Normalization of GOP for Chinese Mispronunciation Detection

Wenwei Dong* and Yanlu Xie*
* Beijing Advanced Innovation Center for Language Resources, Beijing Language and Culture University, Beijing, China
E-mail: dongwenwei_blcu@163.com
E-mail: xieyanlu@blcu.edu.cn

***Abstract*—Goodness of Pronunciation (GOP) is a kind of Computer-Assisted Pronunciation Training (CAPT) technique that can provide language learners with scoring feedback, and its accuracy easily suffers from the performance of model alignment and phone classification. In order to reduce the influence of those aspects, this paper proposes two ways to normalize GOP scores. The first is to separate the GOP calculation of Chinese Initials and those of Chinese Finals. The second is to use the corresponding native pronunciation score as a template to scale the non-native one. In 2-hours test set of Japanese speaking Chinese corpus, the experiment results show the average relative improvement of Diagnose Accuracy (DA) in the approach one is 16.9%, and 28.7% in scaling approach comparing to the traditional scoring method. The combination of those two methods achieves the best performance. The result is 35.9% of average relative improvement. Experimental results demonstrate the effectiveness of the two methods.**

I. INTRODUCTION

With the acceleration of globalization, language learning becomes more and more important. CAPT has been playing a significant role in improving language learning. Scoring methods as a key component in CAPT have drawn a lot of attention.

Scoring methods of early stage mostly were based on template [1−4]. Teacher and student read the same script, then the score can be obtained by computing the distance between the acoustic features of their pronunciation. Later on, as the development of Automatic Speech Recognition (ASR), Hidden Markov Models (HMMs) was also applied to CAPT. Kim [5] compared three HMM-based scoring methods and found that log-posterior probability scores have the highest correlation with human scores than log-likelihood ratio and segment duration scores. The utterance or word level scoring is hard for language learners to correct their pronunciation. Witt and Young [6] introduced the GOP method for phone level scoring, this method widely used in pronunciation evaluation. Some variants of posterior probability were also explored. Zhang et al. [7] used scaled log-posterior probability (SLPP) and weighted phone SLPP improved detection result. In the age of DNN as acoustic model, Hu [8] used target phone's posterior and most competitive phone's ratio as the score of pronunciation. And a lot of works tried to use GOP score as input feature and add a rescoring or verification process to get further improvements [9-12].

Most phone level scores today are based on ASR frameworks, and even if not based on ASR, they still need phones boundary provided by ASR [13]. So it has the following weaknesses: first, the accuracy of model alignment has a great influence on the scores. Yuan's best performance of phone alignment task achieves 93.1% accuracy within 20ms in ASCCD corpus [14]. His system is tested in native speech, but in CAPT systems, we need to handle the situation of the non-native test set. Due to the mismatch between native and non-native corpus, the alignment of non-native speech thus become worse. Second, the quality of acoustic model has a great influence on GOP. If the training of acoustic model is not good, the result of scoring is unreliable. The acoustic model also has different confidence scores when classifying different phones. [6] set different thresholds for different phones to decide correct pronunciation or not.

For those weaknesses, we propose two methods to normalize GOP scores to get a higher agreement with human diagnoses. First, in Mandarin, most syllables consist of an Initial and a Final. For syllables without Initials, we extend the beginning of the syllable when training the acoustic model. We separate Initials and Finals to evaluate. The GOP of Initial (Final) is target Initial and most competitive Initial (Final) posterior probability ratio. Second, we use GOP scores of native speech as a template to scale the non-native. The paper is organized as follows: section 2 presents two way of improving GOP measures and the frameworks of GOP. Section 3 introduces the experiment corpus and setup. Section 4 shows the experiment results and discussions, and the conclusions are drawn in Section 5.

II. IMPROVED MISPRONUNCIATION DETECTION

In this section, we briefly review the traditional GOP method and introduce two improvement methods, then introduce the framework of generating posterior probability.

A. GOP Computing Methods

Witt et al proposed GOP method of phone level scoring which is text-dependent. For the acoustic segment X, target phone p,

$$\mathrm{GOP}(p) = \frac{1}{d} log \frac{P(X|p)P(p)}{\sum_{\{q \in Q\}} P(X|q)P(q)}? \quad (1)$$

where d is the number of frames, *P(X|p)* is the likelihood of X corresponding to phone p. Q is the set of phones. In DNN-HMM based system, assuming all phone's prior probability are equal and the sum of all phone posterior probability can be

approximated by its maximum [6], in this paper we use the GOP,

$$\mathrm{GOP}(p)=\frac{1}{d}log\frac{P(p|X)}{max_{\{q\in Q\}}P(\mathrm{q}|X)+P(p|X)} \quad (2)$$

P(p|X) is the posterior probability of phone p that generates from acoustic model. Then we set a threshold to make the final decision,

$$\mathrm{GOP}(p)>k\begin{cases}\text{yes,} & \text{correct pronunciation}\\ \text{no,} & \text{mispronunciation}\end{cases} \quad (3)$$

those methods can be affected by the alignment result of a model. In practice, we can adjust the threshold for language learners with different level. The main idea of GOP is to use classifier's confidence score as the score of pronunciation.

B. Separating Initials and Finals

In Mandarin, most syllables consist of an Initial and a Final. For syllables without initials, we extend the beginning of the syllable when training the acoustic model. According to annotation and education experience, L2 learners are prone to confuse initials with initials and finals with finals. Traditional GOP method uses the average of frames posterior probability as the phone's posterior probability, but the alignment of GMM-HMM not accurate enough, sometimes when we evaluate initials, the most competitive phone is not initials and a phone's score will be influenced. An example of alignment shows in figure 1.

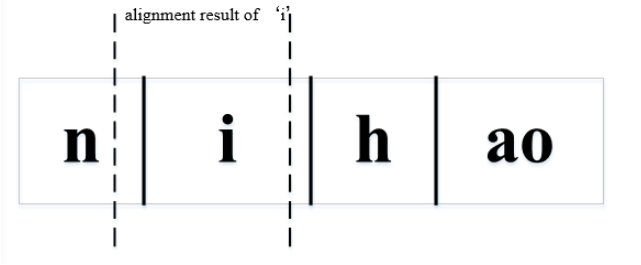

Fig. 1 An example of alignment result

When we evaluate non-native speaker's utterance "nihao", the black solid line is the actual boundary of those initials and finals. The dotted line is the alignment result that generated from the model, from the figure 1 we can see, if we want to get the GOP of Final 'i' in the traditional way, we need average the all frames posterior probability on the basis of the boundary as phone 'i' posterior probability. Due to the alignment error, the posterior probability of 'i' is influenced by the Initial 'n'. Chinese syllables are alternating with Initial and Final, so when we calculate the Initial's score, the most competitive phone posterior probability should be Initial. The same goes for Final.

$$\mathrm{GOP}_1(p)=\frac{1}{d}log\frac{P(p|X)}{max_{\{p\in Qi?q\in Qi:q\in Qf\}}P(q|X)+P(p|X)} \quad (4)$$

where *Qi* means phone set of initial, *Qf* means phone set of final.

C. Scaling with Native Speech GOP scores

In the light of research by Witt [6], a single threshold for all phones is inappropriate. For example, fricatives tend to have lower log-likelihood than vowels. They try to use different thresholds for phones. We also found different phones will have different posterior probability. In GOP method, the main idea is using native speakers as the golden speakers to evaluate pronunciation of non-native speaker, so the native speaker's GOP of each phone should be close to 1. But in practice, native speaker GOP scores are often not close to 1. For example, if the native speaker's score is 0.6 in a phone, then non-native's score is good enough in 0.5. In my research, we use the average of native speech each phone's score as the template to normalize the non-native.

$$\mathrm{GOP}_2(p)=\frac{\mathrm{GOP}_{\text{non-native}}(p)}{\mathrm{GOP}_{\text{native}}(p)} \quad (5)$$

p is the target phone. We use the mean GOP score of different native speakers as a template. The normalization method can partially result in different phonemes that require different thresholds, as well as imperfect model training.

D. Framework of GOP method

The front-end feature extractor converts waveform to filterbank, and that feature are used in acoustic model to generate posterior probability and alignment model to get alignment result of phones. Based on those results, the different GOP measurements can be used to calculate the score of pronunciation. Finally, a threshold can be applied to make the final decision. The choice of the threshold depends on the level of strictness. The framework is shown in figure 1.

III. EXPERIMENTS

A. Speech Corpus

Experiment corpus consists of two parts, the native speech database is provided by the Chinese National Hi-Tech Project 863 for Mandarin continuous speech recognition of large vocabulary system development [15]. It contains 166 speakers about 100 hours. We divide it into the training set about 70 hours and development set about 30 hours, and no speakers overlap. We also use 3600 sentences of native Chinese corpus to test GOP algorithm. The non-native speech corpus is BLCU inter-Chinese speech corpus [16]. It has 19 speakers and each speaker has 301 sentences. We add 12 speakers of it as the training set to reduce the mismatch between native and non-native dataset, and 7 speakers corpus as the test set. The test set has been annotated at the phone level. The details are shown in Table 1.

TABLE 1
JAPANESE L2 INTER-CHINESE CORPUS

Corpus	Description
Text	Conversational Chinese 301
Speaker	7 females
Number of utterances	1899
Number of phones	26431
Average length per utterance	14
Number of annotators	6

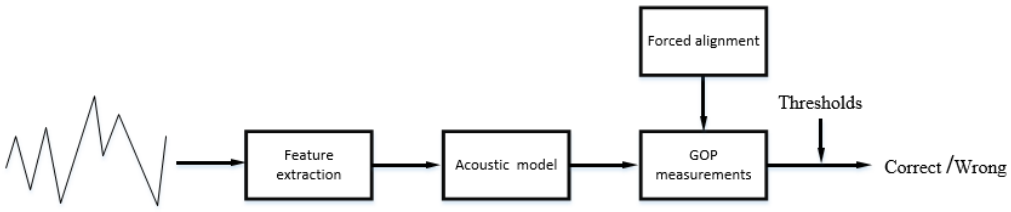

Fig. 2 The framework of mispronunciation detection

B. *Evaluation Metrics*

There are 4 evaluation indicators:

- False Acceptance Rate (FAR): the percentage of mispronunciation phones that are accepted as correct.
- False Rejection Rate (FRR): the percentage of correctly pronounced phones that are rejected as mispronunciation.
- Diagnostic Accuracy (DA): the percentage of correctly detected.
- The Detection Cost Function (DCF):

$$DCF(\tau) = C_{MISS}FRR(\tau)P_{Target} + C_{FA}FAR(\tau)(1 - P_{Target}) \quad (9)$$

where τ is the threshold of GOP. C_{MISS} is the cost of false rejection, C_{FA} is the cost of false acceptance. P_{target} is a prior probability and in practical application, FAR is more important than FRR, because if too many correct pronunciations are rejected as mispronunciations. it will give users a bad experience.

C. *Experiment setup*

The dimension of input feature is 27 including 23-dimensional Fbank, 3-dimensional pitch, and 1-dimensional energy, we use Cepstral Mean and Variance Normalization (CMVN) to reduce speaker difference.

Kaldi toolkit is used to train Gaussian Mixture Modeling (GMM), Hidden Markov Models (HMMs) and Time-Delay Neural Network (TDNN). TDNN has 6 hidden layers and each layer has 850 nodes. The alignments generated by GMM-HMM model.

IV. RESULTS AND DISCUSSIONS

A. *The traditional GOP method for native speaker*

We have tested the traditional GOP method in native speaker corpus. The DA is shown in Figure 3.

X-axis represents different thresholds of GOP. y-axis represents the DA, we can know that even for native speaker, the DA only can achieve 92% in the threshold of 0.1. If we use this method to score non-native, the results will be worse.

Figure 4 shows the GOP scores of different phones in native corpus. The x-axis represents different phones. The y-axis represents the GOP scores. We calculate average score of all speakers in different phones. Figure 4 shows a big difference between GOP scores of phones. Compound Finals tend to have lower scores than other phones. If we use the GOP scores of native speaker in each phone to scale non-native, the phone difference can be eliminated. So the same threshold for all phone's score can be applied.

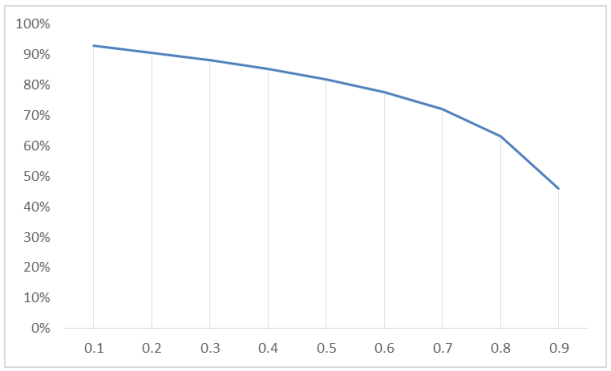

Fig. 3 The DA of native Chinese corpus

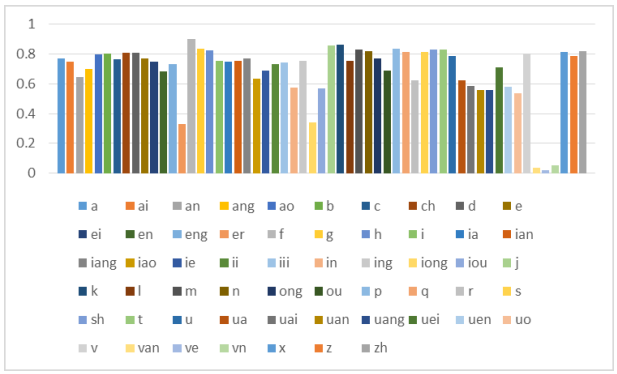

Fig. 4 The GOP scores of different phones

B. *Different Methods of GOP*

Based on TDNN acoustic model, we use three ways to calculate GOP, as the formula in (2), (4) and (5). GOP is the posterior probability ratio of the target phone and most competitive phone. GOP1 is the posterior probability ratio of target initials (finals) and most competitive initials (finals).

GOP2 is the ratio of same phone's GOP from native and non-native. GOP3 is the method that combining formula (4) and (5). The DA is shown in Figure 5.

The result shows GOP3 has the best performance, and it is suitable for high-level language learners because of its robustness. The DA of GOP2 improve 26.08% than traditional GOP at threshold of 0.9, its average improvement is 4.53% in all thresholds. The relative improvement is 16.9%. GOP1 separate initials and finals to evaluate the GOP, it indeed can reduce the influence of alignment error and improve the DA. Acoustic models have different confidence scores in phone classification for native speakers, the average improvement of GOP2's DA is 7.71% than GOP, which indicates that using native's GOP as the template to scale L2's can reduce the influence of it. The relative improvement is 28.7%. GOP3 combined those two methods and get the best performance, its average improvement is 9.62 than GOP. The relative improvement is 35.9%.

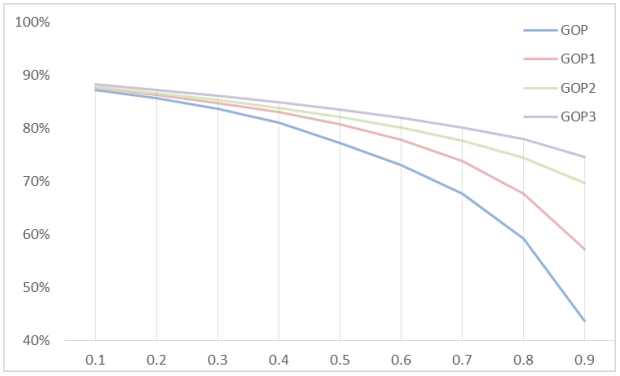

Fig. 5. The DA of different GOP

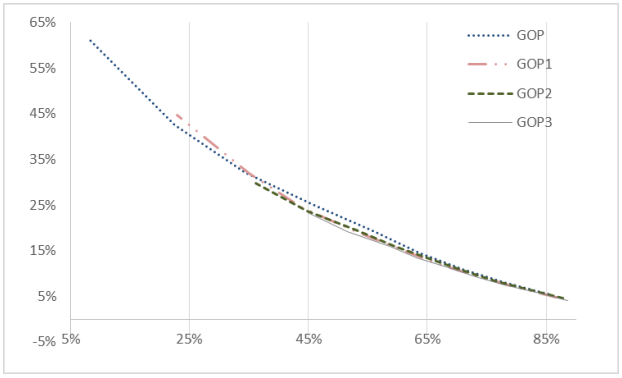

Fig. 6 the ROC of different GOP methods

FAR and FRR need to be a trade-off. From Figure 6. When FAR bigger than 33%, the GOP result is worse than others. GOP1 and GOP2's ROC curve are shorter than GOP's, it means two new methods are more stable than the traditional method in all thresholds. And GOP3 is the shortest, thus it's the best method.

C. *DCF of Different Methods*

The total score is 1, we set τ =0.6, because most tests use 0.6 as the passing score and p_{target} = 0.7. FRR is the percentage of mistake making by model. In practice, we care more about FRR. Table 2 shows the DCF of different methods.

TABLE 2
THE *DCF OF DIFFERENT METHODS.*

Methods	DCF (%)
GOP	31.30
GOP1	29.28
GOP2	29.05
GOP3	28.39

Detection cost of GOP3 is lower than others. The DA of GOP in τ =0.6 is 73.14%, GOP1 is 77.86% and GOP2 is 80.11%. GOP3 is 82.09%.

Thresholds mean the level of strictness. The higher the threshold, the stricter it is. From the experiment results, the GOP method is more suitable for beginner because of its bad performance in the high threshold, GOP3 is more robust in all threshold and is suit for all language learners.

V. CONCLUSIONS

This paper proposed two methods to normalize GOP scores to get a higher agreement with human diagnoses: first is to separate initials and finals to evaluate, initials' (finals) GOP is target initial and most competitive initial's (finals) posterior probability ratio. This method can reduce the influence of bad alignment of phones. Second, we use native speech's GOP as a template to scale the L2. This method can make up for the insufficiency of the acoustic model training to a certain extent, and reduce the calculation amount of different phones using different thresholds.

Experiment results show separating initials and finals to calculate GOP is better than baseline GOP, it can reduce the influence of alignment. The scaling method can further improve the performance of mispronunciation detection, and the combination of those two methods get the best performance. GOP3 is relatively steady than other methods and have the lowest DCF, it robustly fit for L2 learners at all levels.

ACKNOWLEDGMENT

This work is supported by National social Science foundation of China (18BYY124), Advanced Innovation Center for Language Resource and Intelligence (KYR17005), the Fundamental Research Funds for the Central Universities (18YJ030004), the Graduate Innovation Fund of Beijing Language and Culture University (19YCX130), and the project of "Intelligent Speech technology International Exchange". Yanlu Xie is the corresponding author.

REFERENCES

[1] J. Tepperman, and S. Narayanan. "Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners." 2005.

[2] A. Lee, and J. Glass. "Pronunciation assessment via a comparison-based system." *Speech and Language Technology in Education*. 2013.
[3] C. L. Rogers, M. D. Jonathan, and D. Gladys. "Intelligibility training for foreign-accented speech: A preliminary study." *The Journal of the Acoustical Society of America* 96.5 (1994): 3348-3348.
[4] H. S. Wohlert, "Voice input/output speech technologies for German language learning." Die Unterrichtspraxis/Teaching German 17.1 (1984): 76-84.
[5] K. Yoon, H. Franco, and L. Neumeyer. "Automatic Pronunciation Scoring of Specific Phone Segments for Language Instruction", *European Conference on Speech Communication & Technology DBLP*, 1997.
[6] S. M. Witt, and S. J. Young. "Phone-level pronunciation scoring and assessment for interactive language learning", *Speech Communication*, vol. 30, no. 2, pp. 95-108, 2000.
[7] C. Huang, F. Zhang, F K. Soong, et al. "Mispronunciation detection for Mandarin Chinese", *IEEE International Conference on Acoustics*. IEEE, 2008.
[8] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, vol. 67, pp. 154–166, 2015.
[9] G. Huang, J. Ye, Z. Sun, Y. Zhou, Y. Shen and R. Mo, "English mispronunciation detection based on improved GOP methods for Chinese students", *2017 International Conference on Progress in Informatics and Computing (PIC)*, Nanjing, pp. 425-429, 2017.
[10] M. Nicolao, A. V. Beeston, and T. Hain. "Automatic assessment of English learner pronunciation using discriminative classifiers", *IEEE International Conference on Acoustics*, IEEE, 2015.
[11] W. Li, K. Li, S. M. Siniscalchi, N. F. Chen, and C. Lee, "Detecting mispronunciations of l2 learners and providing corrective feedback using knowledge-guided and data-driven based decision trees," in *INTERSPEECH*, 2016.
[12] C. Huang, F. Zhang, F K. Soong, et al. "Mispronunciation detection for Mandarin Chinese", *IEEE International Conference on Acoustics*. IEEE, 2008. H. Huang, H. Xu, Y. Hu, et al. "A transfer learning approach to goodness of pronunciation based automatic mispronunciation detection", *Journal of the Acoustical Society of America*, 142(5):3165, 2017.
[13] A, Lee, N. F. Chen, and J. Glass. "PERSONALIZED MISPRONUNCIATION DETECTION AND DIAGNOSIS BASED ON UNSUPERVISED ERROR PATTERN DISCOVERY." *IEEE International Conference on Acoustics* IEEE, 2016.
[14] J. Yuan, N. Ryant, and M. Liberman.(2014, May). Automatic phonetic segmentation in Mandarin Chinese: boundary models, glottal features and tone. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* pp. 2539-2543. 2014.
[15] S. Gao, et al. "Update Progress Of Sinohear: Advanced Mandarin LVCSR System At NLPR." In proc. *ICSLP*, 2000.
[16] W. Cao, et al. "Developing a Chinese L2 speech database of Japanese learners with narrow-phonetic labels for computer assisted pronunciation training", in *INTERSPEECH*, 2010.