J. CPOL (Journal of Computer Processing of Oriental Languages), 11(3): 221-231, March. 1998
*An extended version of the paper which has been presented in ICCPOL'97 [*1:204-207, Apr. 2, 97, Hong Kong*], **invited.***

A Real-World Speech Recognition System Based on CDCPMs

Fang Zheng, Haixin Chai, Zhijie Shi, Wenhu Wu, and Ditang Fang
Speech Laboratory, Department of Computer Science and Technology,
Tsinghua University (THU), Beijing, 100084, P.R.China
Tel.: +86-10-62784141, Fax: +86-10-62771138
E-mail: *fzheng@sp.cs.tsinghua.edu.cn*, *fzheng@cenpok.net*

Abstract

In this paper a real-world continuous-manner 2000-phrase speaker-independent Chinese speech recognition system is introduced, the vocabulary of which consists of 2000 Chinese phrases and each phrase is made up of 3 to 5 Chinese syllables. This system is based on a robust statistical model named Center-Distance Continuous Probability Model (CDCPM) using the Embedded Multiple-Model (EMM) scheme and efficient knowledge-based search strategies for speech recognition. Users can speak any phrase in the vocabulary in continuous manner. The accuracy of this system is 97.4% on an average.

Keywords: Speaker-Independent Speech Recognition, center-distance continuous probability model (CDCPM), embedded multiple-model (EMM)

1 Introduction

In this paper, a continuous-manner 2000-phrase speaker-independent Chinese speech recognition system is presented.

The model used in this system, namely center-distance continuous probability model (CDCPM) [1], will preserve only the B-matrix of a HMM, and every observation output probability density function (PDF) is replaced by a one-dimensional (center-distance) probability density function (PDF). This replacement will reduce the time and space complexities to a great extent, preserving good performance.

During the scoring procedure of the recognition, many efficient search strategies are adopted to the CDCPMs. In the continuous speech recognition a very natural idea is to segment the input phonic stream into units corresponding to the acoustic models, then to recognize them one by one. But there seldom is a successful system based on this segmentation strategy for time being because it is not easy to find an efficient segmenting method to mark the starting and ending points of each unit exactly. Anyway, there are still many distinctive acoustic features available for segmentation. Based on such a situation, after a long time study we developed a set of segmenting rules, where we do our best to give exact segmentation points and leave unsure parts to the followed frame-based searching procedure. It is very efficient for the followed procedure when using the knowledge obtained previously. For example, most boundaries between connected syllables are determined, which can lower down or eliminate much useless searching cost. With no more than one-tenth time expense to the traditional frame-based searching algorithm, satisfactory results can be got.

This system is flexible in adding phrases to, modifying phrases in, and deleting phrases from the vocabulary without additional training because the system is based on Chinese syllables. It is natural to choose syllables as the recognition units because Chinese is a kind of syllable-based language. As a result, anyone who wants to modify or even replace the whole vocabulary set only needs to edit a plain text file, in which the Chinese character and pinyin string are given. Because the traditional acoustic models are too complex to build a system based on syllable, CDCPMs are used instead.

The established 2000 - phrase continuous - manner Chinese speech recognition system has extremely good performance even under the real-world environment. For testing bed, the average rate is about 97.4%.

The paper is organized as follows. In Section 2, the feature extraction method is presented. In Section 3, the CDCPM is described briefly. In Section 4, a giant real-world Chinese speech database used to train the CDCPMs is briefly described and in Section 5, the system solutions are given. Experimental results and summary are given in Section 6 and 7 respectively.

2 Feature Extraction

In our experiments, speech signal is digitized at 16 kHz sampling rate with 8 kHz cut-off , emphasized using a simple first-order digital filter with the transfer function $H(z) = 1 - 0.95z^{-1}$. The pre-emphasized speech is then blocked into frames of 32 msec in length spaced every 16 msec. Having been weighted by the Hamming Window, each frame is represented by a set of *D*-order (where *D*=16) LPC cepstral coefficients $\{c_d\}_{d=1}^{D}$ [2,3]. Regression analysis [4] is applied to each time function of the cepstral coefficients over several frames every 16 msec and the regression coefficients $\{r_d\}_{d=1}^{D}$ are obtained then.

Each of the two set of coefficients is constructed as a vector in a *D*-dimensional Euclidean space and the weighted Euclidean distance between two vectors $\vec{x}_1$ and $\vec{x}_2$ is defined as

$$y(\vec{x}_1, \vec{x}_2) = \sqrt{\sum_{d=1}^{D} w_d (x_{1d} - x_{2d})^2}\ , \tag{1}$$

where $\vec{x}_1$ and $\vec{x}_2$ can be cepstral vectors or regression vectors and $\vec{\mathbf{w}} = (w_1, w_2, \ldots, w_D)$ is the weight vector. In our experiments, the *d'*th component of the weight vector, i.e., w_d, is chosen to be the reciprocal of the statistical variance of *d'*th cepstral component so that each component contributes statistical equally in distance measure. Actually, this kind of weighted Euclidean distance measure is a Mahalanobis distance measure where the covariance matrix is simplified to a diagonal matrix.

Now the utterance can be represented by time functions of the cepstral vector sequence and the regression vector sequence.

Instead of combining them into one large *2D*-dimensional vector $\vec{\mathbf{v}}(t) = (\vec{\mathbf{c}}(t), \alpha\vec{\mathbf{r}}(t))$ where α is the balance coefficient, we use the cepstral vector $\vec{\mathbf{c}}(t)$ and its corresponding regression vector $\vec{\mathbf{r}}(t)$ separately, these two kinds of vectors are described by their own probability density functions (PDFs), which has been proved better [1].

3 The CDCPM

In this section, the Center-Distance Continuous Probability Model (CDCPM) will be described briefly.

3.1 The CDN Distribution

Denote the PDF of a random variable ξ with a normal distribution by $\mathrm{N}(x;\mu_x,\sigma_x)$, where μ_x is its mean value and σ_x is its standard deviation. Define a new random variable $\eta = |\xi - \mu_x|$, we have the PDF of η as

$$p(y;\sigma_x) = \frac{2}{\sqrt{2\pi}\,\sigma_x}\exp(-y^2/2\sigma_x^2), \quad y \geq 0, \tag{2}$$

where the mean value μ_y of η can be calculated to be $\mu_y = \frac{2\sigma_x}{\sqrt{2\pi}}$. In fact, η is the distance between the normal variable ξ and its mean value μ_x, thus the defined distribution is referred to as a Center-Distance Normal (CDN) distribution. And the CDN pseudo-PDF can be

$$\mathrm{N}_{CD}(x;\mu_x,\mu_y) = \frac{2}{\pi\mu_y}\exp(-y^2(x,\mu_x)/\pi\mu_y^2) \tag{3}$$

The *D*-dimensional case is similar to the mono-dimensional case. Denote the (weighted) Euclidean distance between a *D*-dimensional normal vector $\vec{\xi}$ and its mean value vector $\vec{\mu}_x$ by another random variable η. Assume η is a CDN variable, then its CDN pseudo-PDF is similarly as

$$\mathrm{N}_{CD}(\vec{x};\vec{\mu}_x,\mu_y) = \frac{2}{\pi\mu_y}\exp(-y^2(\vec{x},\vec{\mu}_x)/\pi\mu_y^2) \tag{4}$$

As a matter of fact, $\mathrm{N}_{CD}(\vec{x};\vec{\mu}_x,\mu_y)$ is not the PDF of $\vec{\xi}$ but that of $y(\vec{\xi},\vec{\mu}_x)$, i.e., the distance between a normal vector and its mean vector, it is just for convenience and comparison purpose.

3.2 The CDCPM

A left-to-right CDCPM is in some sense similar to the HMM except that the CDCPM ignores A matrix, and is based on the CDN distribution instead of the normal distribution. A mixture density CDCPM can be described by the following parameters: (1) *N*: number of states per model; (2) *M*: number of densities per state; (3) *D*: number of dimensions per feature vector; (4) $\vec{\mu}_{xnm} = (\mu_{xd}^{(nm)})$: mean vector of the *m*'th density component in *n*'th state; (5) μ_{ynm} : mean center-distance of the *m*'th density component in *n*'th state; (6) g_{nm} : density gain of *m*'th density component in *n*'th state. Here $1 \leq n \leq N$, $1 \leq m \leq M$, $1 \leq d \leq D$, and the observation PDF has the similar form, which is called a mixed CDN density.

3.3 The Scoring Scheme in Acoustic Models

In a continuous hidden Markov model (CHMM) with mixed Gaussian densities (MGD) [5] in each state, Baum-Welch [6], Viterbi [7,8] algorithms and many better improved versions have been available for the training and recognition. Given an observation feature sequence $\mathbf{O} = (\vec{\mathbf{o}}_1, \vec{\mathbf{o}}_2, \cdots, \vec{\mathbf{o}}_T)$ of *T* frames and a CHMM $\Lambda = \{\pi, A, B\}$ with *N* states where the initial probability distribution is $\pi = (\pi_i)_N$, the state transition matrix is $A = (a_{ij})_{N\times N}$, and the output observation PDF matrix is $B = (b_j(\cdot))_N$, the probability density of the CHMM Λ generating $\mathbf{O}$ is

$$f\{\mathbf{O}|\Lambda\}=\sum_{S} f\{\mathbf{O},S|\Lambda\}=\sum_{S}\Pr\{S|\Lambda\}\cdot f\{\mathbf{O}|\Lambda,S\}=\sum_{S}\left(\pi_{s_1}\prod_{t=2}^{T}a_{s_{t-1}s_t}\right)\cdot\left(\prod_{t=1}^{T}b_{s_t}(\vec{\mathbf{o}}_t)\right)$$

$$=\sum_{S}\left(\pi_{s_1}\cdot b_{s_1}(\vec{\mathbf{o}}_1)\prod_{t=2}^{T}a_{s_{t-1}s_t}\cdot b_{s_t}(\vec{\mathbf{o}}_t)\right) \tag{5}$$

where $S=\{s_t|1\le t\le T\}$ is an arbitrary state transition sequence. The Viterbi algorithm gives a maximum likelihood (ML) state sequence $S^{(ML)}=\{s_t^{(ML)}|1\le t\le T\}$, and takes $f\{\mathbf{O},S^{(ML)}|\Lambda\}$ as the final matching score, which is only one term of the sum in Equ. (5), i.e.,

$$Score\{\mathbf{O}|\Lambda\}=f\{\mathbf{O},S^{(ML)}|\Lambda\}=f\{\mathbf{O}|\Lambda,S^{(ML)}\}\cdot\Pr\{S^{(ML)}|\Lambda\} \tag{6}$$

CHMMs can describe signals very well, but the estimation of model parameters will cost too much. Researches on model distance measures show that the A matrix contributes not too much as B does to the recognition performance [9], so we ignore the A matrix and then the matching score between a given observation sequence $\mathbf{O}=(\vec{\mathbf{o}}_1,\vec{\mathbf{o}}_2,...,\vec{\mathbf{o}}_T)$ with the CDCPM $\Lambda=\{\vec{\mu}_{xn},\mu_{yn},b_n(\cdot)|1\le n\le N\}$ is calculated as

$$Score\{\mathbf{O}|\Lambda\}=f\{\mathbf{O}|\Lambda,S^{(ML)}\}=\prod_{t=1}^{T}b_{s_t^{(ML)}}(\vec{\mathbf{o}}_t) \tag{7}$$

As for the HMMs, something like the mixed Gaussian densities (MGD) [10], tied MGD [11], or other forms [12] can be used as the observation PDF or scoring function $b_n(\cdot)$ except that the Gaussian densities are replaced with the CDN densities. The Bayesian learning method [13] can be employed for a CDCPM to train $b_n^{(c)}(\cdot)$ and $b_n^{(r)}(\cdot)$ in Equ. (8).

The mixed CDN densities have the following equations for cepstral and regression representations:

$$b_n^{(c)}(\vec{\mathbf{c}})=\sum_{m=1}^{M}g_{nm}^{(c)}\mathrm{N}_{CD}(\vec{\mathbf{c}};\vec{\mu}_{xnm}^{(c)},\mu_{ynm}^{(c)}) \tag{8-1}$$

$$b_n^{(r)}(\vec{\mathbf{r}})=\sum_{m=1}^{M}g_{nm}^{(r)}\mathrm{N}_{CD}(\vec{\mathbf{r}};\vec{\mu}_{xnm}^{(r)},\mu_{ynm}^{(r)}) \tag{8-2}$$

where $1\le n\le N$, $1\le m\le M$, $1\le d\le D$, and $b_n^{(c)}(\cdot)$ and $b_n^{(r)}(\cdot)$ are the PDFs of cepstral and regression features in state n. The scoring function for vector $\vec{\mathbf{v}}=(\vec{\mathbf{c}},\alpha\vec{\mathbf{r}})$ is defined as

$$b_n(\vec{\mathbf{v}}(t))=b_n^{(c)}(\vec{\mathbf{c}}(t))*b_n^{(r)}(\vec{\mathbf{r}}(t)) \tag{9}$$

where α is of no use.

In our system, the following scoring function is proposed which is based on Nearest-Neighbor rule

$$b_n^{(c)}(\vec{\mathbf{c}})=\max_{1\le m\le M}\mathrm{N}_{CD}(\vec{\mathbf{c}};\vec{\mu}_{xnm}^{(c)},\mu_{ynm}^{(c)}) \tag{10-1}$$

$$b_n^{(r)}(\vec{\mathbf{r}})=\max_{1\le m\le M}\mathrm{N}_{CD}(\vec{\mathbf{r}};\vec{\mu}_{xnm}^{(r)},\mu_{ynm}^{(r)}) \tag{10-2}$$

The PDFs in Equ. (10) are often different from those in Equ. (8) because of the different modeling method.

The scheme is referred to as an Embedded Multiple-Model (EMM) one and can be explained in this way. Assume there is a well-trained left-to-right CDCPM with N states and M CDN densities each state and an unknown speech feature sequence $\mathbf{O}=(\vec{\mathbf{o}}_1,\vec{\mathbf{o}}_2,...,\vec{\mathbf{o}}_T)$. There exists a segmentation determining which state it belongs to for any $\vec{\mathbf{o}}_t$. For any segmentation, scoring using Eq. (10) leads to choosing a maximal matching

score from M^T one-density CDCPMs. These M^T one-density CDCPMs can be regarded to be embedded in the original *M*-density *N*-state CDCPM. Thus the original CDCPM are called an EMM.

The EMM scheme has been proved efficient and powerful, especially for gender-dependent, accent-dependent, and context-dependent models and so on. If *M* is well chosen, it is enough for one CDCPM to represent several different cases for each vocabulary word [14].

3.4 Recognition in an Isolated System

The given unknown observation sequence **O** is first segmented before it is scored. The matching score with a CDCPM is calculated using equations in Section 3.3.

This kind of scoring strategy is useful and proved efficient for isolated word recognition [1]. But it is not easy to be used in a continuous recognition scheme. Hence come the auto-segmentation strategies adopted in the continuous speech recognition, which will be discussed in Section 5 in detail.

4 Database Description

The speech database used to train here is a giant Chinese database, uttered by 80 people aged from 16 to 25 from all over the country. Speakers consist of 40 males and 40 females. The sub-vocabulary for each person includes mono-syllable word set (11groups by 100 Chinese words), bi-syllable word set (63 groups by 100 words), tri-syllable word set (11 groups by 100 words), quad-syllable word set (10 groups by 100 words), penta-syllable word set (1 group by 76 words), hexa-syllable word set (1 group by 23 words) and hepta-syllable word set (1 group by 10 words). Five sub-vocabularies make up a complete vocabulary. As a matter of fact, the database uttered by 80 people is a 16 times' repetition of the vocabulary. In the vocabulary, 419 Chinese syllables do not occur equally, instead, the occurrence frequency for each syllable depends on the frequency it occurs in all the Chinese words (phrases) found in a Chinese dictionary. After all, each speaker utters ten sentences different from those uttered by any other speakers.

Words or sentences are required to be uttered in Chinese Mandarin with a little local accent under an environment with some background noise, so that the obtained real-world speech database will be more available in practice.

Speech is first filtered to a bandwidth of 8 kHz (cutoff frequency) and then digitized at 16 kHz sampling rate. Such a giant database consists of 25GB speech data, about 230 hours' utterances.

5 System Solutions

There are totally 400 syllables or so in Chinese speech. In a live environment, when people speak freely, the syllable has been proved to be the best choice for the speech recognition unit [1,14,15]. Based on this experience, the 2000-phrase system uses Chinese syllable as the speech recognition unit.

5.1 Training the CDCPMs

There are two kinds of CDCPMs in the system. The first kind is corresponding to the Chinese syllables, each syllable is represented by one CDCPM with state number *N*=6, density number per state *M*=4. Another kind is used to describe the noise, which is called a single-state NOISE model with *N*=1 and *M*=16.

In this system, all CDCPMs are trained using speech data taken from the real-world database described in Section 4.

The training procedure is simple. For each syllable CDCPM: (1) Each observation feature sequence **O** from the training database is first segmented into *N* segments (corresponding to *N* states) using the Non-Linear Segmentation (NLS) method [16]. (2) For segment *n*, vectors of this segment from each observation sequence are collected together and then grouped into *M* classes using some clustering algorithm such as the famous LBG algorithm [17]. (3) Estimation of $\vec{\mu}_{xnm}$ and μ_{ynm} is very easy for each density, namely each class, for the specified segment *n*.

For the NOISE model, the training procedure is similar except that we need not segment any noise segments.

5.2 Recognition Strategy

The frame-based searching algorithm is commonly used nowadays in the continuous speech recognition (CSR). According to the vocabulary and the structure of the acoustic model, a searching accidence tree will be first established. The searching procedure will try to traverse every possible path from root to leaf among the accidence tree and choose the path with the maximum likelihood. The phrase corresponding to this path is the final recognition result. For a large vocabulary recognition system it is almost impossible to traverse every path, so some pruning strategy must be adopted. In fact, the frame-based algorithm [18] is also a Viterbi decoding procedure by frames [7]. Our first system was developed in this way but the accuracy rate was as low as 89.7%. However the expenses in space and time were very high because many apparently useless paths must be passed through and saved during such a recognition procedure.

Actually, the result of searching algorithm can be regarded as a segmentation of the input speech with maximum likelihood, so it is natural to have an idea of finding the best segmentation method in a CSR system. Traditionally, a segmentation procedure includes the following steps: (1) segmenting the original speech into units corresponding to the acoustic models; (2) scoring and recognizing them separately; (3) connecting the recognized units together to form the final result, i.e., a phrase or a sentence. In addition, it is well known that Chinese is a monosyllable language and every syllable has the unique Consonant-Vowel (C/V) structure, which makes the segmentation easier than any other languages. Most boundaries between syllables can be marked by experienced researcher even only by viewing the original speech wave. Even so, it is still not easy to find a method to segment the phonic stream into syllables exactly because of the variety, complexity and co-articulation of the speech.

We have been studying this in detail and have our knowledge-based searching strategy.

Our motivation here is to join the above two strategies together to get better performance, that is, first using the acoustic knowledge to make marks on the phonic stream as many as possible, then performing a knowledge-based searching procedure to get the correct results effectively. The new strategy is roughly as follows:

Firstly, making the original marks at those points (frame position) where the cepstrum varies greatly according to a pre-defined threshold. By carefully adjusting thresholds we are sure to make all starting and ending points of the occurred syllables included in these marks. Now we get the primitive segments of the phonic stream. These segments are usually three or four times as many as the actual syllables.

Secondly, trying to get more about these segments according to our acoustic rules. Chinese is a syllable-based language and there is often a short silence (gap) between two adjacent syllables. The acoustic features such as the frame energy and the frame zero-crossing rate of the silence are unique to other segments. Also some unvoiced Chinese consonants have their own distinctive phonic characteristics. All these are often very useful and efficient for us to use to point out most of these silence segments. According to our rules, all frames will be checked, but marked are only those that the segmentation procedure is sure about.

Thirdly, performing a frame-based searching procedure using the knowledge obtained previously. During the searching procedure, a single syllable in a path may cover several primitive segments or only one segment. The obtained knowledge is made full use of during the whole searching procedure. For instance, if a silence segment is detected, the obtained information is that it can not be a part of any syllable. Another example is, if an unvoiced consonant is detected, we know that all the paths must have a syllable beginning with this segment.

All the knowledge makes the searching procedure more efficient. As a matter of fact, the knowledge used here is not only for accelerating but also necessary. In the traditional frame-based searching, the pruning strategy is very sensitive. If too many paths are pruned, the recognizer can seldom get right answer, on the other hand, if too many paths are reserved the recognizer will not bear. Our experiments show that the traditional frame-based searching may get wrong answer even if a complete search is performed.

The frame-based searching algorithm can be treated as another segmentation method without any acoustic knowledge. In our system, with previous segmentation results, the choice of frame-based searching algorithm on boundary between syllables is limited. The algorithm can focus its all efforts on trying to find a best path with maximum likelihood. Such an acoustic knowledge based method is proved to be more efficient and with higher performance.

6 Experimental Results

A 2000-phrase continuous-manner speech recognition system has been established, the vocabulary consists of 2000 Chinese phrases of 3 to 5 syllables. Users can change the vocabulary set freely just by editing a plain text file. In Table 1, recognition rates for training and testing sets are listed. Further research is in progress to enlarge the training data amount, data will be taken from the above giant speech database.

Table 1. *Performance of the 2000-phrase real-world system*

Training Set	Rate	Testing Set	Rate
M00	99.65%	M10	97.80%
M01	99.90%	M11	98.00%
M02	99.90%	M20	95.40%
M03	99.95%	M21	98.40%

7 Summary

Through the experiments, we can draw the following conclusions:

(1) CDCPMs with the EMM scheme are powerful in reducing the time and space complexities. For speech recognition, what should be considered first is the simplicity and efficiency instead of the mere mathematical perfection and the complexity. The CDCPM is a simplified version of the HMM, but what is simplified is the relatively not important part. The CDCPM focuses on the more important factors for speech recognition, resulting in the satisfactory results.

(2) The knowledge-based search strategies are extremely useful in a continuous-manner speech recognition system. According to our 2000-phrase system, knowledge-based strategies are useful and important. The blind search without any knowledge will cause two bad effects, low efficiency and low accuracy. The use of knowledge in the continuous-manner speech recognition application is good evidence.

(3) The strategies discussed in this paper are very useful for not only the current system but also the dictation machine. We believe that the knowledge-based strategies are useful in the continuous speech recognition system or the dictation machine. Further researches on continuous speech recognition are in progress.

References

[1] **F. Zheng, W.-H. Wu, D.-T. Fang,** "CDCPM with Its Applications to Speech Recognition," *J. of Software*, Vol.7, No. 10, 1996, pp.69-75 (in Chinese)

[2] **J. Makhoul,** "Linear Prediction: A Tutorial Review," in Proceedings of *IEEE*, Vol. 63, No. 4, 1975, pp.562-580
[3] **B. Gold and C.M. Rader**, *Digital Processing of Signals*. New York: McGraw-Hill, 1969, p.246
[4] **S. Furui**, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol.34, No.1, Feb., 1986, pp.52-59
[5] **L.R. Rabiner, B.-H. Juang, S.E. Levinson, M.M. Sondhi,** "Recognition of Isolated Digits Using Hidden Markov Models With Continuous Mixture Densities," *AT&T Technical Journal*, Vol. 64, No.6, July-August 1985, pp.1211-1234
[6] **L.E. Baum,** "An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of Markov Processes," *Inequalities*, Vol. 3, 1972
[7] **A.J. Viterbi,** "Error Bounds for Convolutional Codes and An Asymptotically Optimum Decoding Algorithm," *IEEE Transactions on IT*, Vol. 13, No. 2, Apr., 1967, pp.
[8] **G.D. Forney,** "The Viterbi algorithm," in Proceedings of *IEEE*, Vol. 61, No. 3, March, 1973, pp.268-278
[9] **B.-H. Juang and L.R. Rabiner,** "A probabilistic distance measure for hidden Markov models," *AT&T Tech. J.,* Vol.64, No.2, Feb. 1985, pp.391-408
[10] **J.G. Wilpon, L.R. Rabiner, C.-H. Lee, and E.R. Goldman**, "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 38, No.11, Nov. 1990, pp. 1870-1878
[11] **J.R. Bellegarda and D. Nahamoo**, "Tied Mixture Continuous Parameter Modeling for Speech Recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 38, No.12, Nov. 1990, pp.2033-2045
[12] **H. Ney**, "Modeling and Search in Continuous Speech Recognition," in Proceedings of *European Conf. On Speech Technology*, Vol.1, Berlin, 1993, pp.491-498
[13] **J.-L. Gauvain and C.-H. Lee**, "Bayesian Learning for Hidden Markov Model with Gaussian Mixture State Observation Densities," *Speech Communications*, Vol. 11, No. 2-3, June 1992, pp.205-213
[14] **F. Zheng, W.-H. Wu, D.-T. Fang**, "Research on the Speech Recognition Model in the Chinese Dictation Machine," *J. of Tsinghua University*, Vol. 37, No.9, Sept. 1997, pp.37-40
[15] **F. Zheng,** *Studies on Approaches of Keyword Spotting in Unconstrained Continuous Speech*: [Ph.D. Dissertation]. Beijing: Department of Computer Science and Technology, Tsinghua University, June 1997
[16] **L. Jiang, W.-H. Wu, L.-H. Cai, and D.-T. Fang**, "A Real-time Speaker-independent Speech Recognition System Based on SPM for 208 Chinese Words," in Proceedings of *International Conference on Signal Processing* (*ICSP'90)*, Beijing, 1990, pp.473-476
[17] **Y. Linde, A. Buzo, and R.M. Gray**, "An Algorithm for Vector Quantization Design," *IEEE Transactions on COM*, Vol. 28, No. 1, Jan., 1980
[18] **C.-H. Lee, and L.R. Rabiner**, "A Frame-Synchronous Network Search Algorithm for Connected Word Recognition," *IEEE Trans. on ASSP*, Vol. 37, No. 11, Nov.1989, pp. 1649-1658