A Novel Discriminative Method for Pronunciation Quality Assessment

Junbo ZHANG\(^{(a)}\), Fuping PAN\(^{(b)}\), Bin DONG\(^{(c)}\), Nonmembers, Qingwei ZHAO\(^{(d)}\), Member, and Yonghong YAN\(^{(e)}\), Nonmember

SUMMARY In this paper, we presented a novel method for automatic pronunciation quality assessment. Unlike the popular “Goodness of Pronunciation” (GOP) method, this method does not map the decoding confidence into pronunciation quality score, but differentiates the different pronunciation quality utterances directly. In this method, the student’s utterance need to be decoded for two times. The first-time decoding was for getting the time points of each phone of the utterance by a forced alignment using a conventional trained acoustic model (AM). The second-time decoding was for differentiating the pronunciation quality for each triphone using a specially trained AM, where 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 quality scores were different. The decoding network in the second-time decoding included different pronunciation quality triphones, so the phone-level scores can be obtained from the decoding result directly. The phone-level scores were combined into the sentence-level scores using maximum entropy criterion. The experimental results shows that the scoring performance was increased significantly compared to the GOP method, especially in sentence-level.

key words: pronunciation assessment, automatic scoring, distinctiveness training, maximum entropy

1. Introduction

The technology of the automatic pronunciation quality assessment (APQA) has been widely used to help non-native students to learn languages, and the technology has been proved effective [1]. APQA can be used in two scenes: the first scene is being a part of computer-assisted language learning (CALL) system, telling the students where their pronunciation is good, and where is not good [2]; the second scene is being used in large-scale oral exam to assist or even replace human teachers to score [3]–[5]. Automatic scoring can avoid many disadvantages of manual scoring, such as high cost and lack of stability. For automatic scoring, the machine score should close to the manual score, that is, the absolute difference between the machine score and manual score should be minimized. A typical oral exam may require scoring for words, sentences, or short passages. A binary “right” or “wrong” score is usually not enough, most exams need a multi-level score, such as a score in the set \(\{0,1,2,3\}\).

One of the most influential achievement for APQA is the Goodness of Pronunciation (GOP) method [6] presented by Witt et al. in 1997. Before the presentation of GOP, most APQA systems typically required several recordings of native utterances to train the models for each word [7]–[10]. The disadvantage of these methods is that they are text-dependent, so they only work for the utterances with the same text of the native recordings, but can not be used on other utterances. Neumeyer et al. presented a text-independent pronunciation assessment framework [11] in 1996, then they improved the method by using the posterior probabilities instead of decoding log-likelihood [12], [13]. Witt et al. combined the advantages of these works and presented the GOP method. At present, GOP has been widely considered to be an effect method [14]. In Sect. 2, we will introduce the GOP method briefly.

After the GOP method has been presented, most scholars who research APQA focused on how to use and improve GOP. Ge et al. designed a confusion network to improve the calculation of the GOP denominator, and improved the methods of normalization for the GOP [15]. Tepperman et al. designed a better scoring system based on GOP, which used Bayesian Networks to post process GOP scores [16]. Yan et al. applied the discriminative Acoustic Model (AM) on APQA, replacing Maximum Likelihood Estimation (MLE) criterion by Minimum Phone Error (MPE) criterion in the training of the AM, and then used GOP to score [17]. However, the essence of all these methods is to calculate the decoding confidence of the student’s utterance, and then map the confidence into pronunciation quality score. All these methods were failed to discriminate the utterances in different pronunciation qualities. Although Yan et al. applied the discriminative AM in the pronunciation assessment, the discriminative training of the AM was for similar pronunciations rather than different pronunciation qualities, so the essence of that algorithm was still mapping the confidence into pronunciation qualities scores. This study tried to differentiate utterances with different pronunciation qualities directly rather than mapping confidence into scores.

This paper is organized as follows: In Sect. 2, we introduced the traditional GOP method and our new method for phone-level scoring; in Sect. 3, we described how to combine the phone-level scores into sentence-level score with
maximum entropy (MaxEnt) criterion; Sect. 4 shows the experimental results, and compared the performance between GOP and this method; at last in Sect. 5 the conclusion was made.

2. Discriminative Based Phone-level Scoring

2.1 The Traditional GOP Method

At present, most APQA systems use GOP method [6] to calculate phone-level pronunciation quality scores [14]. The GOP of any phone \( p \) is defined as the duration normalized log of the posterior probability \( P(p|O^p) \) that the speaker uttered phone \( p \) given the corresponding acoustic segment \( O^p \) as Eq. (1):

\[
GOP(p) = \frac{\log P(p|O^p)}{NF(p)} \approx \frac{\log P(O^p|p)}{\max_{q \in Q} P(O^p|q)} / NF(p)
\]  

(1)

where \( Q \) is the set of all phone models and \( NF(p) \) is the number of frames in the acoustic segment \( O^p \).

GOP method has been widely used in current APQA systems, however, it has shortcomings. As shown in Eq. (1), GOP reflects how much the student’s utterance matches the AM, but it does not reflect the characteristic of different pronunciation quality utterances, that is, GOP tells “how similar of the student’s utterance and the training utterances”, but it does not tell “which score-level utterances are most similar to the student’s utterance”.

Another shortcoming of GOP is that GOP is not a pronunciation quality score but needs to be mapped to pronunciation quality score, as shown in Fig. 1. To map GOP scores to pronunciation quality scores, several presumed thresholds are required, which are not easy to set. In real-world oral examination, the score should not be only a simple “correct” or “wrong”, but a multi-level score such as scoring in \([0, 1, 2, 3]\), which make the thresholds harder to set.

2.2 Overview of the method

This paper presents a Two-pass Discriminative Assessment (TPDA) method. Unlike GOP method, it does not map decoding confidence into scores, but differentiates different quality utterances directly, as shown in Fig. 2. The student’s utterance was decoded for two times with different AMs. The first-time decoding was to obtain each phone’s time points by processing a forced alignment. The AM used in the forced alignment is referred to “AM1” in this paper. AM1 was trained in conventional method. After the time points of each phone were obtained, the pronunciation quality score of each phone was calculated with another AM which is referred to “AM2” in this paper. AM2 was trained with different quality triphones. We will describe the training process of AM2 in Sect. 2.3. As the score of each triphone had been calculated, those scores of triphones were combined to a sentence-level score for the whole sentence.

Because the scores of each triphone are not all equal to the score of the whole sentence, the MaxEnt criterion was used to combine the scores of each triphone. The score combination method will be described in Sect. 3.

The workflow of our assessment system is shown in Fig. 3. The PLP features [18] of student’s utterance were extracted firstly. Then forced alignment was processed using AM1 to obtain the time points for each phone. After that, TPDA scoring for each phone was processed using AM2. Then the score of each phone was combined to the score of the whole sentence using MaxEnt model, therefore the sentence-level machine score was obtained.

2.3 Training AM2

Suppose \( \lambda \) is the set of all HMM parameters defining AM2. Because the AM for forced alignment has been served by AM1, AM2’s sole purpose was to judge the pronunciation quality for each triphone. To do this, triphones with different pronunciation quality 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 [19] is especially effective for this maximization. That is, making \( \lambda \) meet Eq. (2):

\[
\lambda = \arg \max_{\lambda} \mathcal{F}(\lambda)
\]  

(2)

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

\[
\mathcal{F}(\lambda) = \sum_{t} \sum_{k} \log P(O_{t,k}|\lambda)
\]

where \( O_{t,k} \) is the \( k \)-th score level of the \( t \)-th triphone.

In the discriminative training, we need to find the optimal \( \lambda \) which maximizes the score \( \mathcal{F}(\lambda) \) of triphones. However, since the set of all possible \( \lambda \) is very large (for example, \( 10^{20} \) for a typical system), we cannot exhaustively search through all \( \lambda \). Instead, we make use of the MMIE criterion and use gradient ascent to find the optimal \( \lambda \). The gradient of \( \mathcal{F}(\lambda) \) can be written as

\[
\nabla \mathcal{F}(\lambda) = \sum_{t} \sum_{k} \frac{P(O_{t,k}|\lambda)}{P(O_{t,k}|\lambda)} \nabla \log P(O_{t,k}|\lambda)
\]

Because the scores of each triphone are not all equal to the score of the whole sentence, the MaxEnt criterion was used to combine the scores of each triphone. The score combination method will be described in Sect. 3.

The workflow of our assessment system is shown in Fig. 3. The PLP features [18] of student’s utterance were extracted firstly. Then forced alignment was processed using AM1 to obtain the time points for each phone. After that, TPDA scoring for each phone was processed using AM2. Then the score of each phone was combined to the score of the whole sentence using MaxEnt model, therefore the sentence-level machine score was obtained.
where $O^{(r)}$ is the observation sequence $r$ of a triphone, $R$ is the number of observation sequences, $H^{(r)}_{ref}$ is the corresponding triphone of $O^{(r)}$ with the correct manual score, $H$ is the all corresponding triphones of $O^{(r)}$. Thus the average log-posterior of the reference, $H^{(r)}_{ref}$, is maximized.

The training process is shown in Fig. 4. As the time points of each triphone had been obtained, 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 [20]. So we got the MLE model which can be used as the initial model for the iteration in MMIE training, and then AM2 can be obtained by the MMIE training which was described in [19]. In this study, the denominator lattice in the MMIE training only included the corresponding triphones of the training voice, as shown in Fig. 5.

### 2.4 Decoding to Score in Phone-level

As shown in Fig. 3, after the forced alignment, the student’s utterance was 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. 5, 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 higher than all other paths, as shown in Eq. (4):

$$S_{core}(s) = \arg \max_i \log(p(O^s|p_i))$$  \hspace{1cm} (4)

where $O^s$ is the observation sequence of the segment $s$, $p_i$ is the $i$-th corresponding triphone of $s$.

Thus we obtained a “Score Item” (SI) for each triphone. The format of SI is:

<table>
<thead>
<tr>
<th>triphone:score:count</th>
</tr>
</thead>
<tbody>
<tr>
<td>s-ih+l:2:1</td>
</tr>
<tr>
<td>ih-z+sil:3:1</td>
</tr>
<tr>
<td>s-ih+l:2:2</td>
</tr>
<tr>
<td>w-ah+z:1:1</td>
</tr>
<tr>
<td>v-l+er:2:1</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

where the “count” means the index of occurrences of a triphone with a certain score.

For each student’s utterance, putting together all the SI of the utterance, we got a “Triphone Score Set” (TSS) of the utterance, which is for the sentence-level scoring. Here is an example of a typical TSS:

s-ih+l:2:1
ih-z+sil:3:1
s-ih+l:2:2
w-ah+z:1:1
v-l+er:2:1
...

For example, the item “ih-z+sil:3:1” in the TSS means that the triphone name was “ih-z+sil”, the score was 3 and this item was the first time appeared in the TSS; the item “s-ih+l:2:2” in the TSS means that the triphone name was “s-ih+l”, the score was 2 and this item was the second time...
appeared in the TSS.

Due to the limited training data, we haven’t trained for all possible triphones. For the untrained triphones, we just did not calculate the SI. Because there were only a small percentage of the triphones which were not trained, the sentence-level scoring should not be affected much by the untrained triphones.

3. Sentence-Level Scoring

In the training of AM2, we used the manual scores of the whole sentence as the triphone manual scores, for we did not have the phone-level manual scores, however, pronunciation quality scores of triphones were not all equal to the sentence pronunciation scores, that led to imprecision for some triphones’ models, and it was hard to know which specific triphones’ models were imprecision.

For this reason, we do not take the average of the triphones scores as the sentence score. Instead, we try to use a data-driven method to find out the relationship between the phone-level scores and the score of the whole sentence. So we can enhance the weight of triphones which can contribute much in sentence-level scoring, and weaken the weight of triphone whose contribution was little in sentence-level scoring.

As described in Sect. 2.4, after TPDA scoring, we had obtained a TSS for each sentence. TSS is composed of a number of SI, each SI contains a triphone-level scoring information. In this work, we used the maximum entropy method to make better use of those information in TSS. The principle of the method is simple “model all that is known and assume nothing about that which is unknown”, that is, makes the fewest possible assumption [21].

Accoding the process of the maximum entropy method, we defined an indicator function \( f(x, y) \) as Eq. (5):

\[
f(x, y) = \begin{cases} 1, & \text{if } y = s_h \text{ and } x = SI \\ 0, & \text{otherwise} \end{cases}
\]

where \( s_h \) is the manual score of the sentence.

For any \( f(x, y) \), the mathematic expectation in the training data is:

\[
\hat{E}[f(x, y)] = \sum_{x,y} \hat{p}(x, y)f(x, y)
\]

where \( \hat{p}(x, y) \) is the empirical distribution of \((x, y)\) in the training set.

And for any \( f(x, y) \), the mathematic expectation in the testing data is:

\[
E[f(x, y)] = \sum_{x,y} p(x, y)f(x, y) = \sum_{x,y} p(y|x)p(x)f(x, y) \approx \sum_{x,y} p(y|x)\hat{p}(x)f(x, y)
\]

where \( p(x) \) is the empirical distribution of \( x \) in the testing set.

We constrain this mathematic expectation values in training data and testing data to be the same, that is, we require:

\[
E[f(x, y)] = \hat{E}[f(x, y)]
\]

Combining Eq. (6), Eq. (7) and Eq. (8), we get a constraint:

\[
\sum_{x,y} \hat{p}(x, y)f(x, y) = \sum_{x,y} p(y|x)\hat{p}(x)f(x, y)
\]

Accoding the maximum entropy criterion, the best model of the data satisfies certain constraints and makes the fewest possible assumptions. The “fewest possible assumptions” is closest to the uniform distribution [22]. Therefore, we should select the \( p(y|x) \) which is most uniform under the constraint Eq. (9). A mathematical measure of the uniformity of a conditional distribution \( p(y|x) \) is provided by the conditional entropy:

\[
H(p) = -\sum_{x,y} \hat{p}(x)p(y|x)\log p(y|x)
\]

Then we need to find \( p^*(y|x) \) to maximize Eq. (10) under the constraint condition of Eq. (9). According to the algorithm described in [21], we need to make \( p^*(y|x) \) meet Eq. (11):

\[
p^*(y|x) = \frac{e^{\sum \lambda_i f_i}}{\sum_y e^{\sum \lambda_i f_i}}
\]

where \( f_i \) is any \( f(x, y) \), and each \( \lambda_i \) in Eq. (11) was obtained by the improved iterative scaling (IIS) [23] algorithm. The result of the IIS algorithm is the series of \( \lambda \) which constituted a MaxEnt model [21]. With this MaxEnt model, we can use Eq. (12) to calculate the sentence score:

\[
Score = \arg \max_y p(y|x_1, x_2, \ldots, x_M)
\]

\[
= \arg \max_y \frac{e^{\sum \lambda_i f_i}}{\sum_y e^{\sum \lambda_i f_i}}
\]

where \( M \) is the size of the TSS.

4. Experiment

4.1 Data and Experimental Setting

In this study, three models were need to be trained, that were AM1, AM2 and the MaxEnt model. AM1 has been trained in our previous studies, and some optimized measures specifically for pronunciation quality assessment had been taken in the training process, which was described in details in our previous paper [24]. The training data of AM2
was from an English oral exam in China. The entire data set contained about 90,000 short sentences, each sentence contained about 10 words, and each sentence has a manual score. The manual scores were in \{0,1,2,3\}, in which the score 3 means great pronunciation, the score 2 means average pronunciation, the score 1 means bad pronunciation, and the score 0 means the student spoke nothing. All the training data was segmented to triphones by forced alignment using AM1, then the HMMs were trained using HTK with the Isolated Word Training Strategy [20] and MMIE training. Considered the size of the training set, we trained the GMMs as 4-mixed Guassians. The MaxEnt model was trained using the data from another English oral exam in China, the characteristics of this data set was similar to the data set training AM2, which also contained about 90,000 short sentences with manual scores. 80% of these data was used to train the MaxEnt model and the other 20% was for testing. The GOP method was used for contrast, which used AM1 as the acoustic model, and the GOP value was calculated following the Sect. 2.1 of Witt’s paper [6].

4.2 Performance Metrics

This paper uses three indicators to measure the performance of the system. The indicators are: “Average Scoring Absolute Difference” (ASAD), “Serious Error Rate” (SER) and “Correlation Coefficient” (CC).

To describe ASAD, “Scoring Absolute Difference” (SAD) should be introduced firstly. SAD is the difference of the absolute value between the manual score and machine score as shown in Eq. (13).

\[
SAD_i = |sc_i - sh_i|
\]  

(13)

where \(sc_i\) is the machine score of the \(i\)-th sample, and \(sh_i\) is the manual score.

ASAD is the average value of SAD, as shown in Eq. (14).

\[
ASAD = \frac{1}{N} \sum_{i=1}^{N} SAD_i
\]  

(14)

where \(N\) is the number of samples of the testing data set.

In the real examinations, if the computer give a student a score with large error, it would be very unfair to that student. Those large error scores are called “serious error scores”. SER is used to measure the percentage of serious error scores, as shown in Eq. (15).

\[
SER = \frac{1}{N} \sum_{i=1}^{N} sgn(SAD_i - T)
\]  

(15)

where the \(T\) is the threshold value, and the function \(sgn(x)\) is defined as Eq. (16). We set \(T = 1.0\) here, which means that the scores whose error were larger then 1.0 were treated as serious error scores.

\[
sgn(x) = \begin{cases} 
1, & x > 0 \\
0, & x \leq 0 
\end{cases}
\]  

(16)

CC refers to the correlation coefficient between machine score and manual score. In the three indicators, we shall decrease ASAD and SER, and increase CC as much as possible.

4.3 Experimental Results

Firstly we compared the phone-level scoring result of some randomly selected utterances by using 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. 6. From Fig. 6, the GOP scoring method can differentiate some of the utterances with different manual scores. The GOP score was spread around the region \(-5\) to \(0\). In order to let the GOP scores match the manual scoring levels, it requires to draw \(N - 1\) lines to separate the GOP scores into \(N\) groups. The machine scores from TPDA method were in \{1,2,3\}, which was match the manual scoring levels directly.

The phone-level scoring performance of GOP and TPDA methods was compared in Table 1, where the performance was measured by the indicators described in Sect.4.2. From Table 1, the performance of the TPDA was better than GOP on all the three indicators. 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 advan-

![Fig. 6 Phone-level scoring result of randomly selected utterances. (a) Scoring by GOP; (b) Scoring by TPDA.](image-url)
Table 1 Performance of phone-level scoring.

<table>
<thead>
<tr>
<th></th>
<th>ASAD</th>
<th>SER</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOP</td>
<td>0.646</td>
<td>0.114</td>
<td>0.645</td>
</tr>
<tr>
<td>TPDA</td>
<td>0.628</td>
<td>0.101</td>
<td>0.653</td>
</tr>
</tbody>
</table>

Table 3 Performance of sentence-level scoring.

<table>
<thead>
<tr>
<th></th>
<th>ASAD</th>
<th>SER</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOP</td>
<td>0.281</td>
<td>0.192</td>
<td>0.797</td>
</tr>
<tr>
<td>TPDA+MaxEnt</td>
<td>0.219</td>
<td>0.085</td>
<td>0.843</td>
</tr>
</tbody>
</table>

Fig. 7 The percentage of triphones with different ASAD.

Table 2 Performance of the MaxEnt models.

<table>
<thead>
<tr>
<th># of Iterations</th>
<th>ASAD</th>
<th>SER</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.295</td>
<td>0.108</td>
<td>0.796</td>
</tr>
<tr>
<td>10</td>
<td>0.227</td>
<td>0.088</td>
<td>0.831</td>
</tr>
<tr>
<td>15</td>
<td>0.219</td>
<td>0.085</td>
<td>0.843</td>
</tr>
<tr>
<td>20</td>
<td>0.220</td>
<td>0.086</td>
<td>0.841</td>
</tr>
</tbody>
</table>

In this study, the goal of scoring was on the sentence-level rather than phone-level, and the phone-level scores were just intermediate results. Table 2 shows the sentence-level scoring performance. We tested different numbers of iterations to train the MaxEnt model. The iterative training methods of the MaxEnt model was introduced in [23]. Comparing Table 1 and Table 2, it is obvious that the scoring performance on sentence-level is much higher than phoneme-level. The result suggests that, the statistical methods trained in data-driven way, can automatically increase the sentence scoring weights of some triphone which is more useful on sentence scoring.

Finally, we used the best results in Table 2, i.e., the number of iterations was 15, to compare the scoring performance with the GOP method. The result is shown in Table 3. Table 3 shows that the scoring performance of our method was significantly higher than that of the GOP method on sentence-level scoring, that was because by two reasons: first, TPDA can differentiate the triphones among different pronunciation quality triphones; second, the MaxEnt method reduced the impact of the triphones whose pronunciation qualities were different to the whole sentence. In contrast, GOP scoring on the phone-level does not reflect the discriminative of the triphones with different pronunciation quality, and on the sentence-level, it only use a simple average of phone scores as the sentence score. Therefore, the performance of GOP method was significantly lower than this method.

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 MMIE criterion was used to maximize the discrimination ability of the training utterances with different pronunciation qualities. While on sentence-level scoring, a data-driven method was used to calculate the sentence-level score to reduce the impact of the triphones whose pronunciation qualities were different to the whole sentence. The experimental results shows that the scoring performance of our method was higher than the GOP method on both phone-level and sentence level, especially on the sentence-level, that proved the method described in this paper is effective.

Acknowledgment

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

References

dren’s read-aloud text passages and word lists,” Proc. Fourth Work-
shop on Innovative Use of NLP for Building Educational Applica-
ing of non-native spontaneous speech in tests of spoken English,”
speech recognition,” Fifth European Conference on Speech Com-
munication and Technology, 1997.
pronunciation using hidden Markov models,” Fifth European Con-
system for computer-aided pronunciation teaching,” Speech Com-
lish pronunciation based on speech recognition techniques,” IEICE
[10] C. Rogers, J. Dalby, and G. DeVane, “Intelligibility training for
text-independent pronunciation scoring of foreign language student
speech,” Proc. Fourth International Conference on Spoken Lan-
ciation scoring for language instruction,” 1997 IEEE International
matic scoring of pronunciation quality,” Speech Commun., vol.30,
different approaches for automatic pronunciation error detection,”
pronunciation quality assessment system,” 2011 International Sym-
poum on Computer Science and Society (ISCCS), pp.235–239,
2011.
student model for scoring word reading skills,” IEEE Trans. Audio
on discriminatively refined acoustic models.” Int. J. Information
Technology and Computer Science (IJITCS), vol.3, no.2, pp.18–23,
2011.
[18] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,”
of large vocabulary recognition systems,” Speech Commun., vol.22,
Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland,
The HTK book (for HTK version 3.4),” 2006.
[22] A. McCallum, D. Freitag, and F. Pereira, “Maximum entropy
Markov models for information extraction and segmentation,” Proce.
Seventeenth International Conference on Machine Learning,
[23] A. Berger, “The improved iterative scaling algorithm: A gentle in-
[24] F. Ge, C. Liu, F. Pan, D. Bin, and Y. Yan, “Effective acoustic model-
ing for pronunciation quality scoring of strongly accented mandarin

Junbo Zhang received his BE from Beijing University of Chemical Technology. Now he is a
Ph.D. candidate of Key Laboratory of Speech Acoustics and Content Understanding, Institute
of Acoustics (IOA), Chinese Academy of Science (CAS). His research interests include pro-
nunciation quality assessing and speech recognition.

Bin Dong received his Ph.D. in Information and Signal Processing from IOA, CAS, 2007. He is currently an Assistant Researcher in
IOA. His research is focused on automatic pronunciation evaluation, speech signal processing and
speech recognition.

Qingwei Zhao received his Ph.D. in electronic engineering from Tsinghua University in
1999. Now he is an Associate Professor at
ThinkIT Lab, IOA, CAS. Before joining CAS,
he was with Intel as a Senior Researcher. His
research interests include spontaneous speech
recognition, keyword spotting and automatic
pronunciation evaluation.

Yonghong Yan received his BE from Ts-
inghua University in 1990, and his PhD from
Oregon Graduate Institute (OGI). He worked in
OGI as Assistant Professor (1995), Associate
Professor (1998) and Associate Director (1997)
of Center for Spoken Language Understanding.
He worked in Intel from 1998–2001, chaired
Human Computer Interface Research Council,
worked as Principal Engineer of Microprocessor
Research Lab and Director of Intel China Re-
search Center. Currently he is a professor and
director of ThinkIT Lab. His research interests include speech processing and recognition, language/speaker recognition, and human computer
interface. He has published more than 100 papers and holds 40 patents.