An Effective Approach for Chinese Speech Recognition on Small size of Vocabulary

Feng-Long Huang

Computer Science and Information Engineering, National United University

No. 1, Lienda, Miaoli, Taiwan, 36003

flhuang@nuu.edu.tw

Abstract:

In this paper, an effective approach for Chinese speech recognition on small vocabulary size is proposed - the independent speech recognition of Chinese words based on Hidden Markov Model (HMM). The features of speech words are generated by sub-syllable of Chinese characters. Total 640 speech samples are recorded by 4 native males and 4 females with frequently speaking ability. The preliminary results of inside and outside testing achieve 89.6% and 77.5%, respectively. To improve the performance, keyword spotting criterion is applied into our system. The final precision rates for inside and outside testing in average achieve 92.7% and 83.8%. The results prove that the approach for Chinese speech recognition on small vocabulary is effective.

Keywords:

Hidden Markov Model, Mel-Frequency Cepstral Coefficients, Sub-syllable, Viterbi Algorithm, Keywords Spotting.

1. Introduction

Natural language processing (NLP) is one of the most rapidly developing areas of technology in recent years. NLP focus on the speech and languages. In Speech processing, automatic speech recognition (ASR) is capable 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) [12, 13], and so on. In real world, there are several commercial ASR systems, for example, IBM's Via Voice [18], Voice Portal on Internet and speech keyword's 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.

The approaches adopted on ASR can be categorized as:

1) Hidden Markov Model (HMM) [4, 5, 6, 28],
2) Neural Networks [7, 8, 9, 10],
3) Combination of two approaches above [10, 11].

During the past several years, Hidden Markov Model 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.

DOI : 10.5121/sipij.2011.2205

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.

This paper is organized as follows. In Section 2, the foundational preprocesses for ASR are described. In Section 3, we illustrate acoustic model of speech recognition based on HMM. The empirical results and improving method are presented in Section 4. Conclusion and future works are summarized in last section.

2. The Approach for Speech Processes

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

The analog voice signals are recorded thru microphone. It should be digitalized and quantified. 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.

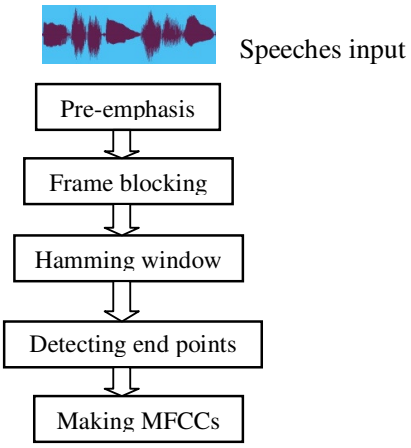

Fig. 1: Preprocess of speech recognition

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

$$x_p(t) = x_a(t)\, p(t)$$

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

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.

$y[0] = x[0]$

$y[n] = x[n]-0.95x[n]\ ,\quad 1<= n <= N$ (1)

where N is the sampling size.

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 Fig. 2.

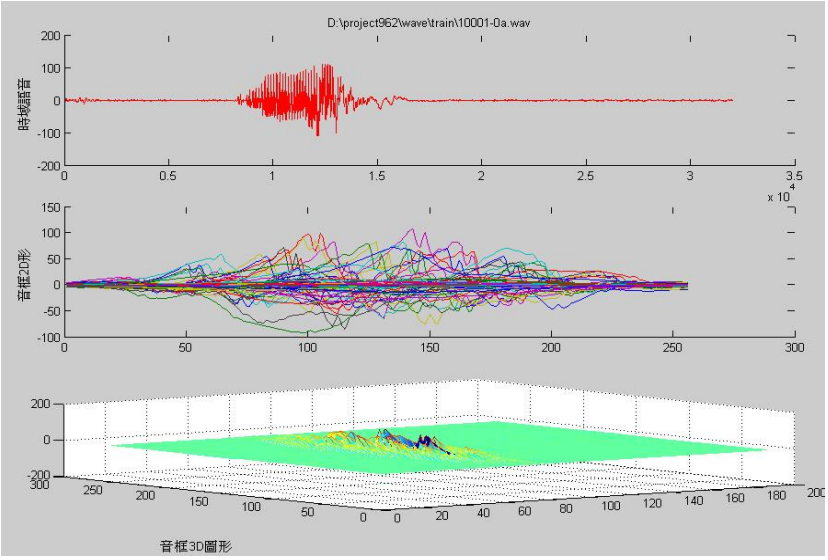

Fig. 2: 3D curves of Chinese number speech 0.

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.

$s[n] = y[n]*w[n]$

$$= y[n] * \left\{0.54 - 0.46cos\left\langle\frac{2n\pi}{n-1}\right\rangle\right\} \quad (2)$$

The curves with respect to various α are shown in Fig. 3. It is apparent that different hamming curve will affect the signal for overlapping frame.

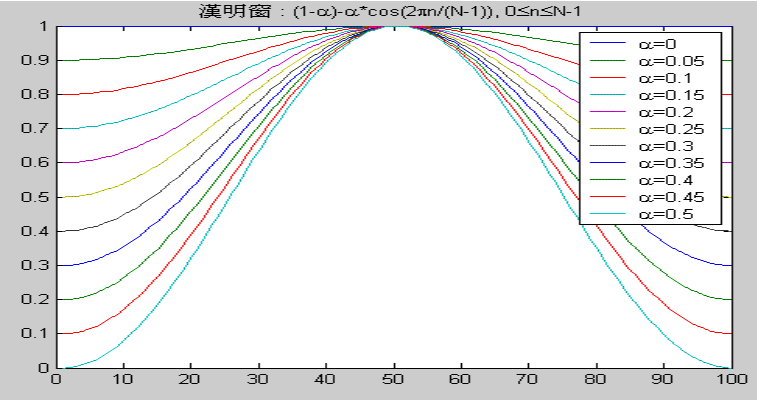

Fig. 3: Hamming curves for various α values.

Zero Crossing Rate

Zero crossing rate (ZCR) is another basic acoustic feature. Our algorithm for detecting end points of speech is based upon measurement of two parameters: the short-time energy and zero crossing rates. 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 Fig. 4, 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. Several novel methods for speech enhancement can be found in [25].

The ZCR for Chinese character "9" (jiu3 - ㄐㄧㄡˇ[1]) is presented in Fig. 4. It is so obvious that ZCR is higher for the consonant "ㄐ" of speech "9", however relatively lower for vowel "ㄧ", "ㄡ" and some noise in speech signal.

[1] The systems of Taiwan phonetics symbols and hanyu pinyi are used in the paper.

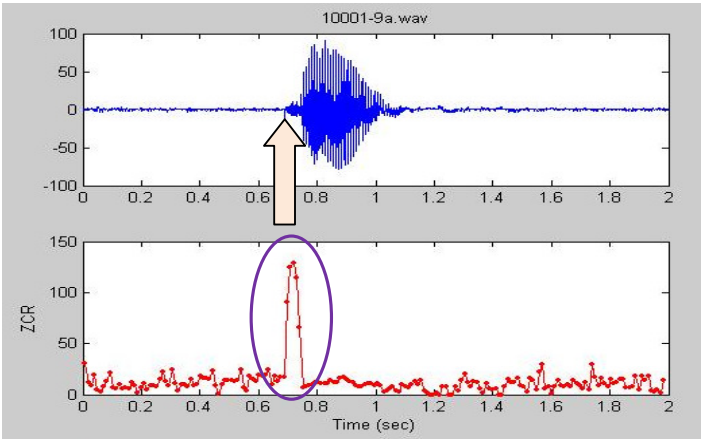

Fig. 4: ZCR for Chinese number "9".

Mel-frequency cepstral coefficients (MFCCs)

MFCCs [15, 24, 26] are based on the human ears' non-linear frequency characteristic and perform a high recognition rate in practical application. The detection scheme used for assessment objective assessment of stuttered disfluencies based on the feature of MFCC [27]. Mel-Frequency Cepstral Coefficient is presented in Fig. 5:

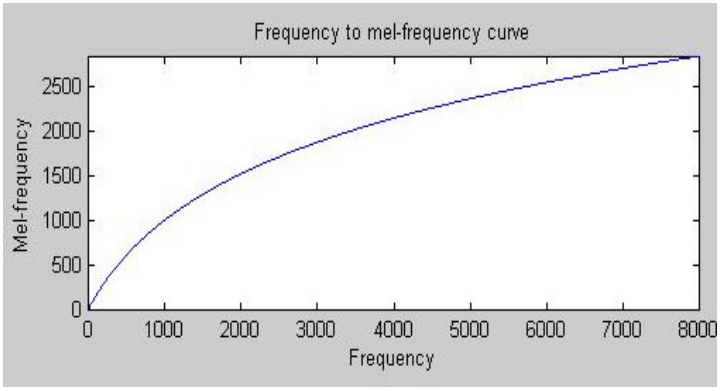

Fig. 5: feature curve of MFCCs.

3. Acoustic Model of Recognition

3.1 Vector Quantification

Foundational vector quantifications (VQ) were proposed by Y. Linde, A. Buzo, and R. Gray - 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 Fig. 6, *X* is the training vectors, *O* is the centroid and *Gi* is cluster *i*.

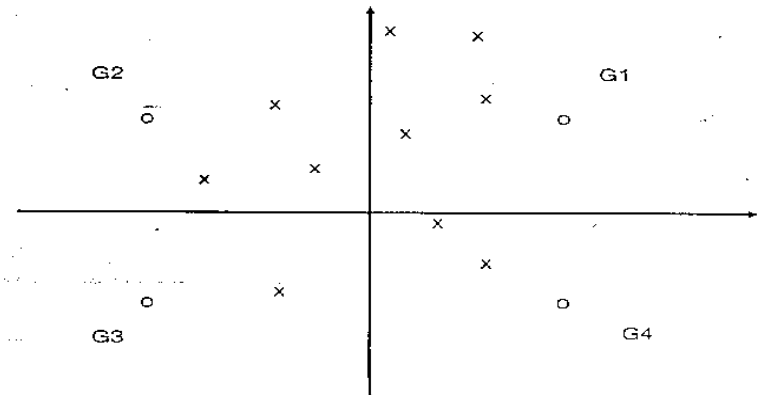

Fig. 6: centroids in VQ clustering.

3.2 Hidden Markov Model

A Hidden Markov Model (HMM) [4, 14, 15] 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:

$$\lambda = (\pi, A, B) \quad , \tag{3}$$

where π, *A* and *B* denote **i**nitial state probability, state transition probability and observation symbol probability, respectively.

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

$$\mathbf{D_i}(\boldsymbol{\tau}_T) = (2\pi)^{-\frac{N}{2}} \left|\mathbf{R}_i\right|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(\boldsymbol{\tau}_\mathbf{T} - \boldsymbol{\tau}_\mathbf{Ri})^T \mathbf{R}_i^{-1}(\boldsymbol{\tau}_\mathbf{T} - \boldsymbol{\tau}_\mathbf{Ri})\right] \tag{4}$$

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 as follows in Fig. 7:

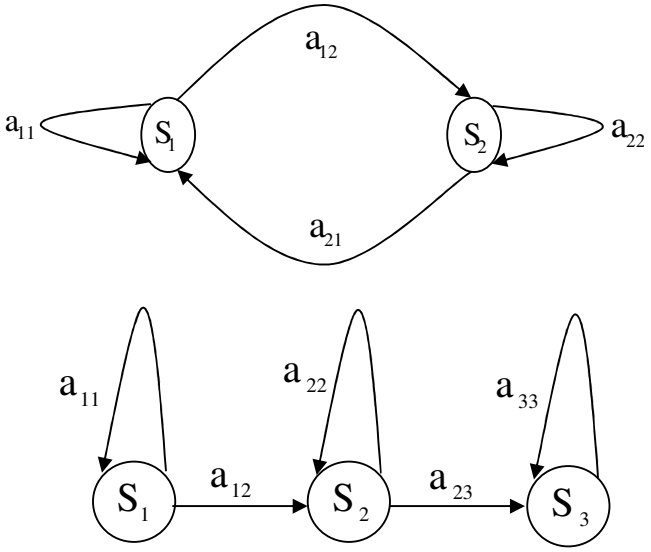

Fig. 7: HMM with 2 states (upper) and 3 states

3.3 Lexical Structure of Chinese

Chinese is a mono-syllable and tonal language. Basically, each syllable of Chinese character is composed of a initial (such as ㄅ, ㄆ, ㄇ) and a final (such as ㄚ, ㄛ, ㄜ and some combinations) with 1,2,3,4 and neutral. There are 21 initials and 36 finals in Chinese speeches. From the linguistic view points, phoneme is regarded as a smallest unit of speeches. A initial comprise one consonant, however each final may be composed of one or two vowels (phonemes), such as ㄧ, ㄚ, ㄧㄚ or ㄩㄢ.

The basic unit of Chinese language is characters, which are furthermore employed to generate the Chinese words. In the systems of Taiwan phonetics symbols, the sub-syllable is regarded as more flexible component for better speech recognition on small size of trained speeches datum. In tour experiments, the sub-syllable contain all the initials and finals, such as ㄧㄢ, ㄧㄚ and ㄩㄝ.

The speech and frequency spectrum for Chinese word: "電話-dian4 hua4" (telephone) are shown in Fig. 8. The speech and frequency spectrum (spectrograms) for Chinese word: "電腦-dian4 nao3" (computer) are shown in Fig. 9. Spectrum is one of important features in speech recognition [19, 20]. According to the Fig. 8 and 9, it is obvious that the correlation between signal strength and lightness in frequency spectrum. The more brilliant of area in the spectrum domain, the stronger of speech. Note that the shadow area on upper picture in Fig. 8 is the speech of first Chinese character "電-dian4 (ㄉㄧㄢ)" in the word "電話". There is also a silence speech with variable length on both starting and ending side as shown in the following figure and "電腦-dian4 nao3" (computer).

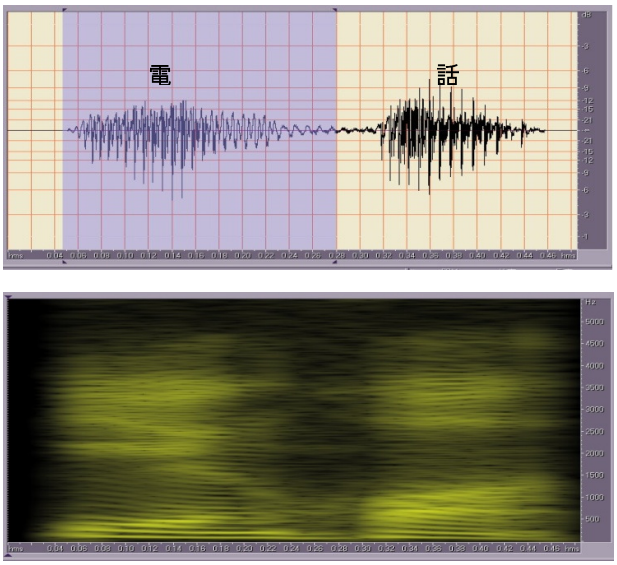

Fig. 8: up: speeches of Chinese word "電話" (telephone)
down: frequency spectrum of "電話"

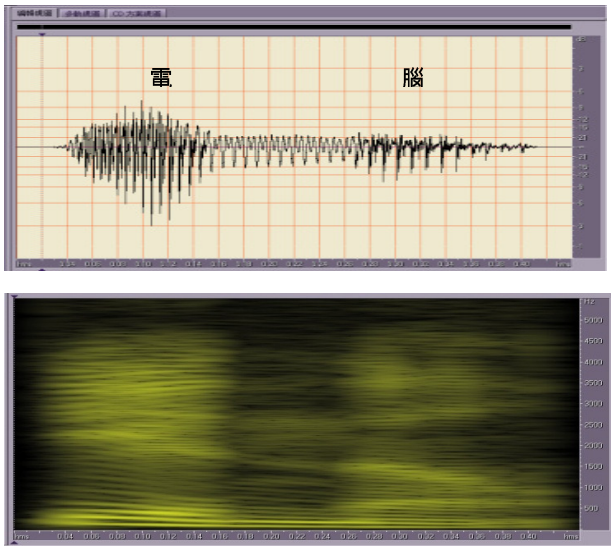

Fig. 9: up: speeches of Chinese word "電腦" (computer)
down: frequency spectrum of "電腦"

3.4 Our Systems

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.

In order to find the best mapping sequence between training model and input speech model, searching methods are usually employed. The Viterbi algorithm (VA) is a recursive optimal solution to the problem of estimating the state sequence of a discrete-time finite-state Markov process observed in memoryless noise. Therefore, the Viterbi algorithm is the frequently used technique for many problems in areas, such as digital communications [1] and speech process of Text-To-Speech (TTS) systems [2].

Viterbi algorithm can be principally described as following steps:

- Step 1: Initialization
- Step 2: Recursion
- Step 3: Termination
- Step 4: Path Backtracking

4. Experiments

4.1 Recognition System Based on HMM

The HMM-based prediction is shown to give considerably more accurate frame classification and a lower RMS pitch error. Increasing the number of clusters in each state of the HMM enables more detailed modeling of the joint distribution of MFCC and pitch [25]. We focus on the Chinese speech recognition on small size of vocabulary. In the paper, the speeches are recorded by 4 native males and 4 females with frequent ability of Chinese speaking. There are totally 640 speech samples for 80 Chinese keywords. The samples rate is 44100 Hz/16 bits/ stereo. All these speeches are divided into two subsets, 480 for training (inside) and 160 for testing (outside).

4.2 Performance Results

Two kinds of frame size can be used in speech recognition: fixed and dynamic size [17]. 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

There are 3 and 6 states for initials and finals of Chinese characters on our HMM models. Based on the Preliminary experiments, comparing the fixed and dynamic frame size, precision rate of the former is better than that of the latter. Therefore the fixed Frame Size of speeches is employed in the paper. The precision rates in average are 89.6% and 77.5% for training and testing, as shown in Table 1.

Table 1: The initial precision rates for the fixed frame size

	Speech Num	Mfcc time	VQ time	HMM training	Precision rate(%)
I	480	32.0	3.54	2.92	89.6
O	160				77.5

According to the experimental results, the final containing two vowels rather than one, such as ㄧㄚor , ㄨㄢ, leads to higher fault and degrade the recognition performance. In our observation, the state number of HMM of the finals can be extended into more states, such as 8 or 9 states. It will improve the recognition for the finals of Chinese words meanwhile the processing time will be longer.

4.3 The Improving Method

4.3.1 Keyword spotting

In the section, we employee one of useful methods, Keyword spotting [21, 22, 23], for improving the recognition can be regarded as a subfield of speech recognition dealing with the identification of keywords in speeches. There are two types of criteria for promising keyword spotting:

- Keyword spotting in unconstrained speech
- Keyword spotting in isolated word recognition

Keyword spotting in isolated word recognition appears when the keywords are separated from other texts by silences. In the paper, the technique that applied in such problems is dynamic time warping (DTW).

It is always apparent that the more candidates the lower the recognition rate. The keywords spotting technique is improving methods for speech recognition. Each word is composed of two or more characters, such as the Chinese word 電話 (Telephone) or 電腦 (computer). The

word can be organized based on the character chain, in which the preceding and successive characters will be generated as chain-like structure, presented in Fig. 10.

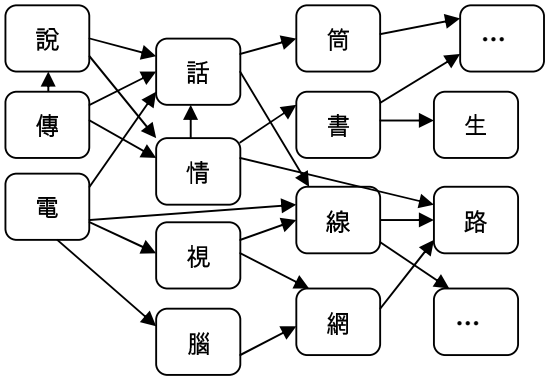

Fig. 10: the lattice of keyword spotting for some Chinese words.

4.3.2 Precision Enhancement

In the chain structure, the possible successive characters of one character in a word are always finite and more definite. The possible candidates for recognition will be decreased and it usually leads to higher precision rate. The keyword spotting criterion is furthermore applied into our system. The final precision rates for inside and outside testing in average achieve 92.7% and 83.8%. Comparing Table 1 with Table 2, the net results are upgraded +3.1% and +6.3%, respectively.

Table 2: The improved precision rates based on improving technique of keyword spotting

	Speech Num	Mfcc time	VQ time	HMM training	Precision rate(%)
I	480	32.0	3.54	2.92	92.7 (+3.1)
O	160				83.8 (+6.3)

The experiments show the keyword spotting will enhance the accuracy of speech recognition by reducing the number of possible candidate speeches and alleviates the prediction errors effectively. The results prove that this approach for speech recognition on small vocabulary is effective.

5. Conclusions

An effective approach for Chinese speech recognition on small vocabulary is proposed in the paper. The features of speech words are generated by sub-syllable of Chinese characters. Total 640 speech samples are recorded by 4 males and 4 females. There are 3 and 6 states for initials and finals of Chinese characters on our HMM models.

Based on the experiments, the preliminary results of inside and outside testing achieve 89.6% and 77.5%. In order to improve the recognition rate, keyword spotting criterion is applied into the system. Final precision rates achieve 92.7% and 83.8%. The results prove that out proposed approach for Chinese speech recognition on small vocabulary is effective and flexible.

Several works will be furthermore researched in future:

1) Considering the meticulous structure of speech syllable for HMM.
2) Employing other methods to enhance the recognition performance.
3) Expanding the methods into Chinese speech on large vocabulary.
4) Analyzing Contextual structure of Chinese words for speech recognitions.

ACKNOWLEDGE

Author would like to thank Liao-Ho Foundation of Education Taiwan for financially supporting the project.

REFERENCES

[1] Tsao-Tsen Chen, Shiau-He Tsai, 2008, Reduced-complexity wrap-around Viterbi algorithm for decoding tail-biting convolutional codes, 14th Wireless Conference, 2008, pp. 1-6.

[2] Hanna Sil´en, Elina Helander, Jani Nurminen, Konsta Koppinen and Moncef Gabbouj, 2010, Using Robust Viterbi Algorithm and HMM-Modeling in Unit Selection TTS to Replace Units of Poor Quality, INTERSPEECH 2010, pp. 166-169.

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

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

[5] Liqin Fu Xia Mao Lijiang Chen , July 2008, Speaker independent emotion recognition based on SVM/HMMS fusion system , International Conference on Audio, Language and Image Processing-ICALIP 2008, pp. 61-65.

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

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

[8] Hsien-Leing Tsai, 2004, Automatic Construction Algorithms for Supervised Neural Networks and Applications, PhD thesis of NSYSU.

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

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

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

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

[13] Wald, M., 2006, 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.

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

[15] Maitreyee Dutta, Renu Vig, 2006, AN IMPROVED METHOD OF SPEECH COMPRESSION USING WARPED LPC AND MLT-SPIHT ALGORITHM, Proceedings of the 6th WSEAS International Conference on Signal, Speech and Image Processing, Lisbon, Portugal, 2006 Sep. 22-24, pp.85-94.

[16] D-Furui, S., Feb. 1986, Speaker-independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 34, Issue 1, pp. 52-59.

[17] Haamid M. Gazi, Omar Farooq, Yusuf U. Khan, Sekharjit Datta, 2008, Wavelet-based, Speaker-independent Isolated Hindi digit Recognition, International Journal of Information and Communication Technology, Vol. 1 , Issue 2 pp. 185-198

[18] http://www.ask.com/questions-about/Via-Voice

[19] Liang-che Sun, Chang-wen Hsu, and Lin-shan Lee, 2007, MODULATION SPECTRUM EQUALIZATION FOR ROBUST SPEECH RECOGNITION, IEEE workshop on Automatic Speech Recognition and Understanding (ASRU-2007).

[20] Li Deng, 2007 Roles of high-fidelity acoustic modeling in robust speech recognition, IEEE workshop on Automatic Speech Recognition and Understanding (ASRU-2007).

[21] K. Thambiratnam and S. Sridharan, 2005, Dynamic Match Phone-Lattice Searches for Very Fast and Accurate Unrestricted Vocabulary Keyword Spotting, IEEE ICASSP-2005.

[22] Jansen, A. Niyogi, P., 2009, Point Process Models for Spotting Keywords in Continuous Speech, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 17, Issue 8, Nov. 2009, pp. 1457.

[23] Shuyong Bai, Linlin Li and Chew Lim Tan, 2009, Keyword Spotting in Document Images through Word Shape Coding, 2009 10th International Conference on Document Analysis and Recognition.

[24] ZHEN Bin, WU Xihong, LIU Zhimin, CHI Huisheng, 2001, On the Importance of Components of the MFCC in Speech and Speaker RecognitionActa Scientiarum Naturalium Universitatis Pekinensis (ASNUP), 2001, Vol.37 No.3 pp.371-378.

[25] Ch.V.Rama Rao, M.B.Rama Murthy and K.Srinivasa Rao, 2011, Noise Reduction Using mel-Scale Spectral Subtraction with Perceptually Defined Subtraction Parameters - A New Scheme, Signal & Image Processing - International Journal (SIPIJ), Vol.2, No.1, March 2011, pp.135-149.

[26] Xu Shao and Ben Milner, 2004, PITCH PREDICTION FROM MFCC VECTORS FOR SPEECH RECONSTRUCTION, IEEE ICASSP 2004, pp. I-97-100.

[27] K.M Ravikumar, R.Rajagopal, H.C.Nagaraj, 2009, An Approach for Objective Assessment of Stuttered Speech Using MFCC Features, DSP Journal, Volume 9, Issue 1, June, 2009, pp.19-24.

[28] Ibrahim Patel and Y. Srinivas Rao, 2010, Speech Recognition Using HMM with MFCC-An Analysis Using Frequency Specral Decomposion Technique , Signal & Image Processing - International Journal (SIPIJ), Vol.1, No.2, December 2010, pp.101-110.