Computational Linguistics and Chinese Language Processing
Vol. 12, No. 4, December 2007, pp. 443-458

Automatic Pronunciation Assessment for Mandarin Chinese: Approaches and System Overview

Jiang-Chun Chen*, Jyh-Shing Roger Jang*, and Te-Lu Tsai+

Abstract

This paper presents the algorithms used in a prototypical software system for automatic pronunciation assessment of Mandarin Chinese. The system uses forced alignment of HMM (Hidden Markov Models) to identify each syllable and the corresponding log probability for phoneme assessment, through a ranking-based confidence measure. The pitch vector of each syllable is then sent to a GMM (Gaussian Mixture Model) for tone recognition and assessment. We also compute the similarity of scores for intensity and rhythm between the target and test utterances. All four scores for phoneme, tone, intensity, and rhythm are parametric functions with certain free parameters. The overall scoring function was then formulated as a linear combination of these four scoring functions of phoneme, tone, intensity, and rhythm. Since there are both linear and nonlinear parameters involved in the overall scoring function, we employ the downhill Simplex search to fine-tune these parameters in order to approximate the scoring results obtained from a human expert. The experimental results demonstrate that the system can give consistent scores that are close to those of a human's subjective evaluation.

Keywords: CAPT, CALL, Speech Recognition, Tone Recognition, Speech Assessment, GMM, Mandarin Chinese, Downhill Simplex Method, Phoneme, Intensity, Rhythm, Forced Alignment.

1. Introduction

With the fast-growing power of personal computers and the advances in speech and language processing technologies, software systems for CALL (Computer Assisted Language Learning) now allow a person to learn a language by interacting solely with computers, especially for second language (L2) learning. In general, a CALL system involves testing procedures for both the receptive and productive skills of a given subject. To evaluate receptive skills such as

* Department of Computer Science, National Tsing Hua University, Taiwan
E-mail: {jtchen; jang}@cs.nthu.edu.tw
+ Innovative DigiTech-Enabled Applications & Services Institute, Institute for Information Industry

[Received December 7, 2006; Revised October 16, 2007; Accepted October 17, 2007]

reading and listening, the procedure is relative simple, since the evaluation is usually based on exams containing questions of single or multiple choices. On the other hand, to evaluate the productive skills of speaking or writing, the procedure is relatively difficult and time-consuming, since a human expert is usually required to evaluate the speech or writing in a subjective and time-consuming manner. With advances in automatic speech recognition, a Computer-Assisted Pronunciation Training (CAPT) system can evaluate the pronunciation quality using various speech features, and provides high-level feedback (hints for further improvement, *etc.*) to the user. Successful applications of CAPT have been reported in the literature [Neumeyer *et al.* 2000; Kim and Sung 2002; Neri *et al.* 2003].

In this paper, we propose several algorithms for constructing a CAPT system that can assess a test utterance in Mandarin Chinese with respect to a target utterance by a native speaker. Basically, there are four evaluation criteria based on different acoustic features, as explained in the following.

1. **Phoneme:** This is based on the log probabilities of the test utterance with respect to the acoustic models derived from a large speech corpus for speaker-independent speech recognition. Note that the target utterance is not required for this evaluation.
2. **Tone:** Each syllable is associated with a tone in Mandarin Chinese. The pronounced tone of a syllable can be identified by a tone classifier, and the result is then compared with the correct tone for evaluation. Note that we can obtain the correct tones from the text of the utterance; hence, the target utterance is not used directly for this evaluation.
3. **Intensity:** Each syllable has an intensity vector, which is compared to that of the corresponding syllable in the target utterance to ensure it has a similar score.
4. **Rhythm:** The duration of each syllable and the silence in between are compared to those of the target utterance to ensure they have a similar score.

Each of the scoring functions for the above four criteria involves several nonlinear parameters. These four scoring functions are then linearly combined to give a score between 0 and 100. We then employ a search method that can find optimum values of these parameters such that the computed scores can approximate those of a human expert. The experimental results demonstrate the feasibility of the proposed approach, which can give consistent results when compared with human evaluation.

The rest of this paper is organized as follows. Section 2 gives a quick overview of related work on automatic pronunciation assessment. Section 3 explains the speech-related techniques used in our approach, including Viterbi decoding and tone recognition. The method of combining weighted scores based on derivative-free optimization is also explained. Section 4

describes the GUI of our system. Section 5 demonstrates the experimental results and Section 6 gives concluding remarks.

2. Related Work

In general, a CAPT system can evaluate pronunciation quality using various speech features. Moreover, the system is expected to have optimum performance by minimizing the discrepancies between the scores from computers and those from a human expert. However, most of the reported systems do not take the characteristics of tonal languages into consideration. In particular, Mandarin Chinese is a tonal language, and each character is associated with one out of five possible tones. The tone of a given character is also context-dependent according to tone sandhi [Lee 1997]. Hence, the correct pronunciation of the tone of each character in a sentence is the most challenging problem for a Mandarin-learning non-native speaker. The proposed system takes this specific problem into consideration and tries to create a comprehensive Mandarin Chinese pronunciation assessment system.

3. The Proposed Approach

The proposed pronunciation assessment system uses four criteria to evaluate a test utterance with respect to a target utterance. The algorithms of these four criteria are explained in this section.

3.1 Syllable/Phone Segmentation using HMM-based Forced Alignment

HMM (Hidden Markov Models) has been used for speech recognition with satisfactory performance over the past few decades [Rabiner and Juang 1993; Huang *et al*. 2001]. Our system employs a speaker-independent HMM-based recognition engine, which was trained on a balanced corpus of Mandarin Chinese recorded by 70 subjects in Taiwan. Each speech feature vector contains 39 dimensions, including 12 MFCC (Mel-Frequency Cepstral Coefficients) and 1 log energy, along with their delta and double delta values. 174 right-context-dependent (RCD) biphone models are derived from the speech corpus. In other words, two phones are regarded as different models if their right phones are different. For example, the right phone of “a” in the syllable “dan” is “n”, while the right phone of “a” in the syllable ‘jiang’ is “ng”. As a result, the phone “a” in the syllable ‘dan’ is defined as “a+n” in the RCD context, to distinguish it from, say, “a+ng” in the syllable “jiang”. The number of RCD biphone models is much larger than that of the context-independent monophone models, thus a large corpus is required for the reliable training of RCD biphone models. Furthermore, we have also designed an efficient pruning method for our speech recognizer [Jang and Lin 2002].

For pronunciation assessment, we need to build a lexicon net consisting of the models of the uttered text. Then, Viterbi decoding is used to do forced alignment between the speech signals and the models in the lexicon net. The final results include frame indices of isolated syllables/phones and the corresponding log probability. The log probability is an absolute measure of how closely the utterance matches the acoustic models identified from the speech corpus. Consequently, the log probability varies considerably among different models due to their different phonetic characteristics, and thus cannot be used directly for phoneme assessment. Instead, we use a ranking-based confidence measure to be explained later.

To deal with characters having multiple pronunciations in Mandarin, we use a sausage-like lexicon net. Take the sentence "朝辭白帝彩雲間" in Tang Poetry for example, where the third character can be pronounced as "bai" or "bo". Both pronunciations are commonly used. Therefore, our lexicon net has two branches for this character to accommodate both pronunciations, as shown in Figure 1 where Hanyu Pinyin is used for phonetic transcription.

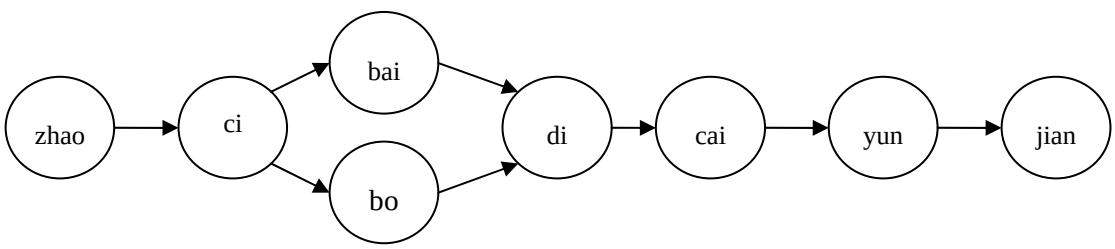

Figure 1. The lexicon net for "朝辭白帝彩雲間".

Figure 2 shows a typical result of forced alignment for the sentence "但使龍城飛將在" in Tang Poetry. The solid lines in the waveform plot indicate the boundaries of phones. The score for each syllable is labeled under each Chinese character, while the score for each phone is labeled under the phone name. The scores depend on the quality of the pronunciation as well as the correctness of forced alignment. In particular, we can see that the fricative phone "sh" in the second character is correctly segmented with a score of 100, while the phone "ch" in the forth character is badly segmented (due to its mispronunciation as a non-retroflexed consonant) with a score of 29. The details of phoneme assessment will be described in the next section.

Figure 2 also shows the pitch vector of this utterance. The dotted lines represent the pitches of the voiced parts, while the union of the dotted and the solid lines represents the pitch curve of the whole utterance derived by dynamic programming. Details of the pitch tracking method can be found in Chen and Jang [2007].

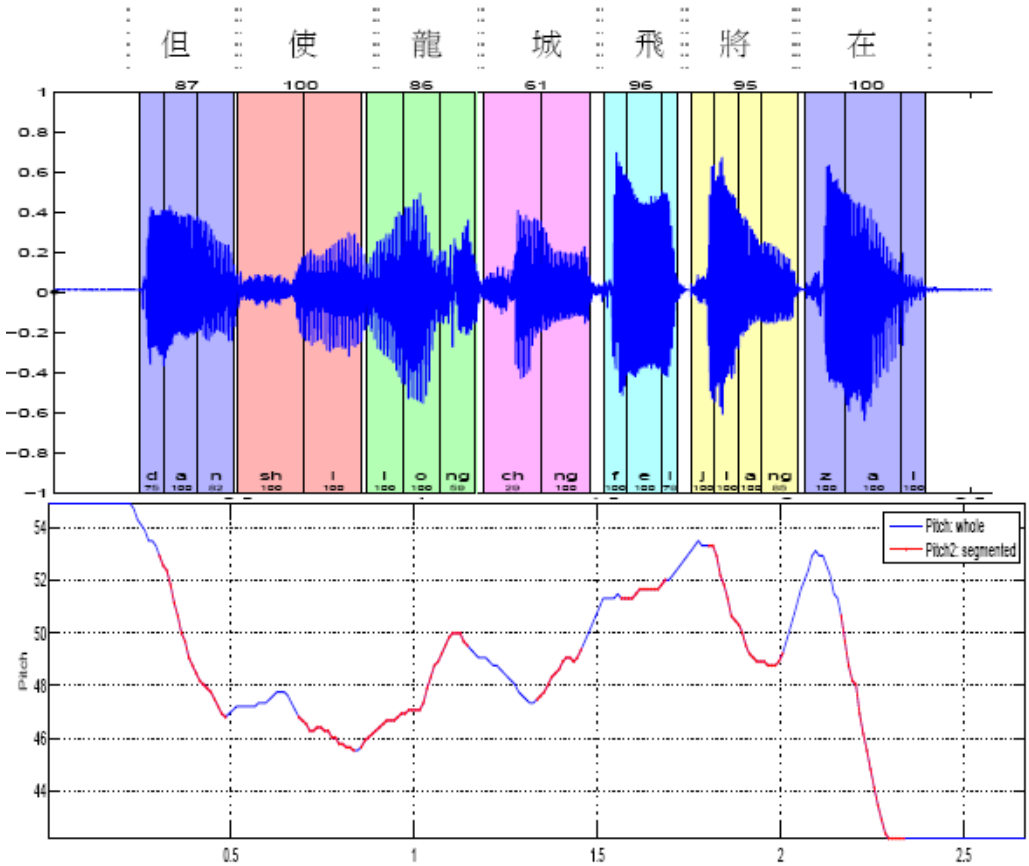

Figure 2. An example of forced alignment for the test utterance "但使龍城飛將在". The upper panel shows the phone segmentation results and the corresponding phoneme scores. The bottom panel plots the pitch vector where the dotted lines correspond to the voiced parts in the utterance.

3.2 Ranking-Based Confidence Measure for Phoneme Assessment

The log probability represents an absolute measure of how closely a pronunciation approximates a given phone model, which does not take into consideration the effect of other competing models. As a result, the log probability varies considerably among different phone models due to their different phonetic characteristics. To deal with this problem, we used a relative measure based on the ranking among all competing biphone models. This is an improved version of our previous approach to confidence measure, based on the ranking among 411 syllables in Mandarin [Chen *et al.* 2004]. By using phone-based ranking, our

system is able to track down phone-level pronunciation errors for detailed and better assessment.

The phone-based phoneme assessment proceeds as follows.

1. For a given biphone model of "x+y", we define the set of competing models as "*+y" where * is a wildcard representing all the possible phones that form a legal biphone with y.
2. After forced alignment, we can obtain the speech signals corresponding to the biphone "x+y". We then send the speech signals to the competing models for a log probability evaluation and find the rank (zero-based) of "x+y" in the competing models.
3. Since each biphone has a different set of competing models, we divide the rank of "x+y" by the size of its competing models to obtain a rank ratio between 0 and 1. Once the rank ratio is obtained, the phoneme score of the *i*-th phone in an utterance is then determined by the following formula:

$$s_{\text{phoneme},i} = \frac{100}{1+\left|\frac{rr_i}{a}\right|^b}, \quad (1)$$

where rr_i is the rank ratio of the i th biphone, and *a* and *b* are the tunable parameters of this scoring function. In particular, when rr_i is equal to 0, a perfect score of 100 is obtained. On the other hand, if rr_i is larger, the score is lower. The values of parameters *a* and *b* are empirically set to 0.1 and 2 respectively. A typical plot of function $s_{\text{phoneme},i}$ is shown in Figure 3. (An on-going research focus is to make the parameters *a* and *b* model-dependent and to determine their values automatically, which will be covered in our future publication.)

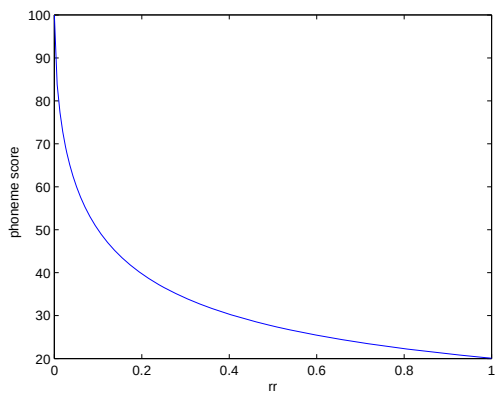

Figure 3. A typical plot of s_{phoneme} with respect to rr_i, where a and b are set to 0.1 and 2 respectively (For simplicity, we have removed the subscript i from the x-axis label of the plot).

The overall score of a test utterance with m phones can be expressed as a duration-weighted average of all of the phones' scores, as follows:

$$score_{\text{phoneme}} = \sum_{i=1}^{m} \frac{dur_i^p}{\sum_{j=1}^{m} dur_j^p} s_{\text{phoneme},i} \,, \tag{2}$$

where dur_i is the duration of the i th phone in test utterance and the weighting factor $\frac{dur_i^p}{\sum_{j=1}^{m} dur_j^p}$ is parameterized by p. In other words, in the test utterance, a phone with a longer duration will have a larger weighting factor. (For a clearer notation definition, we shall use p_{phoneme} to denote the free parameter p in the subsequent discussion.) Similar usage of log-probability is commonly adopted in the research of utterance verification [Sukkar and Lee 1996].

3.3 Tone Recognition Using GMM

Mandarin Chinese is a tonal language and each character is associated with a syllable (out of 411 possible syllables) and a tone (out of 5 possible tones). The tone/pitch information plays an important part in assessing a given utterance. As a result, we need to use isolated syllables for tone recognition.

Previous work of tone recognition for Mandarin Chinese is briefly described. [Chen and Wang 1993] classifies tones based on neural networks. More recently, Lin and Lee [2003] introduces a new set of inter-syllabic features to identify tones. To determine the tones of a given utterance, we use the pitch vector of each syllable segmented via the aforementioned forced alignment. Each pitch vector of a syllable is identified via the autocorrelation method [Huang *et al*. 2001], and then expanded into the Chebyshev polynomials, as explained in Li [2002].

In our experiment, a pitch vector is normalized to the interval [-1, 1] at the time axis and then approximated by the Chebyshev polynomials of degree 6. Moreover, before using the Chebyshev approximation, we have subtracted the mean from the pitch vector of a syllable. In other words, each pitch vector has a mean value of zero before the Chebyshev approximation is applied. As a result, we do not need to use $coef_0$ of the Chebyshev polynomials for tone recognition. Hence, the dimension of the feature vector for the tone recognizer is 5. In fact, more features can be employed for tone recognition, such as the volume of the syllable, the slope of the pitch vector, the duration of a syllable *etc*., as suggested in Huang [2006].

The polynomial coefficients are then used as the feature vectors for the GMM (Gaussian Mixture Model) tone classifier. We used the Tang Poetry microphone corpus [Tang Poetry 2002] of 3211 utterances recorded in our lab for the experiment. The corpus was recorded by eight males and two females, with a sample rate of 16 kHz and a bit resolution of 16 bits. To train the GMM, 2500 utterances (15144 syllables) are used as the training data and the other 711 utterances (4422 syllables) are used as the test set. All the utterances were segmented into syllables via forced alignment. We discarded syllables with durations of less than four frames or more than sixty frames, since the alignment might have been performed incorrectly on these syllables. The result demonstrates that when the number of Gaussian density functions for each tone is 128, we can obtain a recognition rate of 80.25%. For tone recognition, we only consider the most common four tones, the high flat (tone 1), the low rising (tone 2), the high low rising (tone 3), and the high falling (tone 4). Unlike these four lexical tones, the neutral tone does not have a specific pitch pattern; therefore, it is easily confused with the other four tones. As syllables with the neutral tone usually are shorter in duration and lower in energy, short-time energy is found useful in addition to the apparent key features derived from pitch-frequency contours [Lee 1997]. Considering the similarity in the lower energy between tone 1 and tone 5, we put tone 5 in the same class as tone 1 in our experiment. Table 1 lists the confusion matrix of the tone recognition result, where we can observe that tone 3 is mostly likely to be misclassified as tone 4. This is because the rising end of tone 3 is usually missing when the speech rate is high, causing tone 3 to be confused with tone 4.

Table 1. Confusion matrix of the test set for tone recognition

Recognized / Answer	Tone 1	Tone 2	Tone 3	Tone 4	Recognition Rate
Tone 1	1079	87	14	61	86.95%
Tone 2	105	961	89	54	79.49%
Tone 3	39	108	334	163	51.86%
Tone 4	65	27	32	1055	89.48%

Once the pitch vector of syllable *i* is classified into one of the four possible tones, the rank of the correct tone is converted into a score by the following equation:

$$s_{tone,i} = \frac{1-\frac{r_i}{3}}{1+k\cdot\frac{r_i}{3}}\cdot 100, \qquad (3)$$

where *k* is a tunable parameter and r_i is the rank of the desired tone for syllable *i*. When r_i =0 (where the desired tone appears at the top of the output ranking list), we have a perfect

score of 100. On the other hand, if r_i =3 (where the desired tone appears at the bottom), we have a score of 0.

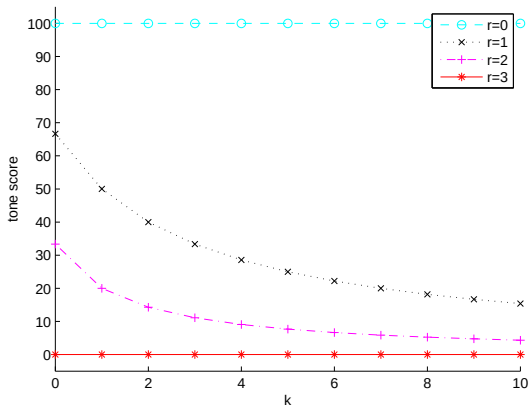

Figure 4. Tone score curves for r_i =0, r_i =1, r_i =2 and r_i =3, respectively, with the value of k varying from 0 to 10. (We have removed the subscript i from the x-axis label for simplicity.)

Figure 4 shows the tone score curves for r_i =0, r_i =1, r_i =2 and r_i =3, respectively, with the value of *k* varying from 0 to 10. The overall score of an utterance with c syllables is once again computed as the average of each syllable's score weighted by its duration:

$$score_{\text{tone}} = \sum_{i=1}^{c} \frac{dur_i^p}{\sum_{j=1}^{c} dur_j^p} s_{\text{tone},i}\,, \qquad (4)$$

where dur_i is the duration of the i th syllable in the test utterance and the weighting factor $\frac{dur_i^p}{\sum_{j=1}^{c} dur_j^p}$ is parameterized by *p* (For a clearer notation definition, we shall use k_{tone} and p_{tone} to denote the free parameters *k* and *p*, respectively, in the subsequent sections).

3.4 Intensity

Intensity and rhythm are two other important factors in forming the prosody of an utterance. We shall describe how to determine the score of the intensity curves for measuring similarity in this subsection. The treatment of rhythm is covered in the next subsection.

Intensity is also referred to as the magnitude or the volume of a given utterance, which is an important cue for pronunciation and its assessment [Chen *et al*. 2004]. Since the intensity

of a given text can only be found in the target utterance, the similarity score is defined between the intensity curves of the test and the target utterances. In comparison, the phoneme score is obtained from the acoustic model and the tone score is obtained from the tone classification; both scores do not require the use of a target utterance.

In order to account for the variation in microphone gain, we need to normalize the signal amplitude of the test utterance before computing the intensity score. This is achieved by the least-squares estimate [chapter 5 of Jang *et al.* 1997] to find the best scaling factor on the test utterance, such that the squared error between the test and the target utterances is minimized. More formally, we define $\mathbf{r}=[r_1,r_2,....,r_N]$ and $\mathbf{t}=[t_1,t_2,....,t_N]$ as two intensity vectors of the reference (or target) and test utterances, respectively, after length normalization via interpolation. Then, the best scaling factor is computed, such that the error between the intensity vector of the reference utterance and the scaled version of the test utterance is minimized. The error can be expressed as $\|e\|^2=\|\mathbf{r}-\mathbf{t}\theta\|^2$, which is minimized when $\theta=\hat{\theta}=\left(\mathbf{t}^T\mathbf{t}\right)^{-1}\mathbf{t}^T\mathbf{r}$. We then use the optimum value $\hat{\theta}$ to compute the scaled intensity vector of syllable i in the test utterance. Specifically, the dissimilarity (distance) of syllable i is $\|e_i\|^2=\left\|\mathbf{r}_i-\mathbf{t}_i\hat{\theta}\right\|^2$, where $\mathbf{r}_i$ is the intensity vector of syllable i in the reference utterance, and $\mathbf{t}_i$ is the intensity vector (after interpolation to have the same length as $\mathbf{r}_i$) of syllable i in the test utterance. The intensity score of syllable i is then computed via the following equation:

$$s_{\text{intersity,i}}=\frac{100}{1+k\|e_i\|^2}, \tag{5}$$

where *k* is a tunable parameter. When $\|e_i\|^2=0$, we have a perfect score of 100. On the other hand, if $\|e_i\|^2$ is large, the score will be small. Then, the overall intensity score of an utterance with c characters is computed as a weighted average:

$$score_{\text{intensity}}=\sum_{i=1}^{c}\frac{dur_i^{\,p}}{\sum_{j=1}^{c}dur_j^{\,p}}s_{\text{intersity},i}, \tag{6}$$

where dur_i is the duration of syllable i in the test utterance and the weighting factor $\frac{dur_i^{\,p}}{\sum_{j=1}^{c}dur_j^{\,p}}$ is parameterized by *p*. (For a clearer notation definition, we shall use $k_{\text{intensity}}$ and $p_{\text{intensity}}$ to denote the free parameters *k* and *p*, respectively, in the subsequent sections.)

3.5 Rhythm

We can define the rhythm as the duration vector (obtained from the forced alignment) of all syllables, including the short pause between any two syllables. For an utterance with c syllables, we can obtain a duration vector of size 2c-1, including the durations of c syllables and c-1 short-pauses. We define $\mathbf{p} = [p_1, p_2, ..., p_{2c-1}]$ and $\mathbf{q} = [q_1, q_2, ..., q_{2c-1}]$ as two duration vectors for the reference (target) and test utterances, respectively. The distance between these two duration vectors can be defined as the normalized sum of the absolute difference:

$$dist(\mathbf{p}, \mathbf{q}) = \frac{1}{2c-1} \sum_{i=1}^{2c-1} \left| p_i - q_i \right| / \, p_i \ . \tag{7}$$

The score for measuring rhythm is computed by the following equation:

$$score_{\text{rhythm}} = \frac{100}{1 + k \cdot dist(\mathbf{p}, \mathbf{q})} \tag{8}$$

where k is a tunable parameter. We denote the free parameter k as k_{rhythm} for clarity in later discussions.

A summary of the proposed four evaluation criteria and their corresponding speech/acoustic features, target of comparisons, and dissimilarity/distance measure is shown in Table 2.

Table 2. Summary of evaluation criteria and their corresponding speech/acoustic features, target of comparisons, and dissimilarity/distance measure.

Criteria	Speech/Acoustic Features	Target of Comparisons	Dissimilarity/ Distance Measurement
Phoneme	MFCC	Acoustic models	Ranking of the desired phoneme
Tone	Pitch	GMM-based tone classifier	Ranking of the desired tone
Intensity	Energy	Intensity vector of the reference utterance	Euclidean distance of scaled normalized intensity vectors
Rhythm	Duration	Durations of each syllable and short-pause of the reference utterance	Normalized absolute sum of difference in duration

3.6 Parametric Scoring Function

As mentioned in the previous subsections, we have obtained four scores based on phoneme, tone, intensity and rhythm. The overall scoring function is defined as the weighted average of four scores:

$$score = w_1 \cdot score_{\text{phoneme}} + w_2 \cdot score_{\text{tone}} + w_3 \cdot score_{\text{intensity}} + w_4 \cdot score_{\text{rhythm}}, \quad (9)$$

where $w_1 + w_2 + w_3 + w_4 = 1$. Apparently, the overall scoring function is formed within several parameters, including p_{phoneme}, k_{tone}, p_{tone}, $k_{\text{intensity}}$, $p_{\text{intensity}}$ and k_{rhythm}, and w_1, w_2, w_3, w_4. To fine-tune these parameters to approximate the scores by the human expert, we employ the downhill Simplex method [chapter 7 of Jang *et al*. 1997] to find the optimal values of these parameters. Basically, the downhill Simplex method is a derivative-free optimization method, which is less efficient than some methods, but simple and flexible in implementation. The flowchart of each of the evaluation processes is shown in Figure 5. The experimental results are covered in the next section.

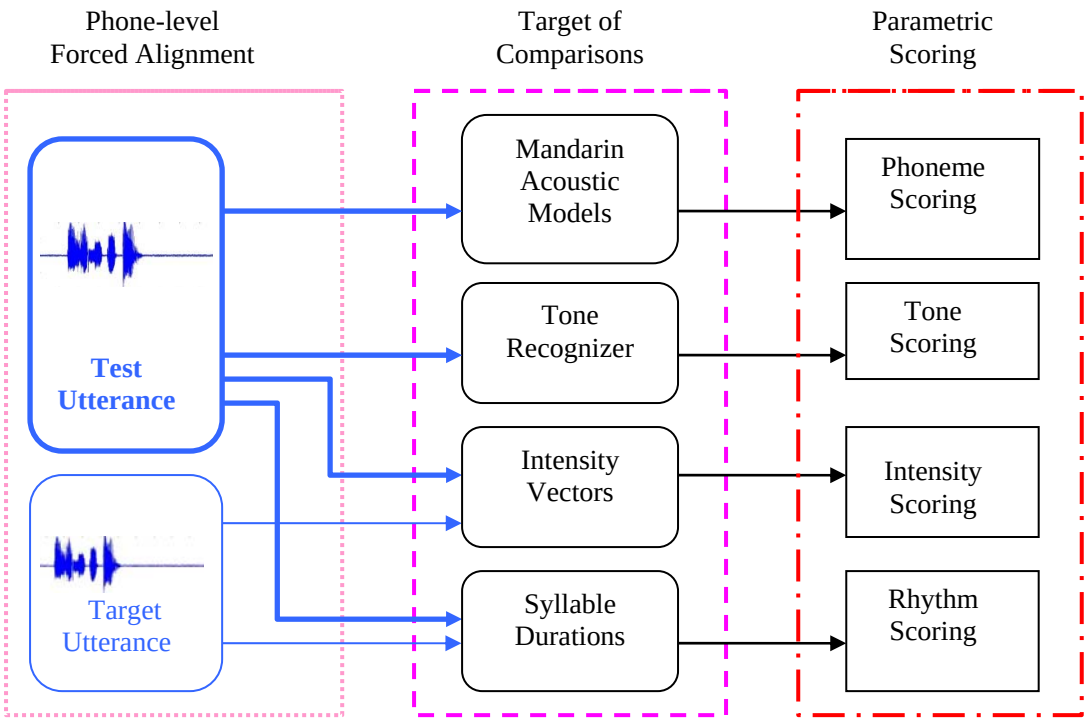

Figure 5. The flowchart of the proposed approach.

4. Experimental Results

To construct the scoring function, we used a dataset containing 400 utterances from 20 speakers, including 10 males and 10 females, each with various levels of proficiency in Mandarin Chinese. Each speaker was asked to utter 20 sentences chosen from the 300 most famous poems of the Tang dynasty. Some of the utterances are purposely pronounced

incorrectly in either content or tone to give "low-score" examples of the training data. These utterances were evaluated by a human expert who subjectively gave a score between 0 and 100 to each utterance, according to the "correctness". We then used the downhill Simplex method to fine tune the parameters of w_1, w_2, w_3, w_4, p_{phoneme}, k_{tone}, p_{tone}, $k_{\text{intensity}}$, $p_{\text{intensity}}$ and k_{rhythm}. The resulting value of w_1 for *phoneme* was 0.52, w_2 for *tone* was 0.22 and w_3 for *intensity* was 0.1 and w_4 for *rhythm* was 0.16, indicating that phoneme and tone were more important factors for speech assessment. This is also consistent with the observation that an utterance with wrong phonemes or wrong tones is much more easily recognized than an utterance with wrong intensity or rhythm.

Table 3. Confusion matrix in terms of three categories.

Machine / Human	Unit: Number of sentences		
	Good	Medium	Bad
Good	121	6	5
Medium	44	67	7
Bad	28	10	112

To evaluate the performance of the system, another set of 400 utterances recorded from 10 subjects was used for an outside test. Each utterance was assigned a category out of three candidates: good (above 80), medium (between 60 and 80), and bad (below 60). Table 3 lists the test results in the form of a confusion matrix in which each column corresponds to a category assigned by our system, and each row corresponds to a category assigned by the human expert. The median category has a smaller score range of [60, 80], therefore the data count is also lower.

In the above table, it is obvious that our system can match the categories assigned by a human expert in a satisfactory manner. The overall recognition rate in terms of these three categories is (121+67+112)/400 = 75%. In addition, the average of the absolute difference between scores from the computer and the human is 5.42, where the standard deviation is 2.31. In a related work, Kim and Sung [2002] reported a recognition rate of about 60% for intonation assessment in English learning, while Neumeyer *et al*. [2000] reported a machine-to-human correlation of 70% for American speakers learning French. Note that the result by Neumeyer is evaluated in the speaker level in their posterior scoring approach, which is different from our evaluation method.

Our goal in this study is to identify effective features/parameters that can approximate the scoring of a single human expert. Actually, having two or more human experts can definitely help shape our system in a more reliable manner. One way to take advantage of multiple human experts is to create a scoring system for each of them, and then combine the results by voting or weighted average. This will be an important direction of our future work.

5. Overview of the Software System Using the Proposed Approaches

Our software system, primarily for Japanese students, provides three different approaches to the learning of Mandarin Chinese, including pronunciation practice, interactive dialogue, and video-aided real-world dialogue. Figure 6 is a screen shot of our system when the pronunciation practice is in action. First, the student can choose a sentence, then the system will show the reference utterance at the bottom. The student can listen to the reference utterance before recording. After finishing the recording, the score of the test utterance is shown at the right-upper corner. The user can click the button labeled "點數結果" to check detailed scores to the phone level, as shown in the popup menu in Figure 6.

Figure 6. The screen shot of the pronunciation practice in action.

To make a vivid visual feedback, low scores for phones/syllables are displayed in red, as shown in the popup dialog, where "這" and "剛" have scores in red. In fact, "這" and "剛" were pronounced as "那" and "不", respectively, in the test utterance. The system was able to detect the mispronunciation and gave both syllables poor scores of 27 and 52, respectively. Note that the target sentence is displayed with Chinese characters, their Hanyu Pinyin transcription, and the corresponding tone symbols, including "－" (tone 1), "／" (tone 2), "∨" (tone 3) and "＼" (tone 4). Hanyu Pinyin without any tone symbol represents "tone 5".

Figure 7 illustrates the screen shot of video-aided real-world dialogue. The scenario in this example is a conversation carried in a hotel checkout procedure. The video stops whenever the student needs to say a sentence in response to the counter clerk. If the scores of the student's utterance are higher than 80, then the video continues. Otherwise the student needs to practice the sentence until the score is higher than 80. The procedure is almost the

same as that of interactive dialogue, except that a video is played to imitate real-world conversations.

Figure 7. A screen shot of the video-aided real-world dialogue.

6. Conclusions

In this paper, we have developed the algorithms to construct a CAPT system for evaluating the pronunciation of Mandarin Chinese. The proposed system uses several techniques from speech signal processing and recognition, including the HMM-based forced alignment, a pitch determination using autocorrelation, and tone recognition using GMM. By using downhill Simplex search, we successfully derived a set of parameters for a scoring function that can approximate the scores from a human expert. Similar approaches can be applied to construct CAPT systems for other tonal languages, such as Taiwanese, Minnan, Cantonese, Tibetan, Punjabi, and so on.

References

Chen J.C. and J.S. R. Jang, "Extended Supratone Modeling for HMM-based Continuous Tone Recognition," *ACM Transaction on Speech and Language Processing*, 2007. [submitted]

Chen J.C. , and J.S. R. Jang, J.Y. Li and M.C. Wu, "Automatic Pronunciation Assessment for Mandarin Chinese," *IEEE International Conference on Multimedia & Expo*, 2004, pp. 1979-1982.

Chen S.H. and Y.R. Wang. “Tone Recognition of Continuous Mandarin Speech Based on Neural Networks,” *IEEE Transactions on Speech and Audio Processing*, 3(2), 1995, pp. 146-150.

Chen J.C., J.L. Lo, and J.S. R. Jang, “以語音辨識與評分輔助口說英文學習,” In *Proceedings of Conference on Computational Linguistics and Speech Processing (ROCLING)*, 2004, available at http://www.aclclp.org.tw/rocling/2004/M25.pdf

Huang S.C., ”Improvement and Error Analysis of Tone Recognition for Mandarin Chinese”, MD thesis, National Tsing Hua University, 2006.

Huang X., A. Acero, and H.W. Hon, Chapter 12 of “*Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*,” Prentice Hall PTR, Upper Saddle River, New Jersey, 2001, pp. 585-636.

Jang J.S. R., and S.S. Lin, "Optimization of Viterbi Beam Search in Speech Recognition," In *Proceedings of International Symposium on Chinese Spoken Language Processing*, 2002, paper 114.

Jang J.S. R., C.T. Sun and E. Mizutani, “*Neural-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*,” Prentice Hall PTR, Upper Saddle River, New Jersey, 1997.

Kim C. and W. Sung, "Implementation of An Intonational Quality Assessment System,” In *Proceedings of International Conference on Spoken Language Processing*, 2002, pp. 1857-1860.

Lee L.S., "Voice Dictation of Mandarin Chinese," *IEEE Signal Processing Magazine*, 14(4), 1997, pp. 63-101.

Li J.Y., “Speech Evaluation,” MD thesis, National Tsing Hua University, Taiwan, 2002.

Lin W.Y., and L.S. Lee, “Improved Tone Recognition for Fluent Mandarin Speech Based on New Inter-Syllabic Features and Robust Pitch Extraction,” In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, 2003, pp. 237-242.

Neri A., C. Cucchiarini, and W. Strik, “Automatic Speech Recognition for Second Language Learning: How and Why It Actually Works,” In *Proceeding of International Congresses of Phonetic Sciences*, 2003, pp. 1157-1160.

Neumeyer L., H. Franco, V. Digalakis, and M. Weintraub, “Automatic Scoring of Pronunciation Quality,” *Speech Communication*, 30(2-3), 2000, pp. 83-93.

Rabiner L. and B.H. Juang, “*Fundamentals of Speech Recognition*,” Prentice Hall PTR, Upper Saddle River, New Jersey, 1993

Sukkar R. A. and C.H. Lee, “Vocabulary Independent Discriminative Utterance Verification for Nonkeyword Rejection in Subword Based Speech Recognition,” *IEEE Transactions on Speech and Audio Processing*, 4(6), November 1996, pp. 420-429.

Tang Poetry Corpus, 2002 Recordings, available at http://mir.cs.nthu.edu.tw/research/corpus/tangPoetry