SUMMARY In this paper we present our investigation into improving the performance of our computer-assisted language learning (CALL) system through exploiting the acoustic model and features within the speech recognition framework. First, to alleviate channel distortion, speaker-dependent cepstrum mean normalization (CMN) is adopted and the average correlation coefficient (average CC) between machine and expert scores is improved from 78.00% to 84.14%. Second, heteroscedastic linear discriminant analysis (HLDA) is adopted to enhance the discriminability of the acoustic model, which successfully increases the average CC from 84.14% to 84.62%. Additionally, HLDA causes the scoring accuracy to be more stable at various pronunciation proficiency levels, and thus leads to an increase in the speaker correct-rank rate from 85.59% to 90.99%. Finally, we use maximum a posteriori (MAP) estimation to tune the acoustic model to fit strongly accented test speech. As a result, the average CC is improved from 84.62% to 86.57%. These three novel techniques improve the accuracy of evaluating pronunciation quality.

key words: CALL, speech recognition, HLDA, speaker-dependent CMN, e-learning

1. Introduction

In most foreign language learning courses, effort is mainly focused on reading, writing and listening comprehension, while much less attention is paid to teaching correct pronunciation. One of the most important reasons for this is that it requires much more expensive resources, such as extensive practice with teachers who are native speakers of the target language. Computer-assisted language learning (CALL) systems aim to improve this situation. They can provide continuous feedback to the student without requiring the full attention of the teacher, and therefore facilitate self-study and encourage interactive practice of the language rather than tedious learning. To be successful, a CALL system must evaluate pronunciation quality accurately. Pronunciation quality assessment is the main component of CALL systems.

Over the last few decades, many research groups have developed interactive language teaching systems incorporating pronunciation quality assessment based on speech recognition techniques. The research has been mainly focused on assessment at the sentence level and speaker level. Evaluation scores are obtained by combining various machine measurements, including scores of Hidden Markov Model (HMM) log-likelihood, timing, phone log-posterior probability and segment duration [1], [2].

In this paper we describe our CALL system, which is dedicated to evaluating the pronunciation quality of Mandarin Chinese syllables and words. Mandarin is a syllabic language, in which each syllable has two parts: the initial part is a consonant and the final part is a vowel. Because of this strict phonetic structure, achieving high segmental pronunciation quality is a major difficulty for some Chinese speakers with strong dialects, such as Hong Kong native speakers. We take the phone as the basic assessment unit for segmental pronunciation quality, and combine phonetic scores to form syllable-level and word-level scores. A classical measurement of segmental pronunciation quality is the phonetic posterior probability. It has been used in several studies [1], [3]–[5]. However, it does not correlate strongly with human assessor. Witt and Young studied pronunciation quality assessment at the phone level and suggested improving the evaluation performance by using a likelihood-based “goodness of pronunciation” (GOP) measure and explicit error modeling [6]. However, little research has been conducted on improving the phonetic evaluation accuracy from the viewpoint of modifying the acoustic model and features.

Our CALL system also operates under the speech recognition framework. It measures the quality of pronunciation (QOP) at the phone level on the basis of two types of phonetic posterior probabilities, which are defined as the average of the frame-based posterior probabilities [3] and the phone log-posterior probability. These two types of QOP are combined into a phonetic score. In this system, there are three outstanding problems: channel distortion, the low discriminability of the acoustic model and the mismatch between the speaker-independent (SI) acoustic model (AM) and strongly accented test speech. These three problems result in unacceptable evaluation performance for strongly accented Mandarin speech. Therefore, three measures are proposed to resolve the problems and improve the accuracy of phonetic evaluation.

First, speaker-dependent cepstrum mean normalization (SD-CMN) is used to normalize the input features, which effectively reduces the effect of the channel on speech. Secondly, we apply heteroscedastic linear discriminant analysis (HLDA) [7] to improve the discriminability of the acoustic
model and better separate easily confusable phones. Finally, maximum a posteriori (MAP) analysis is adopted during the acoustic model training, which can tune the acoustic model to match strongly accented test speech. Experiments show that the three measures can effectively improve the system performance; the average correlation coefficient (average CC) between machine and expert scores is increased from 78.00% to 86.57%. Additionally, HLDA resolves the problem that very high quality pronunciation is usually underestimated and that very low quality pronunciation is usually overestimated by our system, and the speaker correct-rank rate is increased from 85.59% to 90.99%.

The rest of the paper is organized as follows: in Sect. 2 we describe our CALL system; in Sect. 3 we analyze the essential benefits of speaker-dependent CMN, HLDA and MAP; experiments are presented in Sect. 4; and Sect. 5 concludes the paper.

2. Overview of CALL System

Our CALL system evaluates the pronunciation quality of Mandarin speech at the levels of syllables, phrases and sentences. At all levels, the Mandarin syllable is the fundamental assessment unit. All Mandarin syllables can be considered as a combination of initial and final parts. The phonetic structure of a Mandarin syllable can be defined as shown in Fig. 1. The initial part is articulated with a final part to form a syllable. Mandarin is a tonal language. The tone is mainly specified by the pattern of the pitch contour of the vowel part of the syllable.

Three aspects of syllable pronunciation are evaluated: the quality of the consonant, the quality of the vowel and the accuracy of the tone. The first two aspects are evaluated first using the automatic speech recognition (ASR) techniques of a HMM and Viterbi search [8], which form the core of the system. Then the tone quality is evaluated using a Gaussian mixture model (GMM) classifier, which is pre-trained with Mandarin speech. Tone evaluation is not considered in this paper, but is discussed in [9], [10]. A block diagram of the system is shown in Fig. 2. The front-end feature extraction converts the speech waveform to a sequence of mel-frequency cepstral coefficients (MFCCs), which are then fed into the HMM net to undergo one-pass Viterbi decoding. The HMM model net consists of models of the learning text, and the Viterbi decoding is used to ensure force alignment between the speech frames and the HMM models in the net. From the frame index of each HMM state and the accumulated observation probability of the phone segment, the phonetic posterior probability score is calculated as a measurement of the pronunciation quality of each phone. There are two types of phonetic posterior probabilities in our system.

One is the average of the logarithm of the frame-based posterior probability (AFBPP) [3]–[5],

\[
\rho_{\text{AFBPP}}(PH|O) = \frac{1}{e - b + 1} \sum_{i=b}^{e} \log P(s_i|o_i),
\]

where \( O = [o_b, o_{b+1}, \ldots, o_e] \) is the force-aligned observation sequence of the phone PH (which is a consonant or vowel), \( b \) is the initial frame of PH and \( e \) is the final frame of PH. \( S = [s_b, s_{b+1}, \ldots, s_e] \) is the state sequence corresponding to \( O \). \( P(s_i|o_i) \) is the state posterior probability computed by Eq. (2),

\[
P(s_i|o_i) = \frac{p(o_i|s_i)p(s_i)}{p(o_i)} = \frac{p(o_i|s_i)p(s_i)}{\sum_{j \in S} p(o_i|s_j)p(s_j)},
\]

where \( p(o_i|s_i) \) is the output probability of observation \( o_i \) from state \( s_i \), and \( S \) is the global model state set.

The other probability is the phone log-posterior probability (PLPP), which is calculated by Eq. (3),

\[
\rho_{\text{PLPP}}(PH|O) = \frac{1}{\tau} \log \frac{P(q|O^{(q)})}{\sum_{p \in Q} P(O^{(q)}|p)},
\]

where \( \tau \) is the number of frames in the acoustic segment \( O^{(q)} \) and \( Q \) is the set of Mandarin consonants when \( q \) is a consonant and the set of Mandarin vowels when \( q \) is a vowel.

We combine the two posterior probability scores into a more effective confidence value as follows,

\[
P(PH|O) = \omega \ast \rho_{\text{AFBPP}} + \rho_{\text{PLPP}},
\]

where \( P(PH|O) \) is the combined posterior probability of the phone PH, \( O \) is its acoustic segment, and \( \omega \) is the combination coefficient.

The confidence of \( P(PH|O) \) is an absolute measurement describing how close the utterance is to the standard

![Fig. 1](image)

**Fig. 1** Structure of Mandarin syllable.

![Fig. 2](image)

**Fig. 2** Our CALL system structure.
pronunciation. It can be used directly for phonetic pronunciation assessment [1], [4]. We classify phonetic pronunciation quality into three classes: good, fair and poor, corresponding to scores of 2, 1 and 0 respectively. Thus, the final stage of evaluation is to use predetermined thresholds, which are trained by the development corpus annotated with expert grades, to map the posterior probability scores of \( P(PH|O) \) to evaluation grades. The thresholds used for consonants and vowels are context-dependent, i.e., the same phone in different syllables has different thresholds.

Evaluation of the tone is carried out in parallel with the phone pronunciation assessment. The scores for tone are also restricted to 2, 1 and 0, which correspond to good, fair and poor, respectively. Finally, the phone pronunciation scores and the tone score are integrated by Eq. (5) to form the final score for the syllable.

\[
Score_{\text{syllable}} = \min(\text{Score}_{\text{Consonant}}, \text{Score}_{\text{Vowel}}, \text{Score}_{\text{Tone}})/2
\]  

3. Three Measures for Improving the Performance

3.1 Speaker-Dependent CMN

The environment in which the CALL system is used is not necessarily the same as that in which the training data of the acoustic model is recorded. Also, the channel difference can degrade the performance severely. Therefore, the elimination of multiplicative noise is significant for accurate pronunciation quality assessment. The CMN algorithm is an effective method of reducing the impact of multiplicative noise in speech recognition. It normalizes feature vectors using their mean and variance by Eq. (6),

\[
\text{newfeature} = \frac{\text{oldfeature} - \mu}{\sigma},
\]

where oldfeature and newfeature are feature vectors before and after CMN, respectively, \( \mu \) is the mean of the feature vectors before CMN, and \( \sigma \) is the variance.

Generally, CMN is performed on the feature vectors of a single utterance, from which \( \mu \) and \( \sigma \) are obtained. Our CALL system is used in the Putonghua level test, and all the speech data of one speaker is available before assessing his/her pronunciation quality. Therefore, we suggest that CMN should be carried out using all the utterances of one speaker in accordance with Eqs. (7) and (8) to more effectively eliminate the background noise:

\[
\mu = \frac{\sum_{\text{framenum}} \text{oldfeature}}{\text{framenum}},
\]

\[
\sigma = \sqrt{\frac{\sum_{\text{framenum}} (\text{oldfeature} - \mu)^2}{\text{framenum} - 1}},
\]

where framenum is the total number of frames for the utterances of the speaker.

Compared with conventional CMN, speaker-dependent CMN can estimate the mean and variance more accurately, and more effectively normalize the speech features. In particular, when the training data for the acoustic model and the testing data are recorded under very different conditions, performing speaker-dependent CMN on both sets of data can normalize the data more effectively.

3.2 HLDA

In a CALL system, the discriminability of the acoustic model is very important for ensuring assessment accuracy. It must be able to distinguish different qualities of the pronunciation, i.e., it must give a higher score to a correct pronunciation and a lower score to an inaccurate pronunciation. The role of HLDA is to transform the feature space to enhance its discriminability. Because the amount of redundant information in the feature vectors is reduced, there are fewer parameters in the acoustic model to be estimated, and the acoustic model can be trained more accurately.

HLDA is defined in a maximum-likelihood framework [11]. The projection from a \( p \)-dimensional feature space into a \( q \)-dimensional feature space with \( q < p \) is performed by a matrix \( T \), which is estimated iteratively by the EM algorithm. Given matrix \( \hat{T} \) from the previous step and the \( p \)-dimensional feature vectors \( \hat{O} := o_1, \ldots, o_t \), the improved matrix \( \hat{T} \) is chosen such that \( Q(\hat{T}, T) = \sum_m \sum_t p(\eta, \hat{T}) \cdot \log(\eta, N(T_{ot}, \Sigma_m)) \) is maximized. In this equation, \( N(\eta, \Sigma_m, \Sigma_m) \) represents the output density \( \eta \) of a Gaussian mixture. The covariance matrices \( \Sigma_m \) are constrained to be diagonal to allow the row-by-row optimization of \( T \). \( T \) is optimized by an ad hoc iterative numerical procedure [7], [12], utilizing full-covariance statistics for each density \( m \). To ensure that \( T \) is a projection that optimizes rows \( q+1, \ldots, p \) of \( T \), the state-specific statistics are replaced by the corresponding global covariance statistics. Consequently, after estimating \( T \), the dimensions \( q+1, \ldots, p \) of the transformed feature vectors \( T_{ot} \) can be discarded, yielding \( q \)-dimensional feature vectors \( a^T \). Because HLDA diagonalizes the feature space, it is more reasonable to estimate the transformation matrix by jointly updating the covariance matrices of the output densities of the acoustic models \( \Lambda \).

The output is not only a matrix \( T \), but also an updated set of acoustic models \( \Lambda^T \).

3.3 MAP

Ideally, we should adopt standard pronunciations as the evaluation criterion to assess Mandarin pronunciation quality. However, for speakers with a strong accent, their articulation is different from that of native Mandarin speakers. They often have difficulty in pronouncing some initials/finals. Thus they use some of the typical strategies of language learners to compensate for such difficulties, including phonological transfer, overgeneralization, prefabri-
in a way similarly close to standard Mandarin, but not accurately. Accordingly, adjusting the SI AM to match the test speech is necessary for improving the phone segmentation during forced alignment for strongly accented Mandarin speech, and also the speech of nonnative speakers speech. The MAP algorithm is an excellent measure for adjusting the SI AM when adequate data is available.

The MAP algorithm uses the criterion of maximum a posteriori:

\[ \hat{\lambda}_i = \arg \max_{\lambda_i} P(\lambda_i | \chi), \]  

(9)

where \( \chi \) is the feature vector, \( \lambda_i \) is the HMM model of state \( i \), and \( \hat{\lambda}_i \) is its modified HMM model, which can be simplified into the following form.

\[ \hat{\lambda}_i = \arg \max_{\lambda_i} [P(\chi | \lambda_i) P(\lambda_i)] \]  

(10)

Because the MAP algorithm contains information both the acoustic model and the speaker, it can tune the SI AM towards the target speaker or speakers. Therefore, the MAP algorithm works effectively when the SI acoustic model does not match the tester’s speech.

4. Experiments and Results

4.1 Data Corpora

Our system is used to assist the Hong Kong Putonghua Shuiping Kaoshi (PSK) test, which focuses on the pronunciation quality assessment of mandarin speech. The test is taken by Hong Kong native undergraduates with strong southern Chinese dialects. Each test includes 75 utterances, of which the first 50 utterances are isolated syllables and the last 25 utterances are two-syllable words, and there are a total of 100 syllables. The maximum score is 100. The computer-assisted evaluation of each syllable is performed in Sect. 2 of the test, and scores for the 100 syllables are summed to give the final machine score for the test speaker. Meanwhile, five experts are asked to judge the speakers’ pronunciation separately in accordance with given instructions. First, each expert grades the pronunciation quality of the consonant, vowel and tone of each syllable with 2, 1 or 0 for good, fair and poor, respectively. A grade that receives at least three votes is used as the score, otherwise the score is 1. Thus, each consonant, vowel and tone of a syllable are scored by experts. Using Eq. (5), we obtain the expert score for a syllable, and the 100 expert scores for the syllables are summed to give the final expert assessment result for the test speaker. We compare machine and expert scores to estimate the performance of our CALL system.

We collected three groups of PSK test samples, PSK1_eval corpora, PSK2_eval corpora and PSK3_eval corpora, to evaluate the CALL system more comprehensively. All samples comprise speech by Hong Kong native speakers. The speech content of the test samples in each individual group is the same, but those for the three groups are different from each other. Using the development corpora, whose content is the same as the test corpora, and the expert scores, the individual thresholds for each phone depending on the syllable are determined. Thus, the predetermined context-dependent thresholds of phones are different for the three test corpora, and we must test them separately. There are 60 speakers in PSK1_eval, 51 speakers in PSK2_eval and 108 speakers in PSK3_eval.

The acoustic models used in our experiments are gender-independent, continuous-mixture-density, tied-state, within-word triphone HMMs. The basic HMM sets consist of about 5100 tied states, each of which has 16 Gaussian components. The front end uses MFCC analysis to obtain a 39-dimensional feature, including 12-dimensional static cepstra and 1-dimensional energy, with 1st- and 2nd-order derivatives.

The data used for training the acoustic model is standard Mandarin speech totally about 400 hours, including 250 hours of native Chinese speech from the corpus of the National 863 Hi-Tech Project and 150 hours of speech data spoken by native speakers in a quiet room. The Mandarin pronunciation quality of each speaker was above average. For MAP training, we use 50 hours of well-pronounced speech by Hong Kong native speakers that was different from the PSK test samples used for evaluation. The training and testing data do not overlap with each other.

4.2 Evaluation Metrics

There are four evaluation metrics: the correct assessment rate, the average score difference, the average CC between the machine and expert scores and the speaker correct-rank rate.

Consonants and vowels each have their own correct assessment rate, which is the ratio of the number of correctly assessed phones to the total number of phones. The average score difference is defined as the average score difference between the machine and the expert for all the speakers. The average CC between the machine and expert scores is calculated as follows:

\[ R = \frac{1}{total} \sum_{k=0}^{total} \sum_{i=0}^{total} \frac{f_{ki} \cdot g_{ki}}{\|f_{ki}\| \cdot \|g_{ki}\|} \]  

(11)

\[ = \frac{1}{total} \sum_{k=0}^{total} \frac{\sum_{i=0}^{total} (f_{ki} \cdot g_{ki})}{\sqrt{\sum_{i=0}^{total} f_{ki}^2} \sqrt{\sum_{i=0}^{total} g_{ki}^2}} , \]

where \( f_{ki} \) and \( g_{ki} \) are the machine and expert scores, respectively, for test \( k \) and syllable \( i \), and total is the tester number.

In addition, the speaker correct-rank rate is used in Sect. 4.4.2. The Putonghua level test [13] includes six ranks as shown in Fig. 3. The maximum score is 100, and different score ranges are represented by different ranks. If one’s score is below 60, then his/her Putonghua is poor. We project the machine and expert scores onto the ranks, and the speaker correct-rank rate is the ratio of the number of correctly ranked speakers whose machine and expert scores
are ranked are the same, to the total number of speakers.

The four evaluation metrics reflect the performance of our CALL system from different aspects.

### 4.3 Experiment Design

The three techniques discussed in Sect. 3 are used to optimize the acoustic features or model from various aspects, and they do not interfere with each other. We design the following four experiments to demonstrate the improvements in performance of our mandarin pronunciation quality assessment system.

**System 1:** baseline system, in which the acoustic model is trained without speaker-dependent CMN, MAP or HLDA.

**System 2:** we apply speaker-dependent CMN while training the acoustic model and testing the CALL system.

**System 3:** speaker-dependent CMN and HLDA are both adopted. Before training the acoustic model, 52-dimensional feature vectors are extracted from the training data set, which are three-times-delta added to 39-dimensional MFCC features. We normalize the 52-dimensional feature vectors using speaker-dependent CMN. Then a 52-dimensional initial model is trained. Finally, HLDA transforms the initial acoustic model to a 39-dimensional model. During the test, 52-dimensional feature vectors are extracted, and projected to 39 dimensions by the HLDA eigen-matrix.

**System 4:** based on System 3 with the MAP algorithm is added to the acoustic model training after HLDA. The testing process is the same as System 3. In our experiments, we collect the correct pronunciations of the target strongly accented Mandarin speakers to form the data corpora of the MAP algorithm.

### 4.4 System Performance

To compare the four systems, we analyze their performances from three different aspects, the discriminability, the improvement of scoring accuracy stability for various pronunciation proficiency levels using HLDA and the total improvement of system performance. By plotting the phonetic posterior probability distribution and the expert scores, the enhancement of discriminability is presented visually for the different systems. Using HLDA, the problem that very high quality pronunciation is usually underestimated and that very low quality pronunciation is usually overestimated, can be mitigated. The speaker correct-rank rate is increased from 85.59% to 90.99% using HLDA. Finally, we accurately determine the improvements of our system in terms of the three evaluation strategies.

#### 4.4.1 Discriminability Improvement

Figures 4 to 7 show the discriminability of the acoustic model in the four systems using two consonants and two vowels in different syllables. In the two-dimensional plot, the x-axis is AFBPP and the y-axis is PLPP. Circles represent poor pronunciation and asterisks represent good pronunciation according to the expert assessors. All the samples are extracted from test corpora, which comprise all non-native speaker utterances. For a context-dependent phone, we calculate its AFBPP and PLPP as described in Sect. 2 for each sample, then draw the samples in the plot. If its expert score is 0, we mark it with a circle; if it is 2, we mark it with an asterisk; if it is 1, we do not mark it. From System 1 to System 4, the correct and incorrect pronunciations become increasingly separated and the differences between phonetic posterior probabilities of different pronunciation quality in-
crease. This suggests that the proposed measures are helpful for distinguishing different-quality pronunciation, and the enhanced discriminatory power of the acoustic model can contribute to the performance of our CALL system.

4.4.2 Prominent Improvement Obtained from HLDA

As previously mentioned, in our CALL system, very high quality pronunciation is prone to be underestimated and very low quality pronunciation is overestimated (the pronunciation quality referred to here is the expert score), i.e., high-quality pronunciation is scored lower than the expert score and low-quality pronunciation is scored higher by our CALL system. This seriously degrades the system performance, because the discriminability of the acoustic model is unsatisfactory; thus, the system cannot accurately distinguish between pronunciation of different quality. We apply HLDA to solve this problem.

To clearly present the improvement, we smooth the distribution of score difference vs pronunciation quality into a curve that shows the changes in score difference with pronunciation quality. From Fig. 8, we see that the slope of the smoothing curve is reduced significantly from System 2 to System 3. In fact, HLDA can project the feature space into another space with fewer dimensions but equivalent discriminative information; thus, the AM processed by HLDA has more powerful discriminability. Therefore, the system with HLDA can assess pronunciation more accurately and alleviate the above-mentioned problem.

To show the effect of HLDA quantitatively, we use the speaker correct-rank rate described in Sect. 4.2 and give the results in Table 1. The speaker correct-rank rate increases from 85.59% to 90.99% for PSK1_eval, PSK2_eval and PSK3_eval.

4.4.3 Evaluation

In Tables 2 and 3, we show the system performances for consonant and vowel pronunciation in terms of the correct
Table 4 The average score difference between expert score and machine score in the four systems.

<table>
<thead>
<tr>
<th>Test set</th>
<th>System 1</th>
<th>System 2</th>
<th>System 3</th>
<th>System 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSK1_eval</td>
<td>7.20</td>
<td>4.93</td>
<td>4.13</td>
<td>3.82</td>
</tr>
<tr>
<td>PSK2_eval</td>
<td>5.68</td>
<td>3.95</td>
<td>3.51</td>
<td>3.25</td>
</tr>
<tr>
<td>PSK3_eval</td>
<td>8.79</td>
<td>5.64</td>
<td>4.53</td>
<td>4.05</td>
</tr>
</tbody>
</table>

Table 5 Improvement of average CC for the four systems between machine and expert scores.

<table>
<thead>
<tr>
<th>System type</th>
<th>Average CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>System 1: baseline</td>
<td>78.00%</td>
</tr>
<tr>
<td>System 2: SD-CMN</td>
<td>84.14%</td>
</tr>
<tr>
<td>System 3: SD-CMN+HLDA</td>
<td>84.62%</td>
</tr>
<tr>
<td>System 4: SD-CMN+HLDA+MAP</td>
<td>86.57%</td>
</tr>
</tbody>
</table>

assessment rate. The performance monotonically improves from System 1 to System 4. The speaker-dependent CMN algorithm improves the correct assessment rate markedly by 14.8% for PSK3_eval. The techniques of HLDA and the MAP algorithm are helpful for enhancing the evaluation accuracy of our pronunciation assessment system.

Table 4 shows that the average score difference between machine and expert scores monotonically decreases from System 1 to System 4. Speaker-dependent CMN improves the performance markedly, and the other measures decrease the average score difference effectively for the three different test sets.

Table 5 shows the changes in the average CC in the four systems for the test sets of PSK1_eval, PSK2_eval and PSK3_eval. For the combined system with speaker-dependent CMN, HLDA and the MAP algorithm, the accuracy of pronunciation quality assessment for strongly accented Mandarin speech is as high as 86.57% in terms of the average CC.

5. Conclusions

In our user-dependent pronunciation quality assessment system, the acoustic model and features play an important role. Thus, they need to be readapted for students whose native languages are characterized by different acoustic spaces. In this paper we present some measures to improve the acoustic model and features. Our novel evaluation method for matching with pronunciation quality scores can improve the performance of our CALL system markedly in terms of four different metrics, i.e., the correct assessment rate, the average score difference, the average CC between machine and expert scores and the speaker correct-rank rate. In particular, HLDA can decrease the large human-machine scoring difference of the speech data when the pronunciation quality is very good or poor. In the future, we would like to investigate the discriminability of the acoustic model further, with the aim of distinguishing some easily confused phones.

References


Fengpei Ge received her B.E. degree from Tianjin University in 2005. She is currently a Ph.D. candidate at ThinkIT Speech Lab, Institute of Acoustics (IOA), Chinese Academy of Sciences (CAS). Her research interests include pronunciation quality assessment, discriminative training algorithm for acoustic modeling, speech signal processing and speech recognition.
Changliang Liu received his B.S. degree from University of Science and Technology of China. He is currently a Ph.D. candidate at ThinkIT Speech Lab, Institute of Acoustics (IOA), Chinese Academy of Sciences (CAS). His research interests are pronunciation quality assessing, speech signal processing and speech recognition.

Jian Shao received his B.S. degree in electrical science and engineering from Nanjing University in 2003. He is currently a Ph.D. candidate at ThinkIT Speech Lab, Institute of Acoustics (IOA), Chinese Academy of Sciences (CAS). His research is focused on spontaneous speech recognition and spoken document retrieval.

Fuping Pan received his Ph.D. in information and signal processing from IOA, CAS, 2007. He is currently an Assistant Researcher at IOA. His research is focused on automatic pronunciation evaluation, speech signal processing and speech recognition.

Bin Dong graduated with a B.S. and M.S. degrees from Shandong University of Science and Technology in 2000 and 2003. In September 2003, he became a Ph.D. candidate at ThinkIT Speech Laboratory, IOA, CAS, and is carrying out research on speech recognition. He is currently an Associate Researcher at ThinkIT Speech Lab. His research interests include automatic pronunciation evaluation, speech signal processing and speech recognition.

Yonghong Yan received a B.E. degree from Tsinghua University in 1990 and Ph.D. degree from Oregon Graduate Institute (OGI). He worked in OGI as an Assistant Professor (1995), Associate Professor (1998) and Associate Director (1997) at the Center for Spoken Language Understanding. He worked at Intel from 1998–2001, where he chaired the Human-Computer Interface Research Council, worked as the Principal Engineer of the Microprocessor Research Lab and as Director of Intel China Research 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 interfaces. He has published more than 100 papers and holds 40 patents.