326, 2012.

[17] J. S. Bowers, N. Kazanina, and N. Andermane, "Spoken word identification involves accessing position invariant phoneme representations," *Journal of Memory and Language*, vol. 87, pp. 71–83, 2016.

[18] S. A. Phatak and J. B. Allen, "Consonant and vowel confusions in speech-weighted noise," *The Journal of the Acoustical Society of America*, vol. 121, no. 4, pp. 2312–2326, 2007.

[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.

[22] B. Lin, L. Wang, X. Feng, and J. Zhang, "Automatic scoring at multi-granularity for l2 pronunciation," *Proc. Interspeech 2020*, pp. 3022–3026, 2020.