Journal of Theoretical and Applied Information Technology
20th January 2013. Vol. 47 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

OBJECTIVE EVALUATION SYSTEM OF ENGLISH PHONEMES

[1, 2]JING ZHANG, [1]SIMIN YU

[1] Faculty of Automation, Guangdong University of Technology, Guangzhou 510006, China
[2] Cisco School of Information, Guangdong University of Foreign Studies, Guangzhou 510006, China
E-mail: Ha_go@163.com

ABSTRACT

Nowadays it is quite difficult for the traditional teaching methods and resources to meet the widespread and urgent needs of spoken English learning. Computer-assisted Language Learning (CALL) technology can greatly improve the efficiency of language self-learning, providing timely, accurate and objective assessment and feedback which can help learners to find the difference between their own pronunciations and the standard pronunciations, and then correct their own pronunciations. This paper presents an objective evaluation system about English phoneme pronunciation. This system can give an objective evaluation and feedback about student's pronunciation by analyzing the waveform and calculating the Mel Mel-Frequency Cepstrum Coefficients (MFCC) Euclidean distance between student's pronunciation and standard pronunciation using the Dynamic Time Warping (DTW) algorithm, to help students learn English phonemes with no teacher involved.

Keywords: *Computer-assisted Language Learning (CALL), Pronunciation Evaluation, MFCC, DTW*

1. INTRODUCTION

Speech signal processing technology in language learning is an important element for the integration of information technology and language learning, and its goal is to combine the latest speech technology with teaching and learning methods, and establish computer-assisted language learning system. The application of speech processing technology in language learning has become an important research content in the speech signal processing field [1]. The research on computer-assisted language learning system began in the 1950s. Initially, the application of computers in language learning was confined to the university campus research [2]. The birth of the world's first generation of CALL system was used in the learning of Russian at the Stanford University, Dartmouth University, and the Essex University [3]. Later, the learning of other languages joined. The Programmed Logic / Learning for Automated Teaching Operation jointly developed in 1959 by the Illinois University and its business partners, Control Data (PLATO) is considered to be the world's first large-scale computer assisted language learning system, which greatly contributed to the application of the computer in foreign language learning [4]. PLATO integrated the best advantages of the CALL system at that time, and owned the programming language designed specifically for users to facilitate the teaching of foreign languages as well. Other CALL systems in the world include: Watch Me Read[5], ISTRA[6] , LISTEN[7] , etc.

Currently domestic research in the CALL system based on speech technology is still in its infancy, but it has attracted extensive attention from researchers. Electronic Engineering Department of Tsinghua University and Institute of Linguistics of Chinese Academy of Social Sciences, and Institute of Acoustics of Chinese Academy of Sciences, and Institute of Automation as well as Microsoft Research Asia and other research institutions are conducting related work. The in-depth research of speech technology involved in the CALL system will help to improve the current situation of English teaching, and change the mode of teaching, and completely solve the existing problems in the oral teaching, and train qualified personnel quality, the subject has great theoretical value and practical significance.

For the spoken English teaching, teachers should try not only to find ways to help students distinguish English phonemes, English syllable structure, and understand the rhythm of English, English intonation, etc., but also to help students

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

master the complete English speech system. Facing the extensive and urgent needs for spoken English teaching, the current teachers and traditional teaching methods are difficult to meet. While the progress of Speech signal processing and recognition technology, to promote the development and application of Computer-assisted Language Learning techniques (CALL), which can greatly improve the efficiency of language learning. The timely, accurate and objective evaluation and feedback can help learners to find the gap between their pronunciation and the standard pronunciation, and then the pronunciation errors could be corrected in time. And for the computer, the rich function of graphics and speech owned by which promoted the people's language learning powerfully.

Firstly the paper introduced the theory of the English pronunciation and phoneme system, and on the basis of which an objective evaluation system of phonemes pronunciation based waveform analysis and MFCC parameters was proposed, and the specific structure and achievement of the system was described as well. Finally, the system's overall performance was verified by the experimental data.

2. THE THEORETICAL RESEARCH OF ENGLISH PHONEMES PRONOUNCE

2.1 English Phoneme System

English is widely used around the world, its pronunciation also shows large differences due to geographical differences. British English and American English impact the second language teaching most. Both English show some differences in pronunciation, but they are generally consistent, that they have more commons than differences [8]. The English IPA totally has 48 phonemes, of which 20 are vowel phonemes and 28 consonant phonemes. It should be noted in the new textbooks neither British nor American pronunciation include the two pairs of consonants [tr], [dr] and [ts], [dz].

2.2. English Pronunciation Features

English pronunciation characteristics are determined by its phoneme, according to the pronunciation phonetics, vowels is the formed phonemes that the air stream vibrating vocal cords in the oral cavity without any obstacle, different vowels are caused by different oral cavity shapes. Consonant is formed due to the airflow is hindered in the oral cavity, and the different consonant is caused by the different pronunciation place or methods.

The vowel is the core of constituting syllable, and it may both form syllable separately (e.g. indefinite article a), and constitute syllables with the consonants before and after it together (e.g., CAT). English vowel is pronounced by the activities of mouth, tongue, palate (including the soft palate and the hard palate), which constitute the airflow channel and resonate. The air flow has not any hindrance when the Vowel is pronounced, and the vocal cords vibrate at the same time, so the pronunciation is loud, and has longer duration, which is one of the biggest differences in the pronunciation from English consonants [9]. The vowels are divided into two categories: the single vowels and diphthongs, according to movement parts of tongue when pronunciation the vowels can be divided into three categories: front vowels, mid vowels and back vowels.

The air flow has not any hindrance when the Vowel is pronounced, by contrast, consonants is formed due to some obstruction encountered by airflow when it passed through the mouth or nasal, and the airflow broke through the hinder. The tips of consonant pronunciation is to control airflow site where the obstacles that airflow to break through. It is different from the lip we emphasized when we were talking about the vowel [9]. English consonants can be divided into voiceless and voiced consonants according to whether the vocal cords vibrates or not when pronounce, not vibrate as voiceless and vibrate, voiced consonants; according to the different part to control air flow the consonants can be divided into bilabial, labiodentals, tongue and teeth sound, tongue gingival tone , palate sound, soft palate tones and glottal sound; according to the pronunciation methods consonant can be divided into plosive, fricative, affricate, nasal, lingual tone and semi-vowels.

3. THE EVALUATION PRINCIPLE OF PHONEME

3.1 The Selection of Evaluation Target

The differences between English vowel are usually reflected in the length (for example: [i:] and [i], [u] and [u:], etc.) and on the lip, the difference of length is obvious, while that of the tones is difficult to be distinguished. While the difference between consonant is reflected in whether the vocal cords vibrate or not (for example: [p] and [b], [t]and [d], etc.) and the difference of pronunciation position and methods. The correct pronunciation could not be mentioned without correct distinction to the difference of pronunciation, so to achieve the objective evaluation for the phoneme, the right

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

characteristics used to describe the key points and distinction of pronunciation must be found out.

To take the speech waveform as one of evaluation standards can accurately reflect the length difference between vowel and the difference between the consonant vibrations. Mel cepstrum (Mel-Frequency Cepstrum Coefficients, MFCC) is proposed based on the characteristics of the human auditory system, which simulated the perception of human ear to different speech frequencies. The MFCC transformed the speech from the time domain to the cepstral domain, and then to the Mel scale, the transformations were mostly used to speech recognition systems, and achieved better results. As one of the features the MFCC could well reflect the differences between vowels and the differences of pronunciation methods and position between consonants.

Since the length and plumpness of phonemes would change in the specific context , during the pronunciation training of phonemes , there was no unified length standards, so, for the objective evaluation of phoneme pronunciation, the MFCC coefficients was taken as the major score points, while the waveform analysis was taken as a secondary score points.

3.2 System Algorithm

3.2.1 Pre-processing

Before the speech signal is analyzed and processed, the pre-emphasis, and sub-frame, as well as windowing and other preprocessing operations must be done. The purpose of these operations is to eliminate the impact of aliasing, high harmonic distortion, high-frequency and other defect caused by human vocal organ or acquisition equipment. The more uniform and smooth signal should be ensured after subsequent speech processing as much as possible to improve the quality of speech processing.

(1)Pre-emphasis

The average power spectrum of speech signal was affected by glottal excitation and muzzle radiation; it decayed as 6dB/oct (octave) in high frequency about more than 800Hz, for the higher the frequency, the smaller corresponding effect, so the high frequency of speech must be raised before the signal was analyzed. The digital filter is generally measures for the pre-emphasis. The relationship between output of pre-emphasis network s(n) and the input speech signal s(n) can be described as differential equation s(n)= s(n)-α*s(n-1). Where, α is coefficient for the pre-emphasis and usually taken as $0.9 \le a \le 1$, in this system a= 0.9375.

(2) Sub-frame processing

The technology throughout the whole process of speech analysis is “short-time analysis”. Speech signals with the of characteristics time-varying, but within a short period of time, its basic characteristics remain changeless that relatively stable. This kind of characteristics of the speech signal is called “instantaneous ", the short period of time is generally 10~30ms. Therefore, the analysis and processing of speech signal generally based on the basis of “instantaneous”, namely, "short-term analysis ", the characteristic parameters of speech signal was analyzed by sub-segment, wherein each segment is called a "frame", and the frame length is generally taken to be 10 ~ 30ms. Thus, for the whole speech signal, the result of analysis is time sequence of characteristic parameters composed by the characteristic parameters of each frame. The system mentioned in this paper uses the overlapping sub-frame, and each frame owns 256 samples.

(3)Windowing function

Speech signal owns the characteristic of instantaneous stability, so the signal can be sub-frame processed. In order that the speech waveform near sampling n could be strengthened and the rest be weaken, windowing process should be followed. The processing for each short segment of the speech signal is actually to apply a certain transformation or operation to the short segment respectively. The three most commonly used window function is the rectangular window, Hamming window and Hanning window, this system uses Hamming window, which is defined as Eq. (1).

$$w(n) = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi n}{N-1}), 0 \le n \le N-1 \\ 0, others \end{cases} \quad (1)$$

(4) Endpoint detection

Endpoint detection in speech signal processing is primarily in order to automatically detect the starting point and end point of speech. The paper proposed for endpoint detection. Dual- threshold comparison characterized with short-term energy E and short-term average zero-crossing rate, and combined with the advantages of Z and E, so the detection was more accurate, and the processing time of system was effectively reduced and the performance of real-time processing improved, in

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

addition the noise interference from silent segment can be excluded, thereby for speech signal the processing performance was improved.

Figure.1 shows the effect of endpoint detection for the standard speech [v], where the speech segment between the two red lines is effective speech segment.

3.2.2 Waveform analysis algorithm

After the pre-procession of endpoint detection got effective speech segment x [n], of which, the effective speech time that the speech length ST can be extracted as evaluation criteria, the calculation was described as Eq. (2).

$$ST = \frac{n}{f} \tag{2}$$

Where n is the number of the sampling points of effective speech segment x [n], and f is the sampling rate, for the standard wav format files f could be achieved by directly read the file header and for the speech directly sampled by the microphone, it can be defined by the user, generally, f could be 11025Hz (11kHz), or 22050Hz (22kHz) or 44100Hz (44kHz) , the system described in paper used 11025Hz (11kHz) sampling rate.

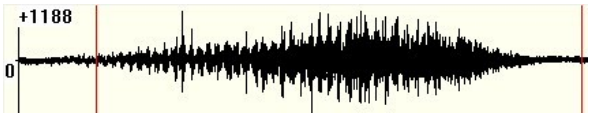

Figure.1 the Effect of Endpoint Detection for the Standard Speech

3.2.3 MFCC parameter extraction algorithm

Mel cepstrum is proposed based on the characteristics of the human auditory system, which simulated the perception of human ear to different speech frequencies. The process for human ear to distinguish frequencies of sounds likes a kind of logarithmic operation. For example: In the Mel frequency domain, the perception to sound is linear, that if the Mel frequency difference of two segments were two times, then difference of perception is also double. Due to space limitations, for more details of MFCC algorithm process see references (3).

3.2.4 DTW algorithm of similarity comparison

The MFCC parameters obtained after features extraction were in frames, it as the feature vector of reference templates, it was denoted by R(1), R(2), ..., R(m). Suppose the MFCC feature vector sequence of input speech was T (1), T (2), ..., T (n). In order to compare the similarity between them, the paper used dynamic time warping algorithm to compute their Euclidean distance dist[T, R], the smaller the distance the higher the similarity. So, dynamic time warping is to find time warping function, $m = w(n)$ which non-linearly mapped the timeline N of input template to the timeline M of reference template and the w meet Eq. (3).

$$dist = \min_{w(n)} \sum_{n=1}^{N} d[n, w(n)] \tag{3}$$

Where, $d[n, w(n)]$ was the distance between the n-th frame of input vector and the m-th frame of reference vector, the dist is the distance measurement between the two templates corresponding to the optimal time warping. From the above analysis it can be known that DTW is a typical optimization problem, which is to solve a corresponding warping function to the minimum total distance when matching the two templates.

According to equation (3), there are so many common points between the reference template and test the template, the computation of searching path of DTW algorithm is still quite large, in order to simplify the calculation, for the specific problems in the DP, according to the characteristics of speech, certain constraints must be defined:

The Boundary conditions as equation (4)

$$w(1) = 1, w(N) = M \tag{4}$$

And the continuous conditions as Eq. (5)

$$\mathrm{w(n+1) - w(n)} = \begin{cases} 0,1,2 \\ 1,2 \end{cases} \tag{5}$$

According to these two constraints, the function w(n) curve is limited to a parallelogram with one side lope is 2, and the other 1/2. In an extreme cases with n increased by 1 and m increased by 2, the end coordinates M= 2N; Otherwise, n increased by 2 and m increased by at least 1, then M=N/2. From the physical significance, it is equal to limit the length difference of two the templates to 1/2 to 2. In this way, since the path through the grid (n_{i-1}, m_{i-1}), then the next gird (n_i, m_i) to be through would be only one of three cases described in Eq. (6)

$$\begin{aligned} (n_i, m_i) &= (n_{i-1} + 1, m_{i-1} + 2) \\ (n_i, m_i) &= (n_{i-1} + 1, m_{i-1} + 1) \\ (n_i, m_i) &= (n_{i-1} + 1, m_{i-1}) \end{aligned} \tag{6}$$

Journal of Theoretical and Applied Information Technology
20th January 2013. Vol. 47 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Thus, by this calculation, the best match path D could be gotten. According to this path, using Euclidean method, as Eq. (7)

$$D(x,y)=\|x-y\|=\sqrt{\sum_{i=1}^{n}|x_i-y_i|^2} \quad (7)$$

Then the Euclidean distance dist [T, R] between the reference template and input template could be achieved.

3.2.5The scoring algorithms

As the speech length of the standard speech template and that of test speech, ST1 and ST2, achieved from the formula (1), could not be taken as the pronunciation score directly, but they could be mapped to the range of scores from 0 to 100 through the conversion by Eq.(8).

$$score1=100-\frac{|ST1-ST2|}{ST1} \quad (8)$$

The dist obtained by the DTW algorithm could not be taken as the pronunciation score directly, a reasonable map from distance to score must be found out. To assume the relationship between the distance and score meet with as Eq. (9).

$$score2=\frac{100}{1+a(dist)^b} \quad (9)$$

Obviously, the distance could be mapped to a range from 0 to 100 score with equation (9). To solve the unknown parameters a, b in the formula, some points and distance couples need to be known, which could be solved by some of the experts' score value and DTW distance obtained by experiments. In the System mentioned here, two groups of experts' score data and system score data were selected and calculated a = 0.0003, b = 2.

The final scores could be obtained through weighting and summing the scores analyzed by the two characteristic parameters according to Eq. (10).

$$Score=w1*score1+w2*score2 \quad (10)$$

Where w1 and w2 are the weighting factor, and w1 + w2 = 1, is the experience, which could be adjusted according to experimental data and the actual situation, and according to experts, the system took w1 = 0.1, w2 = 0.9.

4. SYSTEM IMPLEMENTATION

4.1System Architecture

Using C++ as development languages,then the underlying device such as the sound card could be directly manipulated through calling the Windows API, and the code could be more efficient, which is suitable for a large number of data processing operations. Taking VC6.0 as a development tool and MFC as the whole development framework made development be of high efficiency and easy to debug. The system mainly includes speech collection, and speech playback, and speech signal pre-processing, scoring and pronunciation feedback and other functional modules. The specific structures and processes are shown in Figure.2.

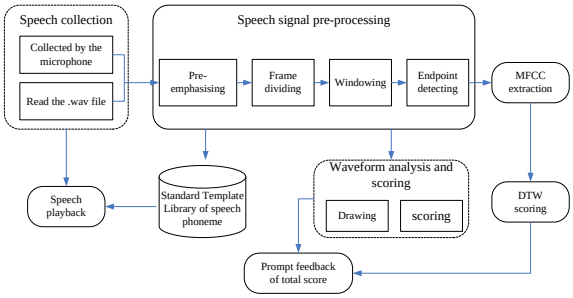

Figure 2. System Framework

The running effect of the system is shown in Figure.3.

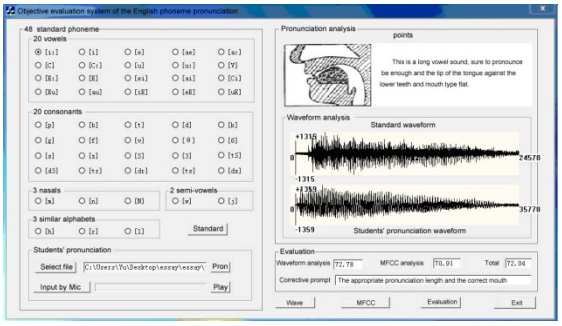

Figure 3. Running Effect Of System

4.2 The Implementation of Each Module in the System

(1)Speech collection module

The speech signal collection can be divided into two ways: collected by the microphone and directly read the .wav files, as shown in Figure.4.

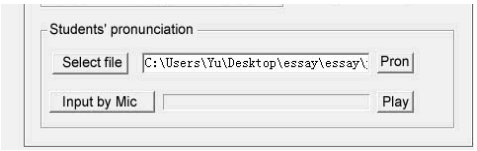

Figure 4. Speech Collection Modules

Where, the microphone collection module was achieved by calling WavIn API, it involved the multi-threaded programming and combined with the endpoint detection which made the user' experience more favorable. The specific processes ,as shown in Figure.5.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

(2)The speech playback module

The speech playback module directly called Windows API Play Sound () and then the .wav files can be played, and users can identify the pronunciation difference between the standard and their own by listening to the file so to practice and correct their pronunciation.

(3)Scoring and pronunciation feedback module

The scoring module using the algorithm described above. The pronunciation feedback is presented by using pronunciation lip-chart and text-tips, as shown in Figure.6.

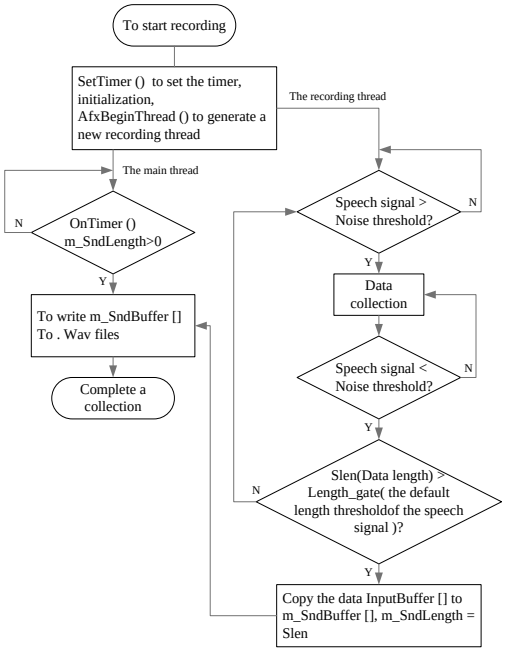

Figure.5. Flow Chart Of Microphone Collection Module

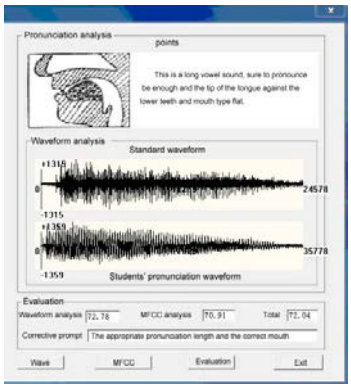

Figure.6. Effects Of Pronunciation Feedback Tips

5. SYSTEM IMPLEMENTATION

For A computer- assisted pronunciation learning system, the performance can be evaluated with four indicators as follows:

(1) Availability: In real auxiliary learning process, whether the system is easy to use and achieve the desired learning target or not. For example, the length of learning times, whether the learning content is rich and the guidance is appropriate or not.

(2) Effectiveness: Compared with the traditional learning methods, whether the system can eventually help learners to improve their pronunciation and significantly improve their level of the target language using or not.

(3) Accuracy: It mainly refers to reliability about the pronunciation scoring and classification, and to determine the error location and the type, and error correction and other aspects as well, the judgments made by the system must be ensured accurate.

(4) Authority: It mainly refers to the information system feedback to the learner's to be absolutely correct instead of any misleading of pronunciation to the learner. It mainly depends on the authority of expert pronunciation the system used, and the correctness of the algorithm that applied this knowledge to judge.

Table.1. Some Experimental Data

Phonemes / ***students***		[i:]	[e]	[ɔ:]	[a:]	[p]	[p]
student A	System score	76.10	68.55	74.25	74.31	67.30	87.71
	Teacher score	80	85	90	90	85	90
student B	System score	70.58	84.32	82.01	69.37	73.46	78.87
	Teacher score	80	90	70	80	90	90
student C	System score	82.64	83.42	82.16	77.63	73.40	69.40
	Teacher score	90	80	85	80	90	85
student D	System score	78.33	73.41	82.82	68.69	71.21	81.85
	Teacher score	90	85	83	70	90	90

When the performance of the system discussed here was analyzed, pronunciations of 48 phonemes were recorded by the English authority teachers as the standard pronunciation library, and 48 pronunciations of 10 students were recorded as test speech. Here Selected the pronunciation of vowel [i:] [e] [ɔ:] [a:] and a pair of consonants [p] [p] of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

four students as part of the experimental data, as shown in Table 1.

It can be concluded through the analysis of system performance, that the system has a friendly user interface and pronounced lip chart, as well as pronunciation correction tips and other functional modules. The system has good usability and effectiveness. However, the accuracy and objectivity in the evaluation still needs to be verified and improved.

6. CONCLUSIONS

The application and development of Computer-assisted Language Learning (CALL) is of great significance to improve English teaching efficiency, the research and design of system meeting the requirements of background, the conclusion and results obtained is of a certain theoretical significance and application value. While the system only for pronunciation of English phonemes evaluation, which has serious limitations, for the post-expansion, the extent can be extended to the evaluation of English word pronunciation, sentence or even a whole article pronunciation.

ACKNOWLEDGEMENTS

This work is partially supported by the ministry of education of humanities and social science project #10YJCZH220 and the school's Youth Project of Teaching Research #GWJYQN11016.

REFERENCES:

[1]Tamara Piankova. "Manual of English pronunciation", BeiJing Language and Culture University press, 2009, pp.320-330

[2] Doh-Suk Kim.,Tarraf An. "Enhanced Perceptual Model for Non-Intrusive Speech Quality Assessment", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP `06), Vol.5, 2006, pp.829-832.

[3] Falk T.H, Wai-Yip Chan. "on-intrusive speech quality estimation using Gaussian mixture models", Signal Processing Lett,Vol.13,2006, pp.108-111.

[4]Williams,S. M.,Nix,D. "Using Speech Recognition Technology to Enhance Literacy Instruction for Emerging Readers", In B. Fishman&S.0'Connor-Divelbiss (Eds.), Fourth International Conference of the Learning Seines, 2000,pp.115-120

[5] Rix A.W. "Perceptual speech quality assessment a review", IEEE International Conference on Acoustics, Speech and Signal Processing, Vol.3,2004, pp.1056-1059

[6] Osaka K."The technique of emotion recognition based on electroencephalogram", Information-An International Interdisciplinary Journal , Vol.11,2008,pp.55-68

[7] Zhang Jie, Huang Zhitong, Wang Xiaolan. "The study and selection principle of model quantity of HMM in Speech recognition",. Computer Engineering and Applications, Vol.1,2009, pp. 67-69.

[8] Feng Yun, Jing Xinxing, Ye Mao. "Improving the MFCC Features for Speech Recognition", Computer Engineering & Science, Vol.31, 2009, pp.146-148.

[9] Yang Xuesong. "The interaction password recognition system of non-specific person for intelligent service robot oriented",. Peking University, 2009, pp.13-14.