A Statistical Modeling Approach to Automatic Evaluation of Mandarin Pronunciation

Su-Kui Xu*, Si Wei**, Zhen-Hua Ling*, Qian-Yong Gao**, Li-Rong Dai* and Qing-Feng Liu*,**

SUMMARY: This paper introduces an automatic Mandarin pronunciation evaluation method, which aims at building a computer-based system to partly replace human examiners in the Putonghua Shuiping Ceshi (PSC) in China. This method learns the mapping relationship between the recorded speech waveforms and the score of pronunciation proficiency by a statistical modeling approach, which is composed of three main modules: the frontend module, the evaluation feature extraction module and the mapping module. In the frontend module, hidden Markov model (HMM)-based acoustic models are constructed to describe the distribution of acoustic features for standard pronunciation. In the evaluation feature extraction module, posterior probabilities are calculated for segmental and tonal acoustic features of speech data from each examinee using the trained acoustic models. These posterior probabilities together with a duration feature compose the feature vector for predicting pronunciation scores. Finally, in the mapping module, piecewise linear regression is introduced to map the evaluation feature vector into a pronunciation score for each examinee. The piecewise linear regression is achieved by cascading an SVM classifier and a linear regression for each class in our implementation. An experiment on evaluating the real PSC test data of 5,420 speakers shows that the system constructed using our proposed method achieved a correlation of 0.901 between the predicted scores and the scores given by human examiners for the first three sections of PSC test. Another experiment which compared the performance of our system with 20 human examiners shows that our system ranked 2nd and outperformed most of the human examiners in terms of evaluation accuracy.

Key words: Mandarin pronunciation, automatic evaluation, statistical model, evaluation accuracy

1. Introduction

Mandarin Chinese is the official spoken language in mainland China, which is called Putonghua (普通话) in Chinese. Considering the numerous dialects existing in China and the difficulties of daily communication among people using different dialects, to promote Putonghua is always an important task of the Chinese government. There is an official test organized by the Chinese government to evaluate a speaker’s pronunciation proficiency of Mandarin Chinese, named Putonghua Shuiping Ceshi (PSC, 普通话水平测试) (PSCSG 2004). To pass this test is a prerequisite in order to qualify some professions in China, such as civil servants and school teachers. In the conventional PSC test, the examinee’s speech is evaluated by human experts. The evaluation process is time-consuming considering the large population of people participating the test every year and it is also difficult to guarantee the consistency of scoring among all examiners. Therefore, it would be very helpful if some computer-based methods could evaluate speaker’s pronunciation proficiency of Mandarin Chinese automatically and accurately.

The computer-based PSC test can be considered as a special case of Computer Assisted Language Learning (CALL), which has attracted the attentions of many research teams on speech technology in the last two decades. The Speech Technology and Research Group of SRI investigated the usage of prosody, rhythm and melody of speech for human-computer dialog systems (Ferrer et al. 2002, 2003, Ang et al. 2002). They used prosodic features to detect the end of utterances and the
A Statistical Modeling Approach to Automatic Evaluation of Mandarin Pronunciation

frustration and annoyance in human-computer dialog systems. The Spoken Language Systems Group of Computer Science and Artificial Intelligence Laboratory at MIT proposed that one of the best ways to learn a foreign language is to communicate with a native speaker. Therefore, they developed a spoken dialogue system using automatic speech recognition technology and natural language understanding for learning spoken English (McGraw and Seneff 2007, Cai et al. 2013). The speech group of Cambridge University focused on pronunciation error detection and phone-level pronunciation evaluation (Witt 1999, Witt and Young 2000, Thomson et al. 2010a, 2010b). They adopted posterior probability as the measurement for evaluation. At Centre for Language and Speech Technology of Radboud University, Nijmegen, web tools were developed to teach languages with corrective feedback on pronounced words and short sentences and the acquisition of syntax (Cucchiarelli et al. 2012). Another well-known example is SpeechRater of Educational Testing Service (ETS) (Evanini and Wang 2013, Wang et al. 2013, Zechner et al. 2012), which has been used in the TOEFL Practice Online test since 2006. It can evaluate the pronunciation fluency, detect grammatical facility and measure the ability of English skills. Now it pays more attention to Natural Language Processing (NLP) based features for text-independent language skill evaluation. Regarding Chinese learning, a software named NTU Chinese (Su et al. 2013) has been developed by National Taiwan University, which aimed at training the listening and oral skills of learners of Chinese by analyzing pronunciation, pitch, timing and emphasis of the learners’ speech.

However, there are several special features of the computer-based PSC compared with the common CALL tasks (Wang et al. 2006). First, CALL systems are mainly developed for non-native speakers learning foreign languages while PSC aims at evaluating the proficiency level of native speakers of Mandarin Chinese. Second, the computer-based PSC is required to achieve much higher precision of scoring than common CALL tasks, which should be as accurate as professional examiners to guarantee the fairness and authority of this test. These two features increase the difficulties of developing computer-based PSC systems.

This paper introduces an automatic Mandarin pronunciation evaluation method developed by our team for the practical application of PSC test (Wei et al. 2006, 2010). First, hidden Markov model (HMM)-based acoustic models are constructed to describe the distribution of acoustic features for standard pronunciation. Then, posterior probabilities are calculated for each segment of input speech to represent the possibility of current segment being pronounced correctly. The evaluation features composed of posterior probability features together with a duration feature are mapped towards a score of pronunciation proficiency using a piecewise linear regression, which are learnt from a training database. Experimental results using real PSC test data show that this method can successfully replace human examiner for the first three sections of PSC test.

The rest of this paper is organized as follows. Section 2 briefly reviews the background of PSC test. Section 3 introduces the details of our proposed method. The experimental results and conclusions are given in Section 4 and 5 respectively.

2. Background of PSC test

Putonghua Shuiping Ceshi (PSC) is a language test in China organized by the State Language Commission, the Ministry of Education, and the State Administration of Radio Film and Television since 1997. The aim of this test is to evaluate an examinee’s accuracy and proficiency of using Putonghua. PSC is conducted through an oral test, which is commonly composed of four sections (PSCSG 2004).

1. Reading 100 mono-syllable words within 3.5 minutes. This section evaluates the accuracy of an examinee’s pronunciation of initials, finals and tones in Putonghua.

2. Reading 50 bi-syllable words with 2.5 minutes. This section evaluates the accuracy of the examinee’s pronunciation of tone sandhi, retroflexed finals, neutral tone in Putonghua besides common initials, finals and tones.

3. Reading a 400-character passage within 4 minutes. This section focuses on evaluating the co-articulation, pauses, intonation and fluency of the examinee’s speech.

4. Speaking on a given topic within 3 minutes. It aims at evaluating the examinee’s proficiency level of Putonghua under the condition where no texts are provided.

The full mark of this test is 100. The results are divided into 3 grades (Grades One, Two and Three) and each grade is further divided into two levels (Levels A and B) according to the test scores as shown in Table 1.

In the common scenario of PSC test, each examinee speaks in front of two examiners and each examinee gives an evaluation score following the specifications
3. Automatic mandarin pronunciation evaluation

An automatic Mandarin pronunciation evaluation method for PSC test will be presented in this section. As introduced in previous section, the reference texts of an examinee’s speech are given for the first three sections of PSC test while no texts are provided for the fourth one. This method works for the sections where reference texts of pronunciations are available. To achieve automatic evaluation on the fourth section will be one of the goals of our future work. The flowchart of this method is shown in Figure 1. The inputs contain the speech of an examinee recorded for the first three sections of PSC test together with their reference texts. The output is an evaluation score within the range of 0 to 60, where 60 is the full mark of the first three sections. This method is mainly made up of three modules: The frontend module, the evaluation feature extraction module and the mapping module. In the frontend module, acoustic features are extracted from the speech waveforms and are aligned towards the phone sequences derived from the texts using acoustic models. Then, the boundaries of each phone can be determined. The second module extracts evaluation features from the segmented acoustic feature sequences. These features are sent into the mapping module to be converted into an evaluation score using a regression model. More details of these three modules will be introduced in this section.

3.1 Frontend processing

In this module, acoustic features and phone transcriptions are extracted from the speech waveforms and reference texts respectively. In Chinese, each syllable is composed of a final part and an optional initial part. Initials and finals in Chinese are treated as basic phone units in this paper although they are not strictly equal because some nasal finals and compound finals may be composed of more than one phonemes. Then, the acoustic feature sequences are aligned to the phone transcriptions using acoustic models to determine the phone boundaries.

Here, the function of acoustic models is to describe the distributions of acoustic features for different speech segments with standard pronunciation. In our implementation, the acoustic features consist of Bottle-Neck (BN) features (Grezl et al. 2007) derived from Mel-Frequency Cepstral Coefficients (MFCC) using deep neural networks (DNN) (Dahl et al. 2012, Hinton et al. 2012) and fundamental frequency (F0) at each frame. These two types of features describe the segmental and tonal characteristic of speech respectively. The reason of using BNs instead of conventional
A Statistical Modeling Approach to Automatic Evaluation of Mandarin Pronunciation

MFCCs is that they are derived from a DNN-based phone classifier and are embedded with more phone-class-related information. BNs are derived by setting the node number of a middle layer in the DNN to be much smaller than other hidden layers and applying linear activation function to this layer. The activities of this layer are expected to be informative for classifying input acoustic features into the tied HMM states, and are named as BottleNeck (BN) features (Grezl et al. 2007). The diagram of extracting BN features is shown in Figure 2. Acoustic models for the BN and F0 features are constructed respectively. Similar to the acoustic modeling in automatic speech recognition, hidden Markov models (HMM) are used as the model structure and Gaussian mixture models (GMM) are adopted to describe the distributions of acoustic features at HMM states (Young et al. 2009). The GMM-HMMs of BN features are estimated for each monophone label and the models of F0 features are constructed for each tone type of syllables. During model training, speaker-independent (SI) models are estimated at first using the training data from various speakers. Only the data from the speakers whose PSC scores are higher than a threshold can be used for model training considering that the acoustic models are expected to represent the characteristics of standard pronunciation and will be used to derive evaluation features in the next module. Then, the SI models are converted to speaker-dependent (SD) models to match specific examinee better by model adaptation using his or her speech pronounced during the test. Maximum Likelihood Linear Regression (MLLR) algorithm (Leggetter and Woodland 1995, Gales 1998) is applied to achieve the model adaptation of BN features and the F0 models are adopted by maximum likelihood based feature normalization (Wei 2008). Considering that the examinee may make some mistakes during pronunciation, such adaptation is conducted in a semi-supervised way, which means that only the speech segments with high confidence of being pronounced correctly are selected during the adaptation.

At evaluation time, the phone boundary segmentation is achieved by aligning the BN feature sequences extracted from input speech towards the phone transcriptions by Viterbi decoding using the trained BN models (Young et al. 1989). In order to deal with the possible mispronunciations of examinees, the decoding network is constructed considering typical mispronunciations for each phone. These typical mispronunciations are summarized from the real data of PSC test containing mistakes made by examinees. Figure 3 shows an example of such network. In this example, the initial zh and final i are the correct pronunciation for the Chinese syllable zhi (ㄗ). Considering that zh is often mispronounced as z, ch, or c and i is frequently mispronounced as ie, all these possible mispronunciations are also included within the network. The Viterbi decoding algorithm automatically finds the path of phones within this network which matches the BN features best. Therefore, we can get reliable phone boundaries even if the examinee has made some mistakes during pronunciation.

3.2 Evaluation feature extraction

In the evaluation feature extraction module, a feature vector is calculated from the segmented acoustic feature sequences of the speech data for an examinee. The feature vector is composed of a duration feature and several posterior probability features. The duration feature is calculated as the average duration of each syllable using the segmentation results given by frontend processing, which is considered to be related with the fluency of pronunciation of the examinee. The posterior probability features are derived from the posterior probabilities calculated for each segment of input utterances to represent the possibility of current segment being pronounced correctly.
Posterior probability is a commonly used evaluation feature in CALL systems (Witt 1999). For a phone with BN feature sequence $O=[o_1, o_2, \ldots, o_n]$, the posterior probability is defined as $p(t|O)$, where $t$ is the phone label of correct pronunciation determined by the reference texts. A larger posterior probability means a higher possibility of pronouncing this segment correctly. $p(t|O)$ can be calculated as

$$p(t|O) = \frac{p(O|t)P(t)}{\sum_{q \in Q_t} p(O|q)P(q)} = \frac{p(O|t)}{\max_{q \in Q_t} p(O|q)}$$

where $p(O|t)$ is the likelihood function of generating acoustic feature sequence $O$ from the acoustic model of phone $t$, and $Q_t$ is the set of phones which $t$ might be mispronounced to. In common CALL systems, the phone set $Q_t$ contains all possible phonemes. However, the pattern of mispronunciation in PSC is more compact because the examinees are native speakers. Therefore, only the typical mispronunciations of phone $t$ are used to construct $Q_t$. As a measurement of pronunciation proficiency, the posterior probabilities should be robust to noise conditions during the pronunciation of examinees. However, we find that the posterior probabilities calculated using (1) always decrease when some noise signals are captured together with the examinee’s speech during recording. In order to deal with this problem, a CDF-matching process is conducted on the calculated posterior probabilities to compensate the influence of noisy conditions (Wei et al. 2010). The posterior probabilities in (1) can also be calculated for F0 features where $O$ becomes the F0 feature sequence of a syllable and $t$ is the correct tone type of this syllable.

Then, for each syllable pronounced by the examinee, $p(t_{\text{init}}|O_{\text{bn-init}})$, $p(t_{\text{fini}}|O_{\text{bn-fini}})$ and $p(t_{\text{tone}}|O_{\text{f0}})$ are calculated respectively, where $t_{\text{init}}$, $t_{\text{fini}}$ and $t_{\text{tone}}$ denote the initial label, the final label and the tone type of this syllable, $O_{\text{bn-init}}$ and $O_{\text{bn-fini}}$ are the BN feature sequences of the initial part and the final part of this syllable, and $O_{\text{f0}}$ denotes the F0 feature sequence of this syllable. Then, the posterior probability feature of each syllable is defined as the average of $p(t_{\text{init}}|O_{\text{bn-init}})$, $p(t_{\text{fini}}|O_{\text{bn-fini}})$ and $p(t_{\text{tone}}|O_{\text{f0}})$. For the first three sections of PSC test, the posterior probability features of all syllables within each section are further averaged to generate one dimension posterior probability feature for this section.

3.3 Evaluation score mapping

Once the evaluation feature vector of one examinee is given, we can map it towards an evaluation score using a regression model. A piecewise linear regression method is adopted here considering it can achieve better accuracy than the globally linear regression method. This regression method is composed of two steps. First, the evaluation feature vector of one examinee is classified into several categories using a multi-class Support Vector Machine (SVM) classifier (Cortes and Vapnik 1995). Second, for each category, a linear regression model is constructed to convert the evaluation features into the final evaluation score. The SVM classifier and the linear regression models are estimated using a large amount of PSC data which contains both speech recordings and corresponding scores given by human examiners for each section.

4. Experiments

4.1 System construction

An automatic pronunciation evaluation system for PSC test was constructed following the method introduced in Section 3. This system can give the evaluation score of the first three sections of PSC test for an examinee. After the fourth section was evaluated manually, the overall score of the examinee can be determined.

In order to build the acoustic models for frontend processing and evaluation feature extraction, a database consisted of 30,000 speakers’ recordings during PSC test were adopted as the training set. The PSC scores of all these speakers were above 90 to ensure the acoustic models can represent the standard pronunciation. This database was also gender- and text-balanced. The waveforms were recorded in 16kHz/16 bit format and the total duration was about 2,000 hours. 13-dimensional MFCCs with their first and second order derivatives and 1-dimensional F0 features were extracted from the waveforms at a frame shift of 10ms. The MFCC features were converted to the same 39-dimension BN features using a 5 hidden-layer discriminative DNN. A 5-state HMM structure with 16-mixture GMM at each state was adopted to train the BN models for each monophone and F0 models for each syllable. The model parameters were optimized under maximum likelihood criterion.

A database of 10,000 examinees with their corresponding evaluation scores were used to train the regression model. The score of each examinee was determined by averaging the scores given by three expert examiners. In this database, the training samples
of different proficiency levels were well balanced. At first, an SVM classifier was built to classify the evaluation feature vector of each speaker into four categories, which corresponded to Grade One Level B or better, Grade Two Level A, Grade Two Level B, and Grade Three Level A or worse, respectively. Then, a linear regression model was trained for each category using the training samples belonging to this category. Some estimated linear regression coefficients corresponding to different evaluation features for these four categories are summarized in Table 2. The variations of the coefficients in different categories show the globally non-linear property of the evaluation score mapping in our system.

4.2 Performance metrics

Two metrics were used to measure the performance of an examiner, which were correlation coefficient and average score difference. The correlation coefficient measures the consistency between scores given by the examiner and the reference scores, which is calculated as

$$r = \frac{\sum_{m=1}^{M} (S_m - \bar{S})(S_{R_m} - \bar{S}_R)}{\sqrt{\sum_{m=1}^{M} (S_m - \bar{S})^2 \times (S_{R_m} - \bar{S}_R)^2}}$$  \hspace{1cm} (2)

where $M$ is the total number of examinees being evaluated, $S_m$ denotes the score of examinee $m$ given by the examiner, $S_{R_m}$ is the reference score of examinee $m$, $\bar{S}$ and $\bar{S}_R$ are the averages of $S_m$ and $S_{R_m}$ respectively.

The average score difference describes the difference between the scoring results of an examiner and the reference scores, which is defined as

$$d = \frac{1}{M} \sum_{m=1}^{M} |S_m - S_{R_m}|$$  \hspace{1cm} (3)

where $m$ and $S_{R_m}$ are the same as the ones used in (2).

4.3 Experimental results

In the first experiment, a real PSC database containing the speech data and corresponding evaluation scores of 5,420 examinees was used as the test set. The examinees came from different provinces in China, including JiangXi, HuBei, ShanDong, ShangHai, Zhejiang, etc. The performance of the constructed automatic pronunciation evaluation system was evaluated by calculating the correlation coefficient and the average score difference on the test set. The scores given by human examiners for these 5,420 examinees were used as the reference scores. The performance of this system was compared with a baseline system. The differences between these two systems are listed in Table 3 and their performances are shown in Table 4. From Table 4, we can see that the proposed system achieved better performance than the baseline system on both metrics, which shows the effectiveness of using BN features during frontend processing, applying noise compensation on posterior probabilities during evaluation feature extraction, and adopting piecewise linear regression instead of globally linear regression during evaluation score mapping. Besides, the evaluation scores given by the proposed system were very close to the reference scores. The correlation coefficient was higher than 0.9. There were only 7% examinees whose score differences given by the proposed system were larger than four points in this experiment. One important reason causing these large score differences is that the regression model used for evaluation score mapping tended to predict scores close to the middle of a category. Therefore, it is difficult to accurately predict the scores of the examinees who have very good or very bad pronunciation.

Another experiment was conducted to further compare the performance of the proposed automatic
evaluation system with the performance of human examiners. In this experiment, 20 human examiners were asked to evaluate the speech data of 100 examinees who were randomly selected from the 5,420 speakers used in the previous experiment. Then, for each one of the 100 examinees, the reference score was obtained by averaging the scores given by 20 examiners after removing the highest and lowest ones. The correlation coefficients and the average score differences were calculated for each human examiner and the proposed system. The results are shown in Figure 4 and Figure 5 respectively. The correlation coefficient of the proposed system was 0.930. This value is very close to the average correlation coefficient of human examiners, which was 0.947 in this experiment. Our proposed system achieved better performance on the metric of average score difference. The reason is that the regression model used for evaluation score mapping was trained to minimize the mean square error of score prediction, which is more similar to the metric of average score difference than to the correlation coefficient. As shown in Figure 5, the average score difference of our proposed system was 0.99, which was only higher than the best human examiners whose average score difference was 0.84 and outperformed the other 19 human examiners. All these results demonstrate that our proposed system can achieve similar performance to that of human examiners on the first three sections of PSC test.

4.4 Practical applications

Since 2007, the automatic pronunciation evaluation system developed by our team has been authorized by the State Language Commission and the Ministry of Education of China to replace human examiners for scoring the first three sections of PSC test. The method introduced in this paper was adopted by the latest version of this system. In 2007, about 120,000 examinees got evaluated by this system. This number increased to about 3,620,000 in 2013. Nowadays, our system has been widely and successfully applied in almost all provinces of China.

5. Conclusions

This paper presents an automatic Mandarin pronunciation evaluation method for PSC test. Compared with common CALL tasks, the computer-based PSC is more difficult because the accuracy of scoring is required to be comparable with human examiners. The proposed method follows a statistical modeling approach. Acoustic models are trained using a large amount of training data with good pronunciation to segment input speech waveforms and to calculate evaluation features.
A regression model is also built using the real PSC data so as to map the evaluation features towards an evaluation score for each examinee. Our experimental results demonstrate that the system built using our proposed method can evaluate the first three sections of PSC test accurately. Its correlation coefficient is close to the average performance of human examiners and its average score difference is smaller than 19 of the 20 human examiners participated in our experiments. At current stage, human examiners are still necessary to evaluate the fourth section of PSC test because our proposed method can’t handle the situation where the reference texts are not provided. One possible solution is to generate the text transcription from speech waveforms using a speech recognizer. However, the accuracy of the state-of-the-art speech recognition systems is still not satisfactory, especially under noisy conditions. The recognition errors may degrade the performance of the proposed pronunciation evaluation method significantly. To investigate solutions to the automatic evaluation of the fourth part of PSC test and to extend the proposed system to evaluating the second language learners of Mandarin Chinese will be the tasks of our future work.

References


Witt, S. M. and S. J. Young (2000) “Phone-level pronuncia-


(Received Nov. 11, 2014, Accepted Apr. 11, 2015)