A NOVEL DISCRIMINATIVE METHOD FOR PRONUNCIATION QUALITY ASSESSMENT

Junbo Zhang, Fuping Pan, Bin Dong, Yonghong Yan

Key Laboratory of Speech Acoustics and Content Understanding,
Institute of Acoustics, Chinese Academy of Science, China

ABSTRACT

This paper presents a novel method for automatic pronunciation quality assessment. Unlike the traditional "Goodness of Pronunciation" (GOP) method, we judged utterance's pronunciation quality directly by a discriminative method. Under this novel framework, we also designed an algorithm to calculate the assessment confidence. We decoded the student's utterance for two passes. The first-pass decoding was just for getting the phone time points, and the second-pass decoding was for differentiating the pronunciation quality for each triphone. In the second-pass decoding, we used a specially trained acoustic model (AM), where the triphones in different pronunciation qualities were trained as different units. The confidence of the phone-level scoring was also calculated, and the low confidence phone-level scores were excluded in calculating the word-level score. The experimental results shows that the scoring performance was increased significantly compared to the traditional GOP method.

Index Terms— Pronunciation assessment, automatic scoring, acoustic model

1. INTRODUCTION

The technology of the automatic pronunciation quality assessment has been proved effective [1], which can be used in two scenes: the first scene is using it in computer-assisted language learning (CALL) systems, telling the students where their pronunciation is good, and where is not good [2]; the second scene is automatic scoring in large scale oral exams [3, 4]. Automatic scoring can overcome many disadvantages of manual scoring, such as high cost and lack of stability. For automatic scoring, the machine score should approximate to the manual score, that is, the absolute difference between the machine score and manual score should be minimized. A binary "right" or "wrong" score is usually not enough, most exams need a multi-level score, for example, a score in the set {1,2,3}.

2. RELATION TO PRIOR WORK

In the early time, there were many studies in the field of automatic pronunciation quality assessment [5]. But one of the most influential achievement in this field is the Goodness of Pronunciation (GOP) [6] method presented by Witt et al. in 1997. At present, GOP has been widely considered to be an effective method [7]. Even in the recent years, most studies in this field still focused on how to use or improve GOP [8, 9, 10].

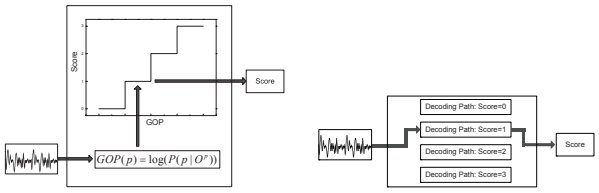
This work is partially supported by the National Natural Science Foundation of China (No. 10925419, 90920302, 10874203, 60875014, 61072124, 11074275, 11161140319).

The GOP of phone p is defined to be the logarithm of the posterior probability $P(p|O_p)$ that the speaker uttered phone p given the corresponding acoustic segment O_p as Eq.1:

$$\begin{aligned} GOP(p) &\doteq \log(P(p|O_p)) \\ &= \log(\frac{P(O_p|p)}{\sum_{q\in Q} P(O_p|q)}) \end{aligned} \quad (1)$$

where Q is the set of all phones and O_p is the corresponding acoustic segment.

However, GOP has a shortcoming. As shown in Eq.1, GOP reflects how much the student's utterance match the acoustic model (AM), but it does not reflect the characteristic of different pronunciation quality utterances, that is, GOP tells "how similar are the student's utterance and the training utterances", but it does not tell "which score-level utterances are the student's utterance most similar to". To overcome this shortcoming, this study tried to differentiate different pronunciation quality utterances directly by a discriminative method. Unlike the GOP method, it does not map the posterior probabilities into scores, but differentiates different quality utterances directly. We named this new method "two-pass discriminative assessment" (TPDA). Fig.1 shows the different ways of phone-level scoring with the GOP method and the TPDA method.



(a) GOP method (b) TPDA method

Fig. 1. Phone-level scoring in GOP and TPDA method.

3. THE METHOD

3.1. Overview

The workflow of the "two-pass discriminative assessment" (TPDA) system is shown in Fig.2. The student's utterance was decoded for two passes with different acoustic models. The first-pass decoding was to obtain each phone's time points by a forced alignment. In this paper, the acoustic model (AM) in the first-pass decoding is referred to as "AM1", which was trained using conventional method [11]. After the time points of each phone were obtained, the utterance of each phone was decoded with a specially trained AM, which is referred to "AM2" in this paper. AM2 was trained with

different quality triphones, that is, the triphones in different pronunciation qualities were trained as different units, and the model was trained in discriminative method to ensure the model has the best discrimination among the triphones whose names were same but pronunciation qualities were different. The detailed training process of AM2 was described in Sect.3.2. The decoding network in the second-pass decoding contained different pronunciation quality triphones, so the phone-level scores can be obtained from the decoding result directly. Because the scores of each triphone may not all equal to the score of the whole word, we calculated the confidence of the phone-level scoring to judge whether a certain phone-level score was credible. Those phone-level scores whose confidences were higher than a threshold were selected to calculate the weighted average as the word-level score.

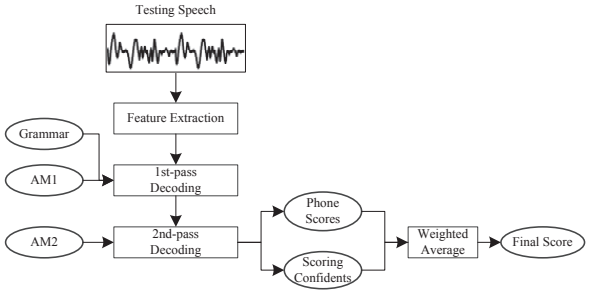


Fig. 2. The TPDA system workflow

3.2. Training AM2

Suppose $\hat{\lambda}$ is the set of all HMM parameters defining AM2. Because the forced alignment has been served by AM1, AM2's sole purpose was to judge the pronunciation quality of each utterance segment within the corresponding phone. To do this, triphones with different pronunciation qualities were treated as different units in training. Assume that there were T triphones and K manual score levels in the training data, there would be $T \times K$ HMMs to train.

To ensure the AM2 have the best discrimination among the different pronunciation quality levels, the posterior probability of each triphone with the correct pronunciation quality level in training data should be maximized, and the Maximum Mutual Information Estimation (MMIE) criterion [12] is especially suited for this maximization. That is, making $\hat{\lambda}$ meet Eq.2:

$$\hat{\lambda} = \arg\max_{\lambda} \mathcal{F}(\lambda) \tag{2}$$

$\mathcal{F}(\lambda)$ is defined as:

$$\begin{aligned}\mathcal{F}(\lambda) &= \frac{1}{R}\sum_{r=1}^{R}\log\left(P(\mathcal{H}_{\mathtt{ref}}^{(r)}|O^{(r)},\lambda)\right) \\ &= \frac{1}{R}\sum_{r=1}^{R}\log\left(\frac{P(O^{(r)}|\mathcal{H}_{\mathtt{ref}}^{(r)},\lambda)P(\mathcal{H}_{\mathtt{ref}}^{(r)})}{\sum_{\mathcal{H}}P(O^{(r)}|\mathcal{H},\lambda)P(\mathcal{H})}\right) \\ &\approx \frac{1}{R}\sum_{r=1}^{R}\log\left(\frac{P(O^{(r)}|\mathcal{H}_{\mathtt{ref}}^{(r)},\lambda)}{\sum_{\mathcal{H}}P(O^{(r)}|\mathcal{H},\lambda)}\right)\end{aligned} \tag{3}$$

where R is the number of observation sequences, $O^{(r)}$ is the $r-$th observation sequence, $\mathcal{H}_{\mathtt{ref}}^{(r)}$ is the corresponding triphone of $O^{(r)}$ with the correct manual score, $\mathcal{H}$ is the all score-level corresponding triphones of $O^{(r)}$.

To obtain the initial model for the iteration in MMIE training, we just need to do single model re-estimation for HMM parameters of each triphone's HMM parameters, without to do the embedded model re-estimation [13], because the time points of each triphone had been obtained in the first-pass decoding. By the single model re-estimation, an initial mode contains $T \times K$ HMMs were trained for T triphones and K manual score levels. And then AM2 was trained by the conventional MMIE training process [12], but the denominator lattice should only contain the corresponding triphones of the training utterances, as shown in Fig.3. Thus the discrimination of the triphones, which have the same name but different pronunciation qualities, was maximized.

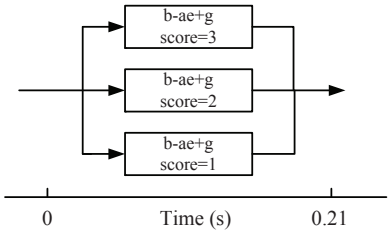


Fig. 3. An example of the denominator lattice in MMIE training

3.3. Decoding to Score in Phone-level

As shown in Fig.2, after the forced alignment, the student's utterance had been cut into several segments, where any segment corresponded a triphone. To calculate the phone-level scores, we processed Viterbi decoding for each segment. The structure of the decoding network was the parallel of the corresponding triphones of the segment, which was similar to Fig.3, but the decoding network did not contain the time information. Then the HMM log-likelihood was calculated for each decoding path, and the score of the segment was obtained from the path whose log-likelihood was highest in the all paths, as shown in Eq.4:

$$PhScore(p) = y \tag{4}$$

where $PhScore_p$ is the score of the phone p, y is the pronunciation quality score with the highest decoding likelihood, as shown in Eq.5:

$$y = \arg\max_{x} \log(P(O_p|p_x)) \tag{5}$$

where O_p is the corresponding acoustic observation sequence of p, p_x is the triphone with the pronunciation quality score x, $P(O_p|p_x)$ is the conditional probability of O_p in p_x.

We also used the posterior probability to evaluate the confidence of the the scoring:

$$\begin{aligned}Conf(p) &\doteq \log P(PhScore(p)|O_p) \\ &= \log(P(p_y|O_p)) \\ &= \log(\frac{P(O_p|p_y)}{\sum_{x\in X}p(O_p|p_x)})\end{aligned} \tag{6}$$

where $Conf(p)$ is the scoring confidence of the phone p, p_y is the triphone p with the pronunciation quality score y, y is calculated from Eq.5, X is the set of every score-level of p.

Eq.6 and Eq.1 have the similar forms, however, their meanings are different: Eq.1 reflects the probability of an utterance correspond a certain phone, while Eq.6 reflects the probability of an utterance correspond a certain pronunciation quality score.

Due to the limited training data, we had not trained HMMs for all possible triphones. For the untrained triphones, we just do not calculate the phone-level score. Because there were only a small percentage of the triphones which were not trained, the word-level scoring should not be affected much by the untrained triphones.

3.4. Word-level Scoring

In the training of AM2, because we did not have the phone-level pronunciation quality manual scores, we had to use the whole word's pronunciation quality manual score as the substitution. However, pronunciation quality of phones may not all equal to the word's pronunciation quality, so there might be some error phone-level scores. Our solution was using the scoring confidences, which was calculated in Eq.6, to judge whether a phone-level score was error. We set an empirical value as the threshold, and any phone-level score whose confidence was lower than the threshold would be judged as error score. The error scores were excluded in calculating the word-level score.

we used Eq.7 to calculate the word score:

$$WdScore = \begin{cases} \dfrac{\sum_i \delta_i \cdot PhScore(p_i)}{\sum_i^N \delta_i}, & \sum_i^N \delta_i > 0 \\ \dfrac{1}{N}\sum_i PhScore(p_i), & \sum_i^N \delta_i = 0 \end{cases} \tag{7}$$

where p_i is the i-th phone in the word, δ_i is for excluding the low scoring confidence phones, which is defined as:

$$\delta_i = \begin{cases} 1, Conf(p_i) \geq thre \\ 0, Conf(p_i) < thre \end{cases} \tag{8}$$

where $thre$ is the excluding threshold.

4. EXPERIMENT

4.1. Data and Experimental Setting

In this study, two acoustic models were trained, that were AM1 and AM2. AM1 was trained in conventional method with 90 hours English native voices, and was adapted using Maximum A Posterior (MAP) method [14] with 10 hours Chinese students spoke English voices. The training data of AM2 contained about 80,000 English words pronounced by Chinese middle school students, and each word has an pronunciation quality score which was scored by human English teachers manually. The manual scores were in the set {1,2,3}, in which the score 3 means great pronunciation, the score 2 means average pronunciation, and the score 1 means bad pronunciation. The 80% of these data was used to train AM2 and the other 20% was for testing. The data was segmented into phones by forced alignment using AM1, then the HMMs of AM2 were trained using HTK with the "Isolated Word Training Strategy" [13] and a modified MMIE training which was described in Sect.3.2. Considered the size of the training set, we trained the GMMs of the AM2 as 4-mixed Guassians. The GOP method was used for contrast, which used AM1 as the acoustic model, and the GOP value was calculated following the Sect2.1 of Witt's paper [6].

4.2. Experimental Results

This paper uses "Scoring Difference" and "Correlation" to measure the performance of the system. "Scoring Difference" is the absolute difference of the machine score and manual score, and the "Correlation" is the Pearson's correlation coefficient of the machine score and manual score.

Firstly we compared the phone-level scoring results of some randomly selected utterances by GOP and TPDA methods. We randomly selected 150 utterances, and drew the machine scores of these utterances on the graph, in which utterances with different manual scores are drawn in different symbols, shown as Fig.4. From Fig.4, the GOP score was spread around the region -15 to 0, which requires to draw 2 lines to separate the GOP scores into 3 groups. Compared with the GDP method, the machine scores from TPDA method were in {1,2,3}, which was match the manual scoring levels directly.

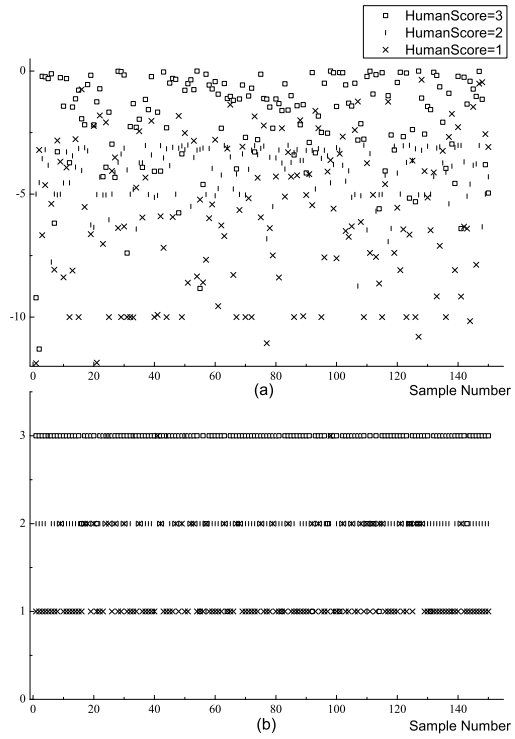


Fig. 4. Phone-level scoring result of randomly selected utterances. (a) Scoring by GOP; (b) Scoring by TPDA

The phone-level scoring performance of GOP and TPDA methods was compared in Table 1. From Table 1, the performance of the TPDA was better than GOP, that is because a discrimination of the different pronunciation quality utterances was made in the model training of the TPDA method, while the GOP method did not do it. This result reflected the advantage of discriminative methods.

The word-level scoring performance is shown in Fig.5. We tested using different threshold to calculate the word-level scores, and the best threshold for this experimental data was -2.0, as shown in Fig.5. The scoring performance on word-level is much better than phone-level, because that the phone-level manual scores were from the word-level manual scores, which have some inaccuracy for phone-level.

Table 1. Performance of phone-level scoring

	Scoring Difference	Correlation
GOP	0.644	0.647
TPDA	0.607	0.676

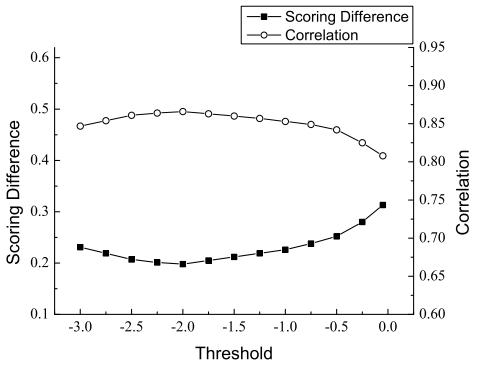


Fig. 5. Performance of the word-level scoring

Finally, we used the best results in Fig.5 to compare the scoring performance with the GOP method. The result is shown in Table 2. Table 2 shows that the scoring performance of our method was significantly better than GOP method on word-level scoring, that was caused by two reasons: first, TPDA can differentiate the triphones among different pronunciation quality triphones; second, the scoring confidence excluded the error phone-level scores. In contrast, GOP scoring on the phone-level does not differentiate the different pronunciation quality triphones directly, and it only use the simple average of all phone scores as the word score.

Table 2. Performance of word-level scoring

	Scoring difference	Correlation
GOP	0.262	0.823
TPDA	0.198	0.869

5. CONCLUSION

The novel method presented in this paper used two separate acoustic models on forced alignment and scoring. Thus AM2, the acoustic model for scoring, can be trained focusing on scoring. In the training of AM2, the triphones in different pronunciation qualities were trained as different units, and the MMIE criterion was used to maximize the discrimination ability of the different pronunciation quality triphones. In addition, the scoring confidences were calculated to exclude the error phone-level scores, which was very helpful for the word-level scoring. The experimental results shows that the scoring performance of our method was much better than the GOP method on both phone-level and the word-level, especially on the word-level, that proved the method described in this paper is effective.

6. REFERENCES

[1] W. Pei-ling, "The effect of computer-assisted whole language instruction on taiwanese university students' English learning," *English Language Teaching*, vol. 4, no. 4, pp. 10–15, 2011.

[2] H. Franco, V. Abrash, K. Precoda, H. Bratt, R. Rao, J. Butzberger, R. Rossier, and F. Cesari, "The SRI EduSpeak system: Recognition and pronunciation scoring for language learning," *Proceedings of InSTILL 2000*, pp. 123–128, 2000.

[3] K. Zechner, D. Higgins, and X. Xi, "SpeechRater: A construct-driven approach to scoring spontaneous non-native speech," in *Workshop on Speech and Language Technology in Education*, 2007.

[4] K. Zechner, D. Higgins, X. Xi, and D.M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[5] F. Ehsani and E. Knodt, "Speech technology in computer-aided language learning: Strengths and limitations of a new call paradigm," *Language Learning & Technology*, vol. 2, no. 1, pp. 45–60, 1998.

[6] S. Witt, S. Young, et al., "Language learning based on non-native speech recognition," in *Proc. of EUROSPEECH*. Citeseer, 1997, vol. 97, pp. 633–636.

[7] H. Strik, K. Truong, F. De Wet, and C. Cucchiarini, "Comparing different approaches for automatic pronunciation error detection," *Speech Communication*, vol. 51, no. 10, pp. 845–852, 2009.

[8] F. Ge, L. Lu, and Y. Yan, "Experimental investigation of mandarin pronunciation quality assessment system," in *Computer Science and Society (ISCCS), 2011 International Symposium on*. IEEE, 2011, pp. 235–239.

[9] J. Tepperman, S. Lee, S. Narayanan, and A. Alwan, "A generative student model for scoring word reading skills," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 2, pp. 348–360, 2011.

[10] K. Yan and S. Gong, "Pronunciation proficiency evaluation based on discriminatively refined acoustic models," *International Journal of Information Technology and Computer Science (IJITCS)*, vol. 3, no. 2, pp. 18–23, 2011.

[11] F. Ge, C. Liu, F Pan, D. Bin, and Y. Yan, "Effective acoustic modeling for pronunciation quality scoring of strongly accented mandarin speech," *IEICE transactions on information and systems*, vol. 91-D, no. 10, pp. 2485–2492, 2008.

[12] V. Valtchev, JJ Odell, PC Woodland, and SJ Young, "MMIE training of large vocabulary recognition systems," *Speech Communication*, vol. 22, no. 4, pp. 303–314, 1997.

[13] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, XA Liu, G. Moore, J. Odell, D. Ollason, D. Povey, et al., "The HTK book (for HTK version 3.4)," 2006.

[14] J.L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 2, pp. 291–298, 1994.