Speaker Independent Recognition of Chinese Number Speeches Based on Hidden Markov Model

Feng-Long Huang, Chan-Hsing Wang, Ming-Shing Yu, Min-Xiong Wu, Yan-Kai Lai
Computer Science and Information Engineering, National United University
No. 1, Lienda, Miaoli, Taiwan, 36003
flhuang@nuu.edu.tw

Abstract–*In this paper, we describe the speaker independent speech recognition of Chinese number speeches 0~9 based on HMM. 560 speech samples are recorded and processed. The results of inside and outside testing achieve 92.5% and 76.79%, respectively. Furthermore, to improve the performance, two important features of speech; MFCC and cluster number of vector quantification, are unified and evaluated with various values. The best performance achieve 96%and 81% on MFCC Number = 20 and VQ clustering number = 64.*

Keywords: Speech Recognition, Hidden Markov Model, LBG Algorithm, Mel-frequency cepstral coefficients, Viterbi Algorithm.

1. Introduction

In Speech processing, automatic speech recognition (ASR) is capable automatically of understanding the input of human speech for the text output with various vocabularies. ASR can be applied in a wide range of applications, such as: human interface design, speech Information Retrieval (SIR) [11,12], language translation, and so on. In real world, there are several commercial ASR systems, for example, IBM's Via Voice, Mandarin Dictation System–the Golden Mandarin (III) of NTU in Taiwan, Voice Portal on Internet and 104 on-line speech queries systems. Modern ASR technologies merged the signal process, pattern recognition, network and telecommunication into a unified framework. Such architecture can be expanded into broad domains of services, such as e-commerce and wireless speech system of WiMAX.

The approaches adopted on ASR can be categorized as: 1)Hidden Markov Model (HMM) [1,2,3,4], 2)Neural Networks [5,6,7], other method is the combination of two approaches above [8,9]. The Hidden Markov Model is a result of the attempt to model the speech generation statistically, and thus belongs to the first category above. During the past several years it has become the most successful speech model used in ASR. The main reason for this success is the powerful ability to characterize the speech signal in a mathematically tractable way.

In a typical ASR system based on HMM, the HMM stage is proceeded by the parameter extraction. Thus the input to the HMM is a discrete time sequence of parameter vectors, which will be supplied to the HMM.

Chinese is a tonal speech. There are 408 base Chinese speeches and more than 1300 various speeches with 5 tones (tone 1, 2, 3, 4 and 0). In this paper, we aimed on the speaker independent recognition of number speeches. Our models are constructed based on the Hidden Markov Model (HMM). First of all, the examples of Chinese speech are recorded, and the processes for detection of end-point and windowing are processed sequentially. The component feature of speech is then extracted for the following process.

The organization of this paper is as follows. In Section II, we introduce the foundational pre-processes for ASR. In Section III, we illustrate our model of speech recognition based on HMM. The empirical results are presented and we will improve furthermore methods unifying two features in Section IV. Our conclusion and future works are presented in last section.

2. Processes of Speech

In this section, we will describe all the procedures for pre-processes.

2.1 Processing Speech

The analog voice signals are recorded thru microphone. It should be digitalized and quantified. The digital signal process can be described as follows:

$$x_p(t) = x_a(t)\,p(t), \tag{1}$$

where $x_p(t)$ and $x_a(t)$ denote the processed and analog signal. $p(t)$ is the impulse signal.

Each signal should be segmented into several short frames of speech which contain a time series signal. The features of each frame are extracted for further processes. The procedure of such pre-process is shown in Fig 1.

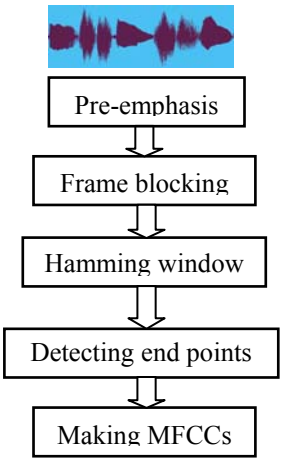


Figure 1: Pre-process of speech recognition

2.2 Pre-emphasis

Basically, the purpose of pre-emphasis is to increase, the magnitude of some (usually higher) frequencies with respect to the magnitude of other (usually lower) frequencies in order to improve the overall signal-to-noise ratio (SNR) by minimizing the adverse effects of such phenomena as attenuation distortion. The results signal before and after pre-emphasis is presented in Figure 2.

$$s2(n) = s(n) - a*s(n-1)\ , \qquad (2)$$

where $s2(n)$ denotes the output signal ,value a ranges between 0.9 and 1.0.

The Z transformation of filter is described as follows:

$$H(z)=1-a*z-1. \qquad (3)$$

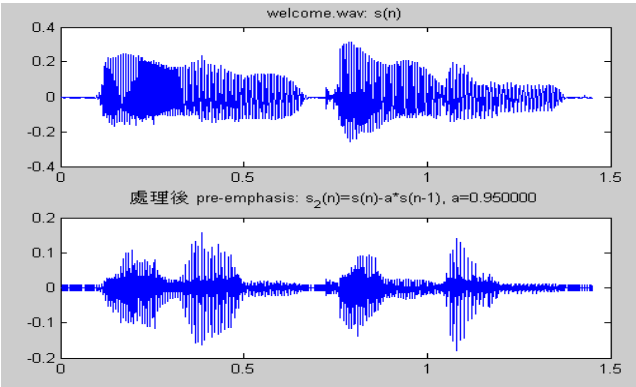


Figure 2: Comparison of before and after pre-emphasis

2.3 Frame Blocking

While analyzing audio signals, we usually adopt the method of short-term analysis because most audio signals are relatively stable within a short period of time. Usually, the signal will be segmented into time frame, say 15 ~ 30 ms.

There are always overlap between neighboring frames to capture subtle change in the audio signals. The overlapping size may be 1/3~1/2 of frame. The 3D curves of speech signal processed with hamming window are shown in Figure 3.

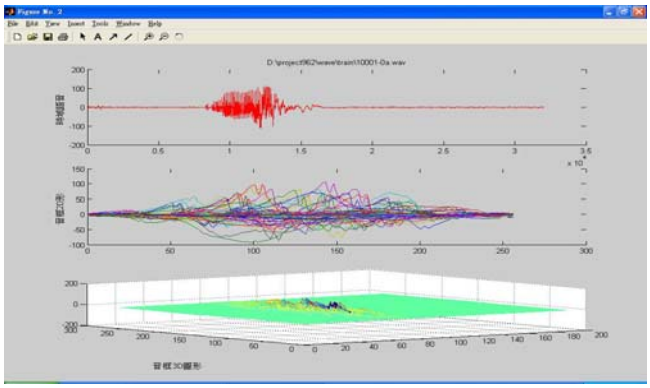


Figure 3: 3D curves of Chinese number speech 0.

2.4 Hamming Window

In signal processing, the window function is a function that is zero-valued outside of some chosen interval. The Hamming window is a weighted moving average transformation used to smooth the periodogram values.

Supposed that original signal $s(n)$ is as follows:

$$s(n),\ n = 0,\ldots N\text{-}1. \qquad (4)$$

The original signal $s(n)$ is multiplied by hamming window $w(n)$, we will obtain $s(n)^*\ w(n)$, $w(n)$ can be defined as follows:

$$w(n)\ =\ (1 - \alpha) - \alpha*\cos(2\pi n/(N\text{-}1)),\ 0 \leqq n \leqq N\text{-}1, \qquad (5)$$

where N denotes the sample number in a window.

The curves with respect to various α are shown in Figure 4. It is apparent that different hamming curve will affect the signal for overlapping frame.

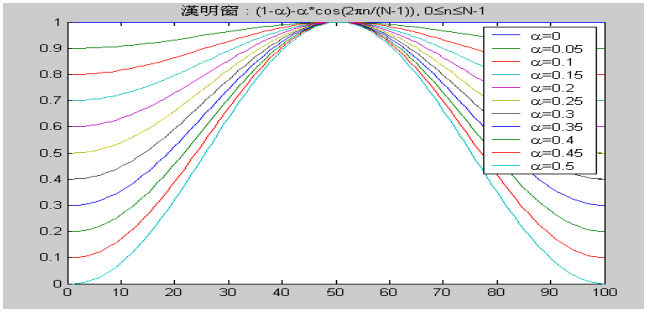


Figure 4: Hamming curves for various α values.

2.5 Zero Crossing Rate

Zero crossing rate (ZCR) is another basic acoustic feature. It is equal to the number of zero-crossing of the waveform within a given frame. The short-time zero crossing rate is defined as the weighted average of the number of times the speech signal changes sign within the time window. The ZCR for number "9" (ㄐㄧㄡˇ) is shown in Figure 5. It is significant that ZCR is higher for the consonant "ㄐ" of speech "9", however relatively lower for vowel "ㄧ", "ㄡ" and some noise in speech signal.

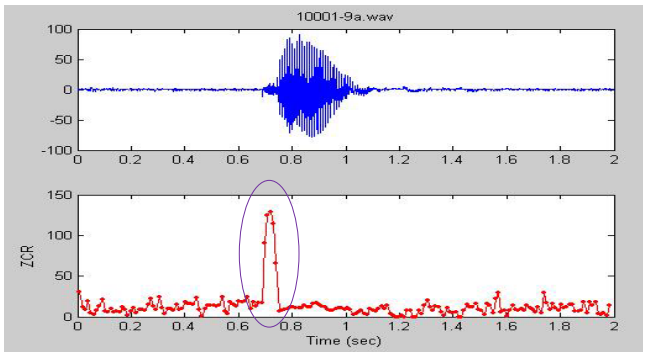


Figure 5: ZCR for Chinese number "9".

2.6 Detection of Ending Points

Our algorithm for detecting end points of speech is based upon measurement of two parameters: the short-time energy and zero crossing rate (ZCR). These measurements are given by the average value and standard deviation of the ZCR figure, as well as average energy. Among which relationships expressed by empirical parameters exist, three thresholds are needed and established: a value for the ZCR

figure and two values (a lower and an upper one) for energy.

The energy and ZCR are subsequently computed for the whole input signal over frames. The execution begins by locating, starting from the first frame, the point at which the energy profile overcomes both the lower and the upper energy thresholds, it should not descend below the lower energy threshold before having overcome the upper one. Such point, upon being identified by the lower threshold is provisionally marked as initial end point of the word. As shown in Figure 6, energy and ZCR are used to detect the end points of speech. In the figure, red and green vertical lines denote the starting and ending location of a number speech.

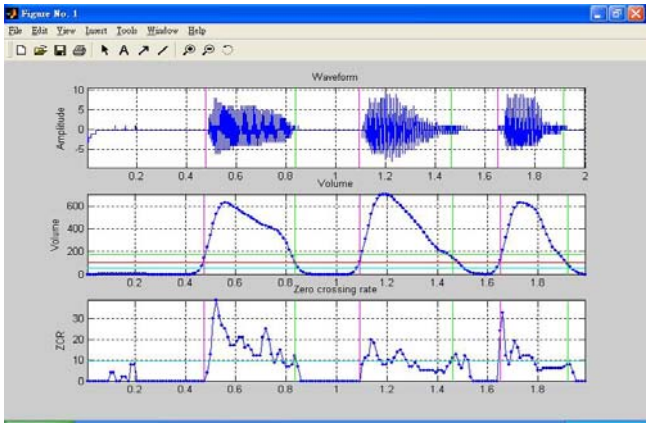


Figure 6: detection of end points.

2.7 Mel-frequency cepstral coefficients

Mel Frequency Cepstral Coefficient (MFCC) is one of the most effective feature parameter in speech recognition. For speech representation, **it** is well known that MFCC parameters appear to be more effective than power spectrum based features. MFCCs are based on the human ears' non-linear frequency characteristic and perform a high recognition rate in practical application.

- **o** lower frequency, human hear more acute.
- **o** higher frequency, human hear less acute.

As shown in Figure 7, MFCC are presented as:

$$mel(f)=1125*\ln(1+f/700) \quad (6)$$

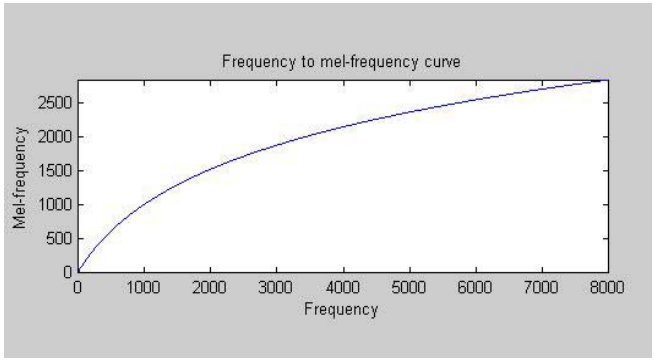


Figure 7: feature curve of mel cepstral frequency.

3. Acoustic Model of Recognition

3.1 Vector Quantification

Foundational vector quantifications (VQ) were proposed by Y. Linde, A. Buzo, and R. Gray in 1980, So-called LBG algorithm. LBG is based on k-means clustering [2,5], referring to the size of codebook G, training vectors will be categorized into *G* groups. The centroid *Ci* of each *Gi* will be the representative for such vector of codeword. In principal, the category is tree based structure. The procedure of VQ can be summarized as follows:

1. All the training vectors are merged into one cluster.
2. Select cluster features and the cluster of lowest level of tree will be divided into two parts, then executing the k-means clustering method.
3. If the number of cluster on lowest level on tree is less than expected number of codebook, go back to step 2.
4. Calculate the centroid *Ci* on lowest level on tree, which can represent each vector in cluster.

In Figure 8, *X* is the training vectors, *O* is the centroid and *Gi* is cluster *i*.

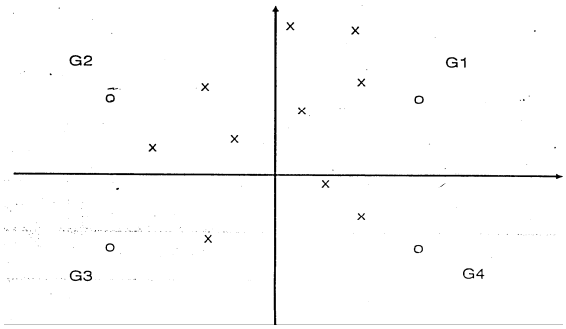


Figure 8: centroid in VQ clustering.

3.2 Hidden Markov Model

A Hidden Markov Model (HMM) is a statistical model in which is assumed to be a Markov process with unknown parameters. The challenge is to find all the appropriate hidden parameters from the observable states. HMM can be considered as the simplest dynamic Bayesian network.

In a regular Markov model, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. However, in a hidden Markov model, the state is not directly visible (so-called *hidden*), while the variables influenced by the state are visible. Each state has a probability distribution over the output. Therefore, the sequence of tokens generated by an HMM gives some information about the sequence of states.

A complete HMM can be defined as follows:

$$\lambda = (\pi, A, B) \quad . \quad (7)$$

HMM model can be defined as (π, A, B) :

1. **Π** (Initial state probability):

$$\pi = \{\pi_i = prob(q_1 = S_i)\} \quad 1 \le i \le N \quad . \quad (8)$$

2. **A (**State transition probability):

$$A = \{a_{ij} = prob(q_{t+1} = S_j \mid q_t = S_i)\} \quad (9)$$

$$1 \le i \le N$$

3. **B (**Observation symbol probability):

$$B = \{b_j(O_t) = \text{prob}(O_t \mid q_t = S_j)\} \quad 1 \le i \le N \,, \qquad (10$$

where $O = \{O_1, O_2, ..., O_T\}$ is the observation.

$S = \{S_1, S_2, S_3, ..., S_N\}$ is state symbols and

$q = \{q_1, q_2, q_3, ..., q_T\}$ is observation states and T denote the length of observation, N is the number of states.

For reducing computation, the Markov Chain can be simplified based on left-right model. Probability density function is defined as follows:

$$\mathbf{D_i}(\boldsymbol{\tau}_T) = (2\pi)^{-\frac{N}{2}} |\mathbf{R}_i|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(\boldsymbol{\tau}_\mathbf{T} - \boldsymbol{\tau}_{\mathbf{Ri}})^T \mathbf{R}_i^{-1} (\boldsymbol{\tau}_\mathbf{T} - \boldsymbol{\tau}_{\mathbf{Ri}})\right] . \qquad (11$$

where N denotes the degree of feature vectors and τ_{Ri} denotes the feature vectors for training or testing signals with respect to i^{th} probability of mixture. R_i is the i^{th} Covariance Matrix.

HMM with 2 and 3 states are shown in Figure 9 and Figure 10, respectively.

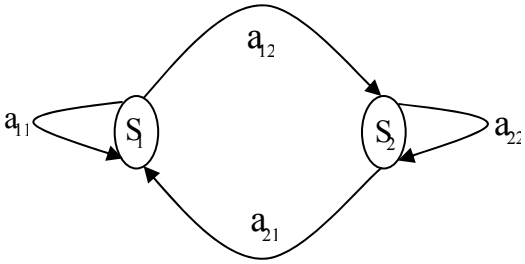


Figure 9: HMM with 2 states

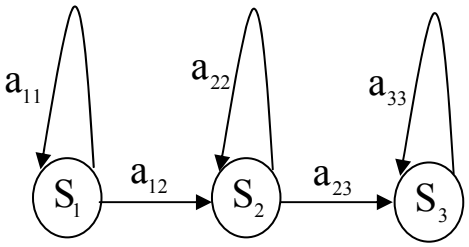


Figure 10: HMM with 3 states

3.3 System Models

The recognition system is composed of two main functions: 1) extracting the speech features, including frame blocking, VQ, and so on, 2) constructing the model and recognition based on the HMM, VQ and Viterbi Algorithm.

It is apparent that short speech signal varied sharply and rapidly, whereas longer signal varied slowly. Therefore, we use the dynamic frame blocking rather than fixed frame for different experiments.

The algorithm of dynamic frame blocking is defined as:

Input: speech vector y(i), i = 1 to n
Output: frame size frameSize
Setup frame size: frameNum = 40;
Calcupating the overlapping size offrame:
overlap = 384 - floor(length(y)/frameNum);
Count the skip size:
frameStep = round((length(y)-overlap)/frameNum);
Getting size frame:
frameSize = frameStep+overlap;

Note that 384 is decided while frameSize =512 and 3/4 overlapping frame. However, supposed that fs = 11025Hz, frameNum is defined 40, all the features will be as follows:

Min. size of speech vector	y(i) = 512
Frame size	frameSize = 376
Number of skip frame	frameStep = 4
Number of overlapping	overlap = 372
Max. size of speech vector	y(i)=15360
Frame size	frameSize = 384
Number of skip frame	frameStep = 384
Number of overlapping	overlap = 0

4. Experiments and Improvement

4.1 Recognition System Based on HMM

In the paper, we focus on speaker independent speech recognition of Chinese number speeches 0~9. All the samples with 44100 Hz/16 bits are recorded by three male adults. Total 560 samples are divided into two parts, 280 for training and 280 for testing. After complete the pre-process, such as pre-emphasis, frame boloking, VQ, the codebook is shown in Figure 11 for number speeches 0~9.

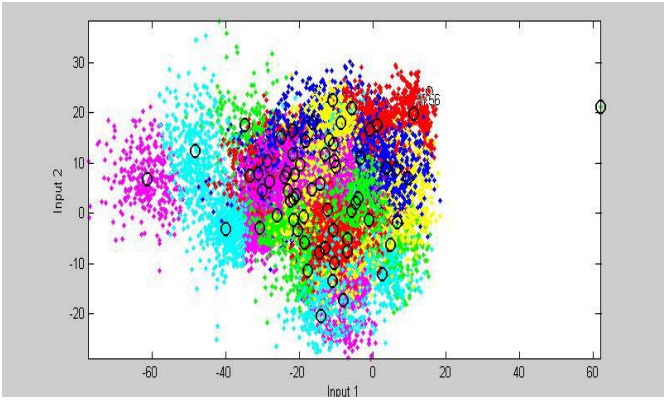


Figure 11: Distribution of Codeword & All MFCCs.

4.2 Comparison for fixed and Dynamic Frame Size

According to our empirical results, comparing the fixed and dynamic frame size, recognition rate of fixed frame size achieves 76.79%, and superior to the other with75.71%, as shown in Table 1.

Table 1: comparing the frame size, **(SymbolNum=64)**

		wave Num	Mfcc time	VQ time	HMM training	Symbol Num	rate(%)
fixed	I	280	32.9	5.77	3.44	64	90.36
	O	280					**76.79***
dynamic	I	280	32.0	3.31	2.42	64	92.50*
	O	280					75.71

PS. I and O denote the inside and outside testing, respectively

4.3. Further Improvement

4.3.1Improving the Samples of Speech

According to our empirical results, recognition rate achieve better results while cluster number=64. Inside and outside testing are 92.5% and 76.79%, respectively.

To improve the performance, we analyze all the speech wavelet. There are many samples affected by boost noise derived from human speaking or environment, as shown in Figure 12. In such a situation, the end points of boosted speech cannot be usually detected correctly. It leads to degrade the performance of system.

Usually, detecting end points judged on ZCR and energy of speech, as shown in Figure 12. However, it is significant that we need extra features to detect for noise situation. Based on experimental results and observation, the improvement rules are summarized as follows:

Input: X(n) , n = 1 to j
Output: Y(m),1 <= m <= j

1. segment the speech X(n): framedY = framed (X(n))
2. calculate the ZCR and energy for each frame.
3. smooth the curves for both ZCR and energy
4. calculate the average of first 10 frames, and multiplying 1.2. The average value will be used as the threshold for detecting process.
5. ZCR is valid only if framedY is larger than 100, as shown in Figure 13.
6. the speech will be effective only if the size is larger than 3ms.
7. the starting energy of speech should be larger than threshold.
8. the energy for continuous 5 frames of speech should be increased progressively.

Referring to the improvement, the speeches number 8 (ㄅㄚ) with boost noise can be detected, as shown in Figure 13. The improvement of detection will leads to better results for following recognition process.

4.3.2 Better Combination of Various Features

To improve furthermore the performance, two features, MFCC and cluster number, of speeches are unified and evaluated. MFCC degree varied from 8 to 36 with interval 4 and cluster number varied on 32 to 256 with interval 32. We evaluated all the combination for these two features with various numbers. The process times needed for computation are shown in Table 2.

Table 2: processed time with VQ = 64.

MFCC degree	8	12	16	20	24	28	32	36
MFCC	15.8	16.9	18.6	23.5	25.3	27.2	28.5	29.9
VQ	1.0	2.6	3.3	3.4	3.8	4.9	5.3	6.6
HMM	1.7	1.7	1.8	1.8	1.8	1.8	1.9	1.9

The best results can achieve on MFCC Number= 20 and VQ clustering number = 64. The inside and outside testing of recognition achieve 96% and 81% shown in Figure 14 and the net. results are upgraded up to 3.5% and 4.2%, respectively. Because of the limit of pages size, we just list the results with VQ = 64.

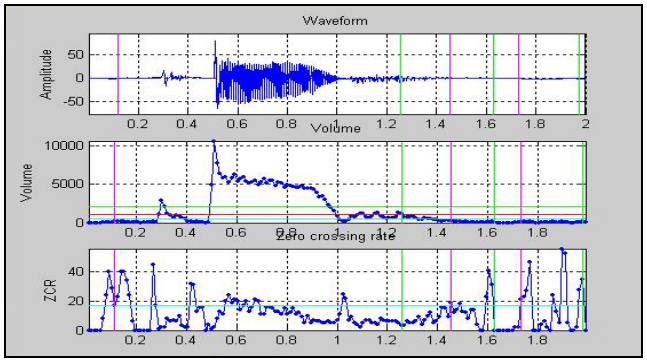


Figure 12: before improvement, Chinese number 8 (ㄅㄚ).

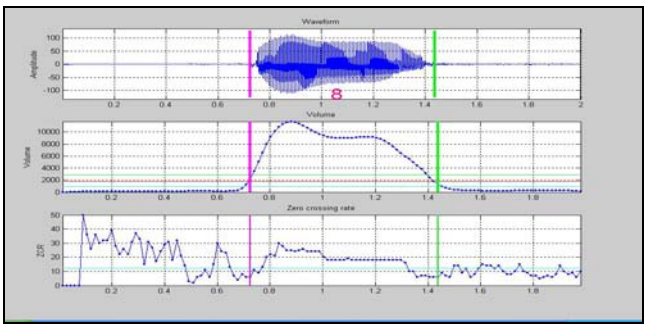
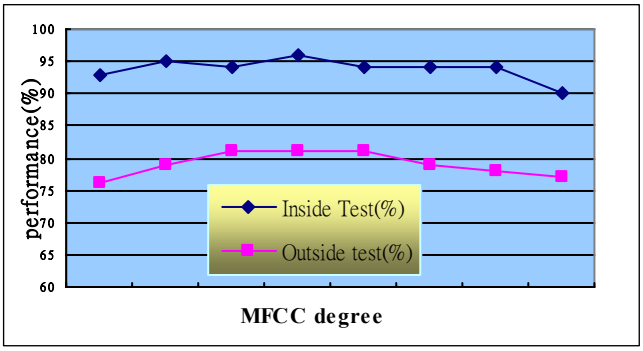


Figure 13: after improvement, Chinese number 8 (ㄅㄚ).

Figure 14: performance with VQ = 64, MFCC degrees varied between 8 and 36.



5. Conclusion

In this paper, we address the speaker independent speech recognition of Chinese number speeches based on HMM. 480 speech samples are recorded and pre-processed. the results of outside testing achieves 76.79%.

To improve furthermore the performance, two features of speeches; MFCC and VQ cluster number, are combined and evaluated. The best performance achieve on MFCC Number = 20 and VQ clustering number = 64. The final inside and outside testing of recognition achieve 96% and 81%

Several works will be researched in future:

1) Improving the selection of speech feature used for recognition.
2) Employing other effective methods to merging our approach in the paper to enhance the performance.
3) Expanding the methods into Chinese speech.

References

[1] Xiaodong Cui, Yifan Gong, “Variable Parameter Gaussian Mixture Hidden Markov Modeling for Speech Recognition”, IEEE Int. Conf. on Acoustics, Speech, and Signal Proc., Vol. 1, pp I_12 -I_15, 2003.

[2] Keng-Yu Lin, “Extended Discrete Hidden Markov Model and Its Application to Chinese Syllable Recognition”, Master thesis of NCHU, 2006.

[3] X. Li, M. Parizeau and R. Plamondon, “Training Hidden Markov Models with Multiple Observations--A Combinatorial Method”, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 22, No. 4, April 2000.

[4] Dimo Dimov and Ivan Azmanov, “Experimental specifics of using HMM in isolated word speech recognition”, CompSysTech 2005.

[5] A. Sperduti and A. Starita, “Supervised Neural Networks for Classification of Structures”, IEEE Transactions on Neural Networks, 8(3): pp.714-735, May 1997.

[6] E. Behrman, L. Nash, J. Steck, V. Chandrashekar, and S. Skinner, “Simulations of Quantum Neural Networks”, Information Sciences, 128(3-4): pp. 257-269, October 2000.

[7] Hsien-Leing Tsai, “Automatic Construction Algorithms for Supervised Neural Networks and Applications”, PhD thesis of NSYSU.

[8] T. Lee, P. C. Ching and L. W. Chan, “Recurrent Neural Networks for Speech Modeling and Speech Recognition”, IEEE ASSP, Vol. 5, pp. 3319-3322, 1995.

[9] Li-Yi Lu, “The Research of Neural Network and Hidden Markov Model Applied on Personal Digital Assistant”, Master thesis of CYU, 2003.

[10] Rabiner, L. R., “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”, Proceedings of the IEEE, Vol.77, No.22, pp.257-286, 1989.

[11] Manfred R. Schroeder, H. Quast, H.W. Strube, “Computer Speech: Recognition, Compression, Synthesis” , Springer, 2004.

[12] Wald, M., “Learning Through Multimedia: Automatic Speech Recognition Enabling Accessibility and Interaction”. Proceedings *of ED-MEDIA 2006:* World Conference on Educational Multimedia, Hypermedia & Telecommunications. pp. 2965-2976.