Automatic Pronunciation Assessment for Mandarin Proficiency Test Based on HMM

Chunting Yang
Zhejiang University of Science and Technology
Email: hzyangct@gmail.com

Yang Liu and Jing Yu
Zhejiang University of Science and Technology, University of Toronto
Email: hzliuyang@gmail.com, jyu@mie.utoronto.ca

***Abstract*—Objective pronunciation assessment plays a very important role in the Mandarin Proficiency Test. But it still has a long way to go before it reaches the level of success. In this paper, the novel Mandarin objective pronunciation assessment pronunciation of is proposed. The standard of Mandarin pronunciation is divided into six levels. The mandarin pronunciation is divided into consonant, vowel and tone. And the feature parameters are explored, and Bark-SD is selected as consonant assessment parameter. Hidden Markov Model (HMM) is used to assess the pronunciation. The objective assessments correlate well with the subjective assessment because perception distortions of human auditory have been taken into consideration. Experimental results show the method performs better than the exiting methods.**

***Index Terms*—mandarin proficiency test; hidden Markov model; objective pronunciation assessment**

I. INTRODUCTION

With the social and economic development, Mandarin, the standard language of the modern Chinese, is great popularized all over China. Now the mandarin test becomes the national test, and more than three million people take part in National Mandarin Proficiency Test (NMPT) each year in China. And many civil servants, teachers, students and other people start to take part in MPT. Up to now MPT is manual. In the process of test, two examiners test one examinee face to face at the same time. Actually the manual testing has many disadvantages, the following are the main [1][2].

It is inefficiency. When the test is in the way of manual work, many examiners are arranged to the examination rooms. There are two examiners in each room. The examinees that are waiting in the waiting room are call to enter the room in turn. At the same time two examiners only test one examinee and the process may last twenty minutes. Two examiners could test about twenty examinees one day, and the workload of examiners is very heavy. On the other hand, lack of mandarin examiners make the situation worse. These causes a lot of people cannot get the chance of the test.

The testing qualities also face the challenge. There are many dialects in China, and the level of examiner is irregular. That result in non-uniform scoring standard. Otherwise, the test is easy to effect by the subjective consciousness of examiners when the test is processed face to face. Sometime two examiners who are working together may discuss each other. In fact the evaluation should be carried through independently by the examiners.

Attention should be paid to the testing cost. The distribution of costs across the different activities in the test of mandarin is clear. The fee of organization, rent of examination room and wages are the main body. The total cost of testing will cut down at least 50%.

In order to remove the human factors improve efficiency and save the cost, NMPT based on computer-aided is expected urgently. If NMPT based on computer-aided become available, the corpus of NMPT can be obtained. With the data mining technology we can get the attributes of defects mandarin pronunciations. These attributes will greatly contribute to mandarin teaching.

General, there are two types of pronunciation assessment methods, subjective assessment and objective assessment. In subjective assessments the observers' judgment based on rules determines the grade. It reflects the observers' subjective impression on speech quality. Different subjective assessment methods test speech quality on the different focus. There are many methods of subjective assessment, such as mean opinion score (MOS), diagnostic rhyme test (DRT), degradation mean opinion score (DMos), Diagnostic acceptability measure (DAM). Which, MOS is a widely used method of subjective assessment. Its measure the speech quality with the average opinion score of observers. It expresses the speech quality with five levels that are excellent (5 mark), good (4 mark), average (3 mark), poor (2 mark), bad (1 mark). MOS should be a true representation of the speech quality because people are the ultimate recipients of speech.

The advantages of subjective assessment methods are simple and easy to understand. These also are true reflection of the actual speech quality. But in subjective assessment methods there are strict requirements on the assessment condition and process. In order to avoid the

Manuscript received August 1, 2009; revised September 1, 2009; accepted September 10, 2009

The research was sponsored by Zhejiang Province Natural Science Foundation of China under Grant No. X106870

doi:10.4304/jcp.5.7.1062-1069

perceived bias of individual assessment, subjective assessments should count the statistical results of many observers. So subjective assessment methods are laborious, time-consuming, high cost, poor flexibility, and poor reproducibility, these are also difficult to apply real-time occasions.

To make up for deficiencies in the subjective assessment methods, various objective assessment methods of speech quality have been proposed and applied [7][8]. Objective assessment methods automatically determine speech quality by computer.

According to feature parameters used in speech assessments, objective assessment methods of speech quality has generally experienced the development from the time-domain analysis to the frequency domain analysis, and from the frequency domain analysis to the perception analysis. Based on frequency domain analysis the perceptual domain analysis usually converted speech signal to internal acoustic feature that reflect the psychological characteristics. To some extent the transformation simulate psycho-acoustic characteristics of human and the speech processing in peripheral auditory system and Cochlea of human. The speech feature parameters generated by the processing are considered contains high-level perceptual processing of the nervous system. Now the main objective assessment methods are based on feature parameters of speech perceptual analysis.

The objective assessment methods can be divided into non-reference and reference methods. Reference methods assess the speech quality with the distortion between output speech signal and reference speech signal. On other hand non-reference methods only use output speech signal to assess the speech quality. Objective speech quality metric is a parameter in a kind of feature space. Linear prediction cepstrum coefficients distance (LPCCD), Mel frequency cepstrum distance (Mel-CD), bark spectral distortion (BSD) are typical metrics.

The basic idea of LPCCD is using a linear combination of the past several speech signals sampling to approximate a speech signal sampling. Then a group of predictive parameters are decided by the optimization of difference between speech signal sampling and liner predictive under some criterion. It is appropriate that LPCCD is used as speech distortion metric. But the characteristics of human ear are not considered when LPCCD is used to assess speech signal. Therefore the assessment of LPCCD and subjective assessment may be relatively large deviation.

To different frequencies of sound waves, human auditory sensitivity is different. Actually the relationship between human auditory and the acoustic frequencies is logarithmic. And human auditory perception has the masking effect. According to human auditory perception, the frequency of the acoustic signal is divided by non-uniform. Then some new metrics, such as Mel-CD, BSD, were proposed.[3][4]

Based on cepstral coefficients in the distortion measure and the non-linear frequency characteristics of human auditory perception, Mel-CD is proposed by R. Kublchek. First the frequency axis is transform into scale for Mel cepstral, and then transform into cepstral coefficients obtained in cepstral domain. Mel-CD use weighted sum of squares to define the distortion metrics. Mel-CD may more accurately reflect the human auditory perception on speech, so it has been widely used in the speech recognition and identification.

Bark Spetral Distortion is also one of objective assessments of speech based on human auditory perception.

BSD construct transformation model to simulate the human perception mechanism of the speech signal on the basis of human psychoacoustic characteristics.

BSD convert speech spectrums into auditory perception spectrum, in the 20 Hz ~ 16 kHz audible region, 24 Barker frequency groups which have different center frequencies are constructed.

BSD metric is defined as Barker spectrum Euclidean distance between the original signal and the speech signal. BSD metric simulate the human auditory characteristics with a wide range of hearing.

Mel-CD and BSD are two typical spectrum distortion assessment methods. When they are used to process the speech signal, parts of the human auditory characteristics have been taken into account. Because the frequencies of the speech signal are divided by non-uniform, the assessment results are closer to the subjective assessment.

II. SUBJECTIVE ASSESSMENT OF MANDARIN PRONOUNCING

The smallest unit of Mandarin pronunciation is syllables; all of more than 10,000 Chinese characters are corresponding to only about 1300 syllables. In accordance with the provisions of “Mandarin Proficiency Test level standards", the standard of Mandarin is divided into three levels, each level divided into two grades.

Phonetic error, phonetic defect, intonational deviance and phonetic systemic defects of consonant or vowel are major error in NMPT. Phonetic error means a syllable is misread for another syllable. When a consonant, vowel or tone is misread, the syllable is misread. There are many types of phonetic defect. One of defects is consonant defect. It refer to the oral part of pronunciation is not accurate enough, but not misreading for another consonant. Vowel defect refers to wrong-shaped mouth or open enough, obviously not listening to a sense of nature; Compound vowels defect are not allowed to place the tongue, not enough movement. Tone defects means the basic tone adjustment and tune trend are correct, but the tone value is obviously low or high, especially the relatively high or low of four tones are obviously inconsistent and so on. When a consonant, vowel or tone is defect, the syllable is defect. Generally certain types of defects are more than 10 times in NMPT, they can determine phonetic systemic defects. That means heavy dialect accent. Intonational deviance are deviance of sentence intonation,or stress.

NMPT consist of four parts that include reading 100 monosyllabic words, reading 100 multiple syllable words, reading a 400-syllable short essay, 3-minute speech.

According to the total score, examinee's Mandarin level is assessed.

III. THE PRINCIPLE OF MANDARIN PRONOUNCING ASSESSMENT

The smallest unit of Mandarin pronunciation is syllable. It is the basic unit of speech structure.Traditional phonology analyzed a syllable into consonant and vowels. The beginning of a syllable is called as consonant. The consonant letters are used to indicate the sound. There are 21 consonant in Mandarin. The pronunciation length of consonant is relatively stable. This feature is very useful to the objective assessment. A part of the back consonant in a syllable is called vowel. There are 39 vowels in Mandarin.

Mandarin is a tone language, tone is one of the important attributes of Mandarin, and plays an important role in the assessment of Mandarin. In addition to consonant, vowel, syllable also includes tone. Tone is an important component of the syllable. The syllable consists of the same consonant and vowel, if the tone is different the meanings of syllable is not same.

The assessment of Mandarin syllable pronunciation accuracy is contained in two aspects. From the assessment point of view, composition of a syllable phoneme can be assessed with the information of the acoustic channel. On the other hand the tone of a syllable is included in the information of acoustic source. Mandarin tone information is contained in the pitch curve of a syllable, and mainly in the segment of vowel [6].

Syllable information in these two aspects is independent of each other. They can be processed separately.

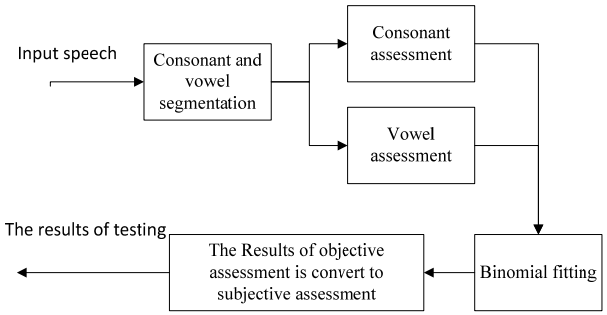

Figure 1. The process of Mandarin objective assessment

The speech assessment principle is shown in Fig.1. Firstly the input speech syllable is segmented into consonant and vowel so as to assess them separately. Because the tone information connote in the vowel, the tone assess with vowel. After consonant and vowel are assessed respectively, binomial fitting method is used to process the results. Then the objective assessment results should be obtained, and finally the results are converted to the testing levels.

Figure 2. Consonant assessment

Fig.2 shows the process of consonant assessment. In this paper the HMM was used to assess the consonant. And the tone assessment is shown as Fig.3. The tone information is contained in the vowel, so we can pick up the phoneme pitch curve from the tone information. Then we can assess the tone with this curve.

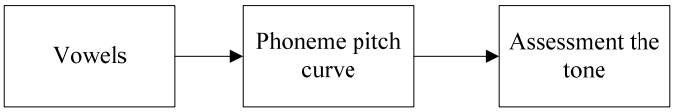

Figure 3. Vowels assessment

General the minimum frequency of the vowel pitch may be 60Hz. One level clipper, reducing the sampling rate and linear interpolation are used to detect the pitch curve.

The one level clipper function is

$$f(x) = \begin{cases} 0, & x(n) < C_1 \\ 1, & x(n) \geq C_1 \end{cases}$$

Autocorrelation function

$$f(n) \cdot f(n) = \begin{cases} 1, & f(n) = f(m) = 1 \\ 0, & f(n) = 0 \text{ or } f(m) = 0 \end{cases} \quad (1)$$

After a one level clipper processing, the peak of auto-correlation function is still very prominent, and the pitch frequency is more accurate, so the sampling rate could be reduced. The output sequence frequency after on level clipper processing can be divided by 3. One sampling data is collected in each three consecutive sampling data. Then the period of the signals can be preserved. The rule is expressed as

$$y(k) = \begin{cases} 1, & f(k) = 1 \text{ or } f(k+1) = 1 \text{ or } f(k+2) = 1 \\ 0, & f(k) = f(k+1) = f(k+2) = 0 \end{cases} \quad (2)$$

General, there are obvious errors scattered points on the pitch curve. Therefore the curve should be smoothed with FIR (Finite Impulse Response) filter. Fig.4 shows the smoothing pitch curve of the first, second, third, and fourth tones.

According to the feature of four tones shown as in Fig.4, the tone can be recognized. Pitch envelope is the key feature for detect the tones. The feature parameter of tones can be defined as:

$$x_t = [\log f_t + \log f_{t+1}, \log f_t - \log f_{t+1}]. \quad (3)$$

Where f_t is the pitch frequency at time t. So $\log f_t + \log f_{t+1}$ means the local amplitude of pitch frequency curve at time t, and $\log f_t - \log f_{t+1}$ means the local slope of pitch frequency curve at time t.

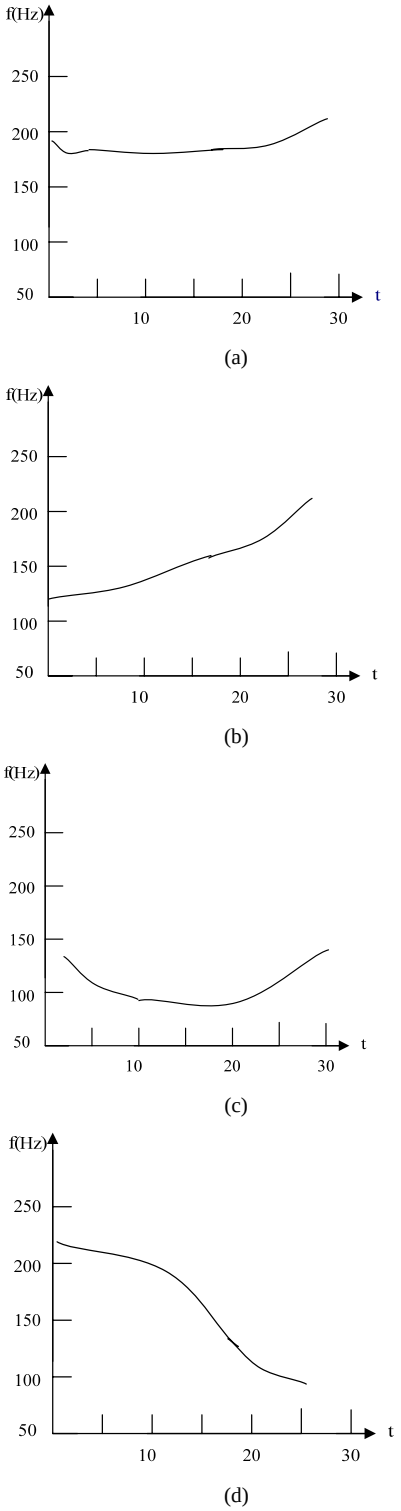

Figure 4. The smoothing pitch curve of the first, second, third, and fourth tones

IV. MANDARIN PRONOUNCING FEATURE EXTRACTION

Feature selection and extraction is the key step in the objective assessment of Mandarin. Feature extraction means extracting feature parameters which represent the nature of speech. Some commonly used feature parameters of speech are described as follows.

A. Linear Prediction Cepstrum Coefficients

Linear prediction cepstrum coefficients are defined as:

$$\mathrm{LPC-CD(l)}=\sqrt{[C_R(l,0)-C_O(l,0)]^2+2\sum_{k=1}^{p-1}[C_R(l,k)-C_O(l,k)]^2}. \quad (4)$$

$C_R(\cdot)$, $C_O(\cdot)$ are the Linear prediction cepstrum coefficients of reference speech signal and distortion speech signal respectively. P Is the order of predictor, set p=15, then for the distortion signal Including the L-frame, We can get the following equation.

$$\mathrm{LPC-CD}=\frac{1}{L}\sum_{l=0}^{L}\mathrm{LPC-CD(l)}. \quad (5)$$

Cepstrum coefficients can improve the stability of feature parameters, and it can separate the excitation signal generated by speech model and response signals of acoustic channel. Feature parameters based on acoustic channel are over-reliant on the accuracy of speech model. The speech signal stationarity assumed by model are not always true. The robustness of feature parameters based on channel model is not very good.

B. Mel Frequency Cepstrum Coefficients

Mel Frequency Cepstrum Coefficients is built on short-time frequency domain analysis of speech signal. Based on frequency domain analysis, Mel frequency cepstrum bends the frequencies of the speech signal to a new non-linear scale in the light of experiment of human auditory perception on the frequency and amplitude. And then the feature parameters are extracted in this scale.

$$C(i,k)=\frac{1}{N}\sum_{k=0}^{N-1}X(i,k)W^{ik}. \quad (6)$$

$$X(k,m)=T_w[\ln|s(k,m)|^2]$$

$$W=\exp\left(-j\frac{2\pi}{N}\right)$$

$$S(k,m)=\sum_{n=0}^{N-1}S(n,m)\exp\left(-j2\pi\frac{k}{N}\right)$$

N is the sampling points, T_w is the operator of Mel frequency scale bending, $S(k,m)$ is the Spectrum for speech signals.

Low-frequency part of the speech signal contains more speech information. And Mel frequency cepstrum coefficients are in particular to strengthen the low-frequency information. Meanwhile there is no any assumption on the premise for the speech signal. So the results of Mel frequency cepstrum coefficients should be good.

C. Bark Spectral Distance

Bark spectral distance distortion metric not only bend the auditory frequency, but also take into account the critical band integration, the effects of frequency on sensitivity of amplitude, the characteristics of subjective loudness.

Because many subjective characteristics of the human auditory system are considered, Bark spectral distance distortion metric should be able to better approximate the real characteristics of the human auditory system.

Through the following three steps, the power spectrum of speech signals are converted into the auditory spectrum decided by auditory perceptual characteristics of human ear system. Bark-SD assessment calculates the distance between the reference signals spectrum and the distortion signals auditory spectrum.

The power spectrum of speech signals is shown in the following equation.

$$P(\omega) = \mathrm{Re}[S(\omega)]^2 + \mathrm{Im}[S(\omega)]^2. \quad (7)$$

1) The critical band analysis:

The linear frequency ω is converted to the Bark frequency Ω by the following equation.

$$\Omega(\omega) = 6\ln\left[\frac{\omega}{1200\pi} + \left[\left(\frac{\omega}{1200\pi}\right)^2 + 1\right]^{1/2}\right]. \quad (8)$$

Hence the $P(\Omega)$ can be obtained. Then $P(\Omega)$ is convoluted with the auditory hidden critical band model$\psi(\Omega)$. We can get the equation.

$$\theta(\Omega_l) = \sum_{\Omega=-1.3}^{2.5} P(\Omega - \Omega_i)\psi(\Omega). \quad (9)$$

which

$$\psi(\Omega) = \begin{cases} 0, & \Omega < -1.3 \\ 10^{2.5(\Omega+0.5)}, & -1.3 \le \Omega \le -0.5 \\ 1, & -0.5 \le \Omega \le 0.5 \\ 10^{-1.0(\Omega-0.5)}, & 0.5 \le \Omega \le 2.5 \\ 0, & 2.5 < \Omega \end{cases}$$

The results of convolution are sampled in about 17 Bark intervals. Then the sampled sequence $\theta[\Omega(\omega)]$ is obtained.

2) Equal loudness curve processing

Acoustic psychology experiments show that the human feel the different loudness to different frequencies. The 40dB equal loudness curve of the human ear is simulated as following equation.

$$E(\omega) = [(\omega^2 + 56.8 \times 10^6)\omega^4]/[(\omega^2 + 6.3 \times 10^6)^2 \times (\omega^2 + 3.8 \times 10^9)]. \quad (10)$$

Then $\theta[\Omega(\omega)]$ is processed on the weighted, we can obtain the equation.

$$E[\Omega(\omega)] = E(\omega) \times \theta[\Omega(\omega)]. \quad (11)$$

3) Acoustic loudness and intensity

The relationship between the human feeling acoustic loudness and its intensity is not a linear, and it can be approximate with 3rd power. The human feeling acoustic loudness at frequency Ω is set as $B(\Omega)$, then

$$B(\Omega) = E(\Omega)^{0.33}. \quad (12)$$

Bark spectral distance of speech signals is defined by the Euclidean distance

$$\text{Bark} - \text{SD} = \sum_{i=1}^{N}[B_x(\Omega_i) - B_y(\Omega_i)]^2. \quad (13)$$

Where, N is the number of frequency band. And $B_x(\Omega_i)$ and $B_y(\Omega_i)$ are the acoustic loudness of reference signals and distortion signals respectively.

4) The experiments on LPC-CD, Mel-CD and Bark-SD

The selection of feature parameters has a great impact on the Mandarin objective assessment methods. In order to choose more effective feature parameters of Mandarin assessment, Experiments on LPC-CD, Mel-CD and BSD are carried out. The results compared with the MOS are shown as Fig.5, Fig.6, and Fig.7.

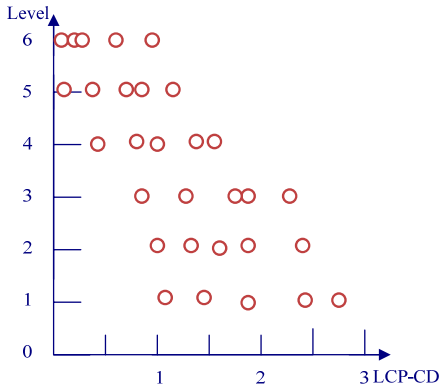

Figure 5. The relationship of LPC-CD and subject assessment

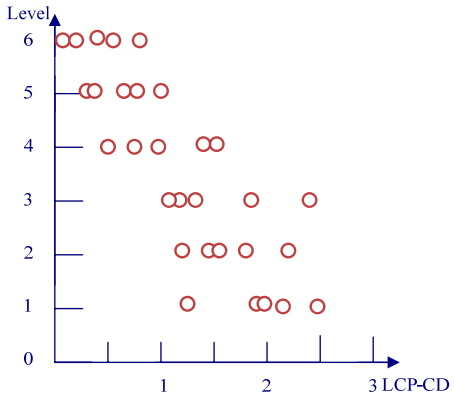

Figure 6. The relationship of Mel-CD and subject assessment

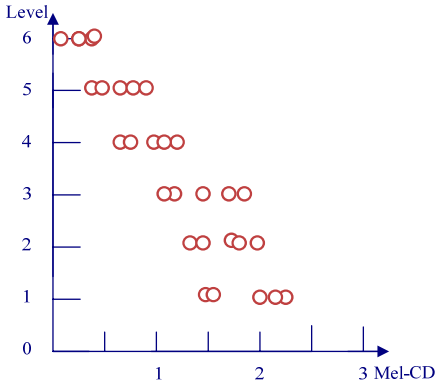

Figure 7. The relationship of Bark-SD and subject assessment

From Fig.5, Fig.6, and Fig.7, we can find that Mel-CD represent the Mandarin feature clearer than LPC-CD, and meanwhile Bark-SD represent the Mandarin feature clearer than Mel-CD. So in this paper the Bark-SD is selected as the feature parameter of Mandarin objective assessment. There are several advantages.

When LPC-CD and Mel-CD etc. are used as the metrics, the reference signals and distortion signal must be strictly aligned at the starting point. But the signals need not be synchronous for the Bark-CD.

The pronunciation length of a word is different for different people. So signals should be strictly normalized for LPC-CD and Mel-CD etc.. This is bound to increase the calculation and some of the signal feature would loss. Of Bark-CD is concerned, the situation is not same.

Bark-SD fully reflects the perception characteristics of the human auditory system.

V. OBJECTIVE ASSESSMENT OF MANDARIN PRONOUNCING BASED ON HMM

A. Hidden Markov Model (HMM)

A HMM is a statistical model in which the system being modeled is assumed to be a Markov process with unobserved state. Hidden Markov models are especially known for their application in temporal signals processing such as speech. In fact HMMs are widely used in speech signal processing.

Although the speech signals impacted by various factors show a strong uncertainty, however, the semantic information implied in the speech signal is determined. The changes of the speech signals are decided by the signal implied the semantics information.

Speech signals can be viewed as a piecewise stationary signal or a short-time stationary signal. That is, one could assume in a short-time in the range of 10 milliseconds, speech could be approximated as a stationary process. Speech could thus be thought of as a Markov model for many stochastic processes.

In speech signals processing, the HMM would output a sequence of n-dimensional real-valued vectors. The vectors would consist of cepstral coefficients, which are obtained by some kind of transform.

The HMM will tend to have in each state a statistical distribution that is a mixture of diagonal covariance Gaussians which will give a likelihood for each observed vector. Each word, or, each phoneme, will have a different output distribution; a HMM for a sequence of words or phonemes is made by concatenating the individual trained HMMs for the separate words and phonemes. HMM can descript the overall non-stationary and local stability of speech signals, is an ideal model of the speech signal processing. Moreover HMMs can be trained automatically and are simple and computationally feasible to use. So HMMS are popular.

When HMM are used to Mandarin speech pronunciation assessment system, the key issue is to select important modeling parameters.[5]

1) Feature parameters

Bark-CD as the speech feature parameters has many advantages. The experiments of Section II have proved that Bark-CD is better than other parameters. Therefore in this paper the Bark-CD is selected as the feature parameter.

2) Distrubutional parameters

HMMs can model complex markov processes where the states emit the observations according to some probability distribution. One such example of distribution is Gaussian distribution, in such a HMM the states output is represented by a Gaussian distribution.

Moreover it could represent even more complex behavior when the output of the states is represented as mixture of two or more Gaussians, in which case the probability of generating an observation is the product of the probability of first selecting one of the Gaussians and the probability of generating that observation from that Gaussian.

Let A denote the State transition probability matrix, then

$$A = \{a_{ij}\}_{N\times N},$$

$$a_{ij} = P(q_{t+1} = S_j | q_t = S_i), 1 \le i, i \le N$$

A is a hidden Markov chain, the mathematical probability that state S_i transit to state S_j is only relevant with the state S_i.

B is observation probability matrix,

$$B = (b_{jk})_{N\times L}$$

Where $b_{jk} = P(O_k | q_t = S_j), 1 \le j \le N, 1 \le k \le L$

L is the possible observation number of each state. For Random vector $O_n = [O_{n1}, O_{n2}, \cdots, O_{nT},]^T$, Each observation vector has its corresponding output probability with different state.

Let π denote initial state distribution,

$$\pi = \{\pi_i\},$$

$$\pi_i = P(q_1 = S_i), 1 \le i \le N, \text{ and } \sum_i \pi = 1$$

3) HMM unit

The hidden Markov process can be modeled at any scale, such as phoneme, word and sentence. A Mandarin syllable consists of consonant, vowels and tone. The assessment unit in NMPT is consonant, or vowels or tone. Therefore the HMM at phoneme scale can represent the main feature of Mandarin pronunciation.

4) The number of state

In speech signals processing, English phonemes are generally used 3 states. But Mandarin assessment needs more states. Different from the speech recognition, Mandarin assessment is not only to recognize words, but also to assess how accurate the pronunciation is. In order to achieve a certain assessment precision, more states are needed. Therefore the number of phonemes states is defined as 5.

Phoneme states: $\{S_1, S_2, S_3, S_4, S_5\}$,

At any time, phoneme states can only be one of the $\{S_1, S_2, S_3, S_4, S_5\}$.

5) The number of mixture Gaussians

The speech parameters are changed with different people, so generally the number of mixture Gaussian is more. The size of mixed Gaussian Mandarin has a significant impact on the mandarin assessment accuracy.

If training sample is sufficient, the greater the number of mixture Gaussian, the more detailed description of the

state speech space, and the state speech space will be closer to physical speech space. However, if the number of mixture Gaussian is bigger, the speech number designated to the state should be sufficient in a state. Moreover the data must be evenly distributed. In this paper the number of mixture Gaussian is set to 3.

HMM include two parts. One is Markov chain that is described by the A and π, and it can emit a sequence of state outputs. Another part is random process, it can output observable sequence. Therefore the HMM can denote $\lambda = (\pi, A, B)$.

VI. EXPERIMENTAL RESULTS

In this paper all of the training samples and test samples are from the NMPT database of Zhejiang Province Mandarin Professional Testing Center. The Pronunciation of 4 Chinese characters for 6 levels are selected, there are 20 samples each pronunciation.

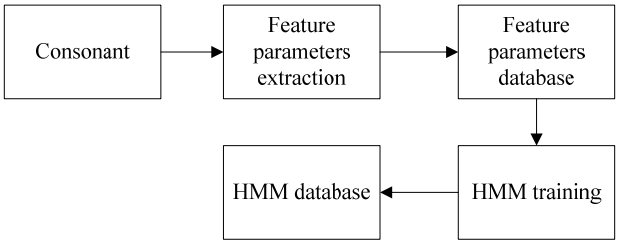

Figure 8. The process of HMM training

Generally the correlativity of the subjective and objective assessment results can be represented by the Pearson coefficient. The correlative coefficient ρ and deviation δ can be calculated using the following equations.

$$\rho = \frac{\sum_{i=1}^{M}(S_O(i)-\overline{S_O})(S_S(i)-\overline{S_S})}{\sqrt{\sum_{i=1}^{M}(S_O(i)-\overline{S_O})^2 \sum_{i=1}^{M}(S_S(i)-\overline{S_S})^2}} \quad (14)$$

$$\delta = \sqrt{\frac{\sum_{i=1}^{M}(S_O(i)-S_S(i))^2}{M}} \quad (15)$$

Where M is the number of distortion speech, and $S_O(i)$ and $S_{OS}(i)$ are the results of subjective assessment and objective assessment of the i^{th} distortion speech respectively. $\overline{S_O(i)}$ and $\overline{S_S(i)}$ are the average results of subjective assessment and objective assessment of the i^{th} distortion speech respectively.

VII. CONCLUSION

The above theory has been applied to the information management system of mandarin test (IMSMT). Zhejiang province mandarin test center is the first user of the IMSMT. The Hangzhou city mandarin test center, Yiwu city mandarin test center and Hangzhou normal university are the new users. They all give a high appraisement. About 7140 examinees finish their mandarin test with the IMSMT. The system has passed Chinese Ministry of Education appraisal in 2009.

TABLE I.
THE ASSESSMENT RESULTS OF HMM

Test Level	Recognition Rates (%)			
	Ba(1)	Lu(4)	Hen(3)	Hang(2)
Level I(A)	91	90	89	89
Level I(B)	91	90	88	87
Level II(A)	90	89	88	86
Level II(B)	88	87	86	86
Level III(A)	85	85	82	83
Level III(B)	83	82	78	76

TABLE II.
THE CORRELATIVE COEFFICIENT

Test Level	ρ			
	Ba(1)	Lu(4)	Hen(3)	Hang(2)
Level I(A)	0.9768	0.9836	0.9728	0.9647
Level I(B)	0.9736	0.9682	0.9686	0.9652
Level II(A)	0.9562	0.9624	0.9472	0.9531
Level II(B)	0.9624	0.9587	0.9513	0.9584
Level III(A)	0.9453	0.9534	0.9427	0.9426
Level III(B)	0.9562	0.9471	0.9468	0.9402

TABLE III.
THE DEVIATION

Test Level	ρ			
	Ba(1)	Lu(4)	Hen(3)	Hang(2)
Level I(A)	0.9768	0.9836	0.9728	0.9647
Level I(B)	0.9736	0.9682	0.9686	0.9652
Level II(A)	0.9562	0.9624	0.9472	0.9531
Level II(B)	0.9624	0.9587	0.9513	0.9584
Level III(A)	0.9453	0.9534	0.9427	0.9426
Level III(B)	0.9562	0.9471	0.9468	0.9402

ACKNOWLEDGMENT

The authors wish to thank Zhejiang province mandarin test center and Hangzhou mandarin test center support for the project. This work was supported in part by a grant from Zhejiang Province Natural Science Foundation of China under Grant No. X106870.

REFERENCES

[1] Chunting Yang, Yang.liu, and Zhigang Cheng, "A Framework for the Information Management System of Mandarin Test", *The Proceeding of IEEE IUCE 2009.* Chengdu, pp. 164–167, May 2009.

[2] Yang Liu, Chunting Yang, and Weifeng Ma, "Automatic Pronunciation Scoring for Mandarin Proficiency Test Based on Speech Recognition", *The Proceeding of IEEE IUCE 2009*. Chengdu, pp. 168–172, May 2009.

[3] Kubichek R. "Melcepstral distance measure for objective speech quality assessment", *IEEE Pacific Rim Conference on Communications , Computers , and Signal Processing* , pp. 125–128, 1993.

[4] Wang S , Sekey A , Gersho A. "An objevtive measure for predicting subjective quality of speech coders", *IEEE Journal on Selected Areas in Communications* , Vol.10 (5), pp. 819 – 829,1992.

[5] Lee Yu- min, Lee Lin- shan, "Continuous hidden Markov models integrating transitional and instantaneous features for mandarin syllable recognition", *Computer Speech and Language*, Vol.7, pp. 247- 263,1993.

[6] Ying Y, Xu S，"A fast method of pitch detection for Chinese four tones recognition"，*Proceeding of ISCP'93* , Beijing, Oct 1993.

[7] H Franco, L Neumeyer, Y Kim, O Ronen ,"Automatic pronunciation scoring for language instruction," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997 vol.2,pp.1471-1474,1997

[8] L Neumeyer, H Franco, V Digalakis, M Weintraub, "Automatic scoring of pronunciation quality,"*Speech Communication*, Volume 30, Issues 2-3, pp. 83-93, February 2000

Chunting Yang was born in Qiqihar City, P. R. China, on January 16, 1964. Yang received a BS in 1986 from Nanjing University of Aeronautics and Astronautics. Then he received his MS in 1991 and PhD in 1996 from the Southeast University and Zhejiang University. He is an associate professor of computer science and engineering at the Zhejiang University of Science and Technology. His research interests are in computer image processing, computer vision and speech signal processing.

Dr. Yang is the member of a council of Zhejiang Computer Society and the member of a council of Zhejiang Electronics Society.

Yang Liu was born in Wuhan, P. R. China, on June 4, 1978. Liu received her BS in 2000 and MS in 2003 from Huazhong Normal University.

She is a full-time lecturer of computer science and engineering at the Zhejiang University of Science and Technology. Her research interests are in computer vision, virtual reality.

Jing Yu received her PhD in Mechanical Engineering in 1998 from Zhejiang University (Hangzhou, China) and worked as an Associate Professor in the Institute of Vibration Engineering Research at Nanjing University of Aeronautics and Astronautics (Nanjing, China) for the period of May, 1998 to March, 2000. Between October 2000 and April 2002, she worked as a Research Associate in Lakehead University, Canada. At present, she is engaged in research in University of Toronto, Canada. Her areas of research interest include random vibration control and signal processing.. .