A Two-Stage Algorithm for Mandarin Consonant Recognition

Ming-Tzau Lin

Assistance Professor, Department of Computer and Communication Engineering,

De-Lin Institute of Technology

Abstract

A two-stage algorithm with multiple features for Mandarin consonant recognition was proposed. We classified Mandarin consonant into seven classes, each with similar phonetic and acoustic characteristics. In the first stage of the proposed algorithm, consonant classification was conducted by computing features to construct mixture density models for each feature and for each consonant class, and then we defined for the decision making of consonant classification. Our multi-speaker experiments showed that a high correct percentage (97.9% for the top two selections) could be achieved in our first-stage consonant classification. In the second stage, with the selected top two consonant classes as candidates, detailed consonant recognition is performed by feature extraction of cepstral coefficients followed by the pattern comparison using the continuous hidden Markov model (CHMM) or the segmental probability model (SPM). The experimental results showed that both SPM and CHMM can achieve comparable accuracy percentage of 89.2%, but the SPM requires only about 1/8 processing time. Furthermore, the proposed two-stage algorithm is superior to the one-stage algorithm (omitting the first stage) by about 3% accuracy rate increasing and 12% processing time saving.

Keywords: Mandarin consonant recognition, consonant classification, pattern comparison.

兩階段演算法之中文聲母語音辨認

林 明 灶

德霖技術學院 電腦與通訊工程系 助理教授

摘要

在聲母辨認方面，鑑於各聲母特性差異頗大，以單一特徵參數及單一樣型比對法不易達成高辨認率，因此我們設計了一個二階段聲母辨認法。第一階段先執行聲母分類，我們依國音聲母的語音及聲學特性將其分為七類，並選定六種時域及頻域特徵參數據以設計輸入語音與各聲母類別間的距離度量法則。實驗結果顯示以本方法判定的前二排名聲母類別含正確選擇者高達 97.9%。第二階段再以選定的前二排名聲母類別為候選進行個別聲母辨認，我們以倒頻係數(Cepstral Coefficients)為特徵參數，再以連續隱藏式馬可夫模型(Continuous Hidden Markov Model, CHMM)或分段機率模型(Segmental Probability Model, SPM)作樣型比對。實驗結果顯示使用前述二法的平均辨認正確率相當，均達 89.2%，但 SPM 僅需約 1/8 的計算量。此外，相對於不經第一階段聲母分類而逕行以第二階段步驟直接進行聲母辨認，我們所提出的二階段聲母辨認法不但辨認正確率提高近 3%，計算時間亦可節省 12%以上。

關鍵字:中文聲母辨認、聲母分類、樣本比較。

1. Introduction

Phone recognition of speech utterances offers a promising approach toward large vocabulary recognition of speech. This is particularly true for Mandarin speech of Chinese language due to its special monosyllabic structure. Chinese word is composed of one to several characters. Each syllable can be decomposed into an "INITIAL/FINAL" format, in which "INITIAL" means the initial consonant of the syllable while "FINAL" means the succeeding vowel or diphthong part, possibly with a medial or nasal ending. The consonant could be absent; in that case, a syllable is just a FINAL [1]. There are a total of 21 consonants in the Mandarin Chinese. Table 1 shows the classification of Mandarin consonants according to their pronunciation places and manners based on the Chinese Phonetic Alphabet (CPA) notation [2]. The selection of INITIALs and FINALs as basic speech units is appropriate for the structure of acoustic-phonetic approach. This paper is proposed for Mandarin consonant recognition.

Low consonant recognition rate will degrade the possibility for extending the system to large vocabulary size. Mandarin consonants, and also the consonants in other languages, are considered highly difficult in machine recognition mostly due to their relative shorter duration and transient characteristic. Previously, only limited researches have been reported about Mandarin consonant recognition. Liu and Wang [3] employed two models, namely, the hidden Markov models and the neural network models to develop a Mandarin consonant recognizer. They used a speech data set of 52 Mandarin syllables with Tone 1 and Tone 4 provided by four male speakers. The consonant parts of the entire database were phonetically hand-labeled in the training stage. The experiment results showed that the recognition accuracy of 95.2% was achieved. However, this result is obtained only for small vocabularies. Poo and Ou [4] developed a large time-delay neural network (TDNN) for recognition of the Mandarin consonants. His database consists of a set of 1345 Mandarin syllables produced by a female speaker. The entire set of database was phonetically hand-labeled in the training stage. His results showed that the recognition accuracy was 92.7%. However, Poo's TDNN algorithm applied only for speaker-dependent task. In addition, for large vocabulary recognition, the strategy of modularity and incremental learning should be further explored for TDNN approach.

The results of our preliminary investigation suggested that using only a single pattern classifier with limited features is nearly impossible to achieve a recognition rate as high as desired in our Mandarin consonant recognizer. Consequently, a two-stage algorithm using extensive features was

Table 1. Articulatory classification of Mandarin consonants (The symbol used is in CPA notation [2])

Manner / Place	Plosive		Affricate		Fricative		Liquid	Nasal
	unaspirate	aspirate	unaspirate	aspirate	unvoiced	voiced		
Labial	ㄅ/b/	ㄆ/p/						ㄇ/m/
Labiodental					ㄈ/f/			
Front part of tongue tip			ㄗ/tz/	ㄘ/ts/	ㄙ/s/			
Tongue tip	ㄉ/d/	ㄊ/t/					ㄌ/l/	ㄋ/n/
Back part of tongue tip (retroflex)			ㄓ/j/	ㄔ/ch/	ㄕ/sh/	ㄖ/r/		
Alveolar palatal			ㄐ/j(i)/	ㄑ/ch(i)/	ㄒ/sh(i)/			
Velar	ㄍ/g/	ㄎ/k/			ㄏ/h/			

proposed in this research. Our phonetic and acoustic study showed that the Mandarin consonants could be classified into seven classes, each with distinct characteristic features. The basic idea of our proposed algorithm thus is, by using extensive time- and frequency-domain features, to identify the input speech data into the most possible candidate classes, and then followed by a detailed consonant recognition by using a pattern comparison method. This proposed algorithm includes stages of consonant classification and consonant recognition. As shown in Fig. 1, this proposed algorithm includes stages of consonant classification and consonant recognition. For the first stage, Mandarin consonants are classified into seven classes based on their phonetic and acoustic characteristics. Five time- and frequency-domain features are computed to construct mixture density model for each feature and each consonant class. The appropriate minimum scorings were experimentally decided to construct the complete consonant classification rule. In the second stage, the detailed consonant recognition is performed by using the pattern comparison method.

The paper is organized into five sections. In Section 2, three research topics that are class assignment, the selection of effective acoustic features and the decision rule for classification processing of first stage are introduced. In Section 3, the CHMM and SPM are employed for the second stage. The experiment results and improvements are reported in Section 4. Finally, a conclusion is made in section 5.

2. Consonant Classification in the First Stage

For consonant classification, three research topics should be emphasized: 1) class assignment, i.e., how to properly decide the total number of classes and the belonging of each consonant, 2) the selection of effective acoustic features with high discrimination for different consonant classes, and 3) the decision rule for classification processing.

2.1 Class Assignment

Since the manner of articulation is the key factor for differentiating various consonants [5], our consonant classification is basically along this line. There are eight kinds of manners of articulation for Mandarin consonants, namely aspirate plosive, unaspirate plosive, aspirate affricate, unaspirate affricate, unvoiced fricative, voiced fricative, liquid, and nasal. The liquid sound, the nasals and the voiced fricative are combined to form one class called sonorant since they are the only four voiced sounds in Mandarin consonants. The detailed grouping is listed and symbolized as below:

(1) C_1 = the set of unaspirate plosives (UP) = {ㄅ/b/, ㄉ/d/, ㄍ/g/},

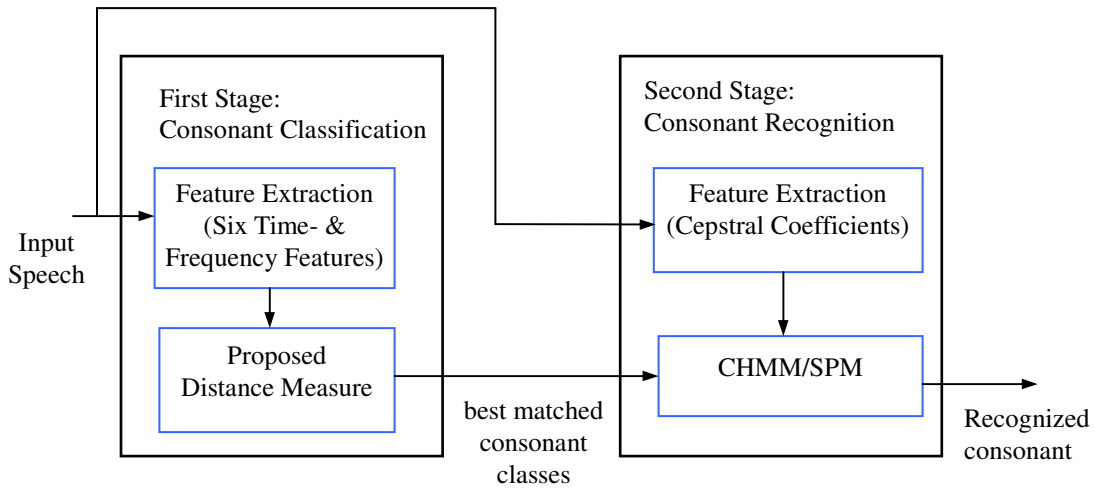

Fig. 1 Block diagram of the proposed two-stage scheme for Mandarin consonant recognition.

(2) C_2 = the set of aspirate plosives (AP) = {ㄆ/p/, ㄊ/t/, ㄎ/k/},
(3) C_3 = the set of unaspirate affricates (UA) = {ㄓ/j/, ㄗ/tz/, ㄐ/j(i)/},
(4) C_4 = the set of aspirate affricates (AA) = {ㄔ/ch/, ㄘ/ts/, ㄑ/ch(i)/},
(5) C_5 = the set of unvoiced fricatives 1 (UF1) = {ㄕ/sh/, ㄙ/s/, ㄒ/sh(i)/},
(6) C_6 = the set of sonorant (S) = {ㄇ/m/, ㄋ/n/, ㄌ/l/, ㄖ/r/ },
(7) C_7 = the set of unvoiced fricatives 2 (UF2) ={ㄈ/f/, ㄏ/h/}.

where "/ /" denotes the corresponding English transcription with a similar pronunciation.

2.2 Feature Selection

The ideal of feature selection is to determine a set of signal features that can describe distinct acoustic properties for different consonant classes. After some preliminary experiments, six time- and frequency-domain features, namely, the consonant duration (F^1), the average power (F^2), the pitch period (F^3), the average zero crossing rate (F^4), the energy ratio between the high- and low-frequency bands (F^5), and the energy ratio between the full- and the middle frequency band (F^6) are selected for our consonant classification. Six feature values are computed to form a feature vector $\boldsymbol{y} = [y^1, y^2, y^3, y^4, y^5, y^6] \in F^1 \times F^2 \times F^3 \times F^4 \times F^5 \times F^6$, where $\times$ represents the Cartesian product. Specifically, for each consonant $\{s(n) \mid n = 1...N\}$, where N is the number of sample points, the values for the six selected features are defined as next paragraph.

In our training stage, the training data set was normalized by maximum and minimum values of each feature, which it guaranteed all feature values are between 0 and 1 for each selected feature F^j $(1 \le j \le 6)$ and each class C_i $(1 \le i \le 7)$ as

$$x^j = \frac{y^j - \min(y^j)}{\max(y^j) - \min(y^j)}. \tag{1}$$

To construct the probability density function (pdf) for each selected feature $F^j (1 \le j \le 6)$, and each class $C_i (1 \le i \le 7)$, in the form of multivariate mixture Gaussian (MMG) function [6].

$$\begin{aligned} p(x_i^j) &= \sum_{m=1}^{M} a_{ijm} p_m(x_i^j) \\ &= \sum_{m=1}^{M} a_{ijm} N(x_i^j; \mu_{ijm}, \upsilon_{ijm}) \end{aligned} \tag{2}$$

where M is the number of mixture components, $N(x_i^j; \mu_{ijm}, \upsilon_{ijm})$ is the Gaussian density function in feature j class i (with mean μ_{ijm} and variance υ_{ijm}) for the m-th mixture component, and a_{ijm} is the weight for the m-th mixture component in feature j class i determined by the K-means algorithm [7]. The mixture coefficient a_m stand for a *priori* probabilities as

$$\sum_{m=1}^{M} a_m = 1. \tag{3}$$

and its mixture component given by

$$a_m = L_m / \sum_{k=1}^{M} L_k \tag{4}$$

where L_m is the number of each subset. The essential advantage of the mixture density is that several maxima in the density function can be modeled that may correspond to different acoustic realizations of the same phoneme due to coarticulatory effects.

The six selected features are defined as follows:

(1) The duration of consonant segment: $y^1 \subset F^1 \subset [0,\infty)$,

$$y^1 = N .$$

The consonant duration is proved to be useful in characterizing different consonant classes. In our preliminary investigation, we find that the maximum duration is 325.3 ms and minimum duration is 3.9 ms in consonants. The set of unaspirate plosives (UP) had the shortest duration (the average length of "ㄅ/b/" is 22.3 ms, "ㄉ/d/" is 22.2 ms and "ㄍ/g/" is 25.8 ms"), and the sets of aspirate affricate (AA) (the average length of "ㄘ/ts/" is 106.6 ms, "ㄔ/ch/" is 107.7 ms and "ㄑ/ch(i)/" is 116.7 ms") and unvoiced fricative-1 (UF1) (the average length of "ㄙ/s/" is 104.2 ms, "ㄕ/sh/" is 125.5 ms and "ㄒ/sh(i)/" is 133.6 ms") had the longest duration in the consonant classes. To utilize this feature to discriminate the UP, AA and UF1 is useful. Fig. 2 illustrates the mixture probability distribution curves of consonant average duration for UP, AP, and UF1 classes. There are significant discriminate between these three curves. It can be seen that the maximum center of average duration of UP is around 25.7 ms, AP is 54.5 ms, and UF1 is 106.1 ms.

(2) The average power: $y^2 \subset F^2 \subset [0,\infty)$,

$$y^2 = \frac{1}{N}\sum_{n=1}^{N} s^2(n). \tag{5}$$

The average energies of unaspirate affricate set is the lowest energy in all-consonant classes. The maximum power is 3.08×10^8 and minimum power is 1.18×10^4 (the resolution of our experiment data is 16-bits; the amplitude values are from –32768 to 32767). Fig. 3 illustrates the mixture probability distribution curves of consonant average duration for UF2, AA, and S classes (the scale is from 0 to 0.02). There are significant discriminate between these three curves.

(3) The pitch period: $y^3 \subset F^3 \subset [0,\infty)$,

The pitch period is used to discriminate the voiced speech that includes the sonorant consonants. The pitch period is obtained by using the Rabiner's autocorrelation pitch detector [19]. The short-term autocorrelation function $r(\eta)$ of the INITIAL segment is first computed as

$$r(\eta) = \frac{1}{N}\sum_{n=1}^{N} s(n)s(n-\eta). \tag{6}$$

The *pitch period* is then determined in samples by the formula:

$$y^3 = \arg\max_{\eta}(r(\eta)). \tag{7}$$

(4) The average zero-crossing rate, $y^4 \subset F^4 \subset [0,1]$,

Zero-crossing rate (ZCR) is proved to be useful in characterizing different audio speech. It has been popularly used in voiced/univoced speech classification algorithms. The set of UF1 and UA had larger ZCR, and the sets of S had smaller ZCR in the consonant classes. In our experiments, we have found that the maximum ZCR is 0.699 and minimum ZCR is 0.0139. Fig. 4 illustrates the mixture probability distribution curves of ZCR for UA and S classes.

ZCR is define as

$$y^4 = \frac{1}{N}\sum_{n=1}^{N}\frac{\left|\text{sgn}\{s(n)\} - \text{sgn}\{s(n-1)\}\right|}{2} \tag{8}$$

$$\text{where}\quad \text{sgn}\{s(n)\} = \begin{cases} +1 & s(n) \geq 0 \\ -1 & s(n) < 0 \end{cases} . \tag{9}$$

(5) The energy ratio between the high frequency band (*HF*=4000-8000 Hz) and the low frequency band (*LF*=200-2000 Hz) : $y^5 \subset F^5 \subset [0,\infty)$,

$$y^5 = \frac{\sum_{\omega \in HF} |S(e^{j\omega})|^2}{\sum_{\omega \in LF} |S(e^{j\omega})|^2} \quad . \tag{10}$$

where $|S(e^{j\omega})|^2$ is the energy magnitude. $|S(e^{j\omega})|$ can be evaluated by the FFT of $s(n)$. The set of S, UP and AP had larger energy ratio, and the sets of UF1, UF2, UA and AA had smaller ratio in the consonant classes. In our experiments, we have found that the maximum energy ratio is 81.77 and minimum ratio is 0.05. Fig. 5 illustrates the mixture probability distribution curves of energy ratio for UF1, AP and S classes between different frequency bands.

(6) The energy ratio between the full band and the middle frequency band ((*MF*=2000-4000 Hz) : $y^6 \subset F^6 \subset [0,\infty)$.

$$y^6 = \frac{\sum_{\omega \in MF} |S(e^{j\omega})|^2}{\sum_{\omega \in FULL\ BAND} |S(e^{j\omega})|^2} \tag{11}$$

The set of AA had larger energy ratio, and the sets of UF2 had smaller ratio in the consonant classes. In our experiments, we have found that the maximum energy ratio is 0.662 and minimum ratio is 0.015. Fig. 6 illustrates the mixture probability distribution curves of energy ratio for UF2 and AA classes between the full band and the middle frequency bands.

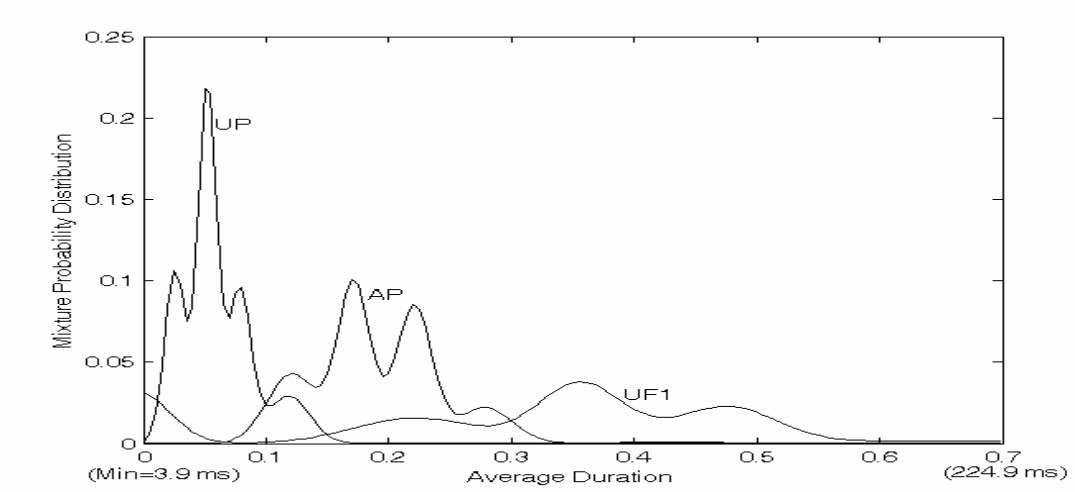

Fig. 2 The mixture Gaussian probability distribution of average duration for UP, AP and UF1.

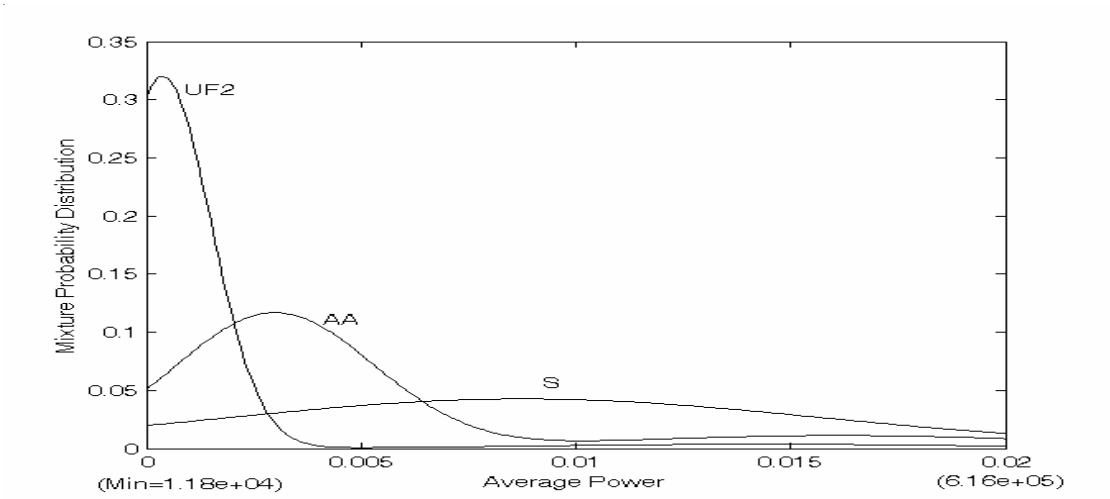

Fig. 3 The mixture Gaussian probability distribution of average power for UF2, AA and S classes.

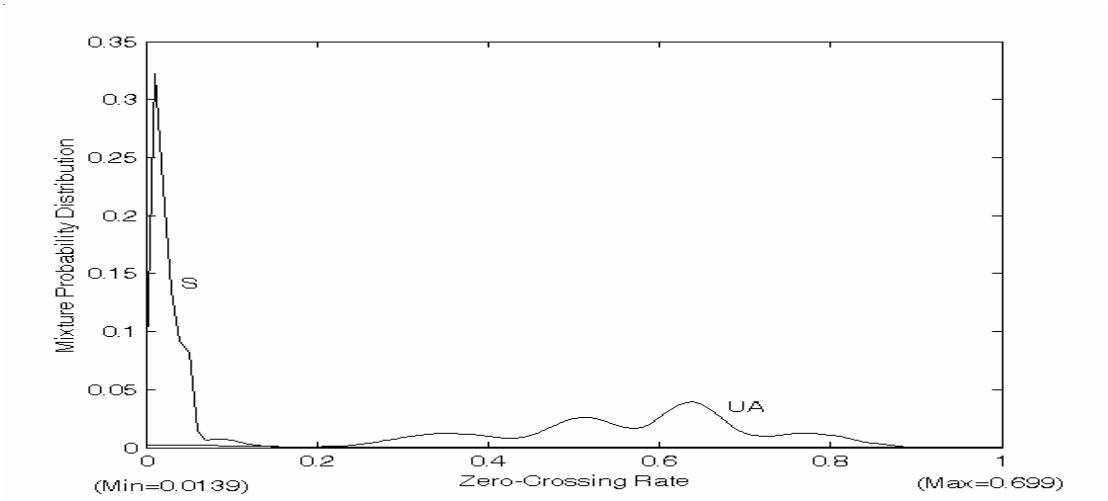

Fig. 4 The mixture Gaussian probability distribution of zero-crossing rate for S and UA classes.

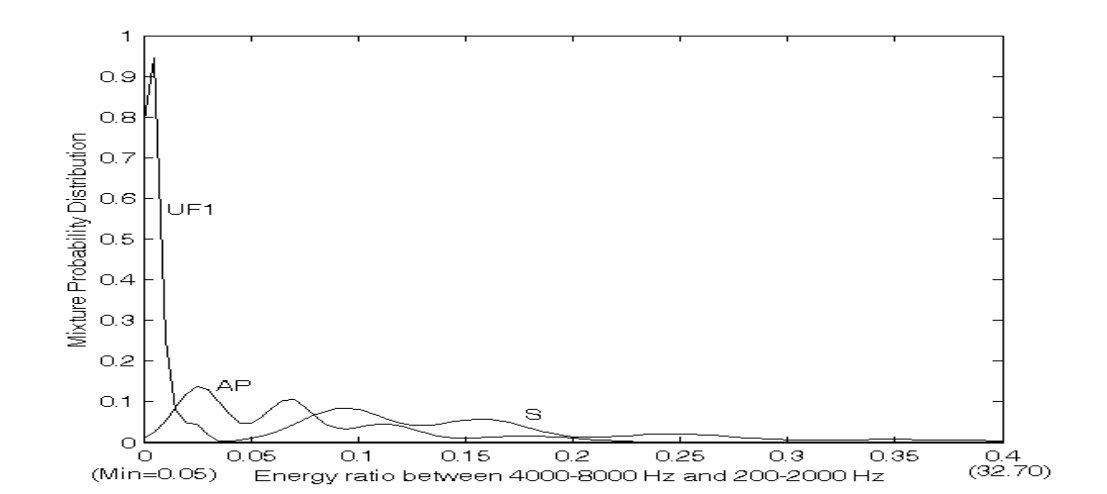

Fig. 5 The mixture Gaussian probability distribution of energy ratio for UF1, AP and S classes between 4000-8000 Hz and 200-2000 Hz.

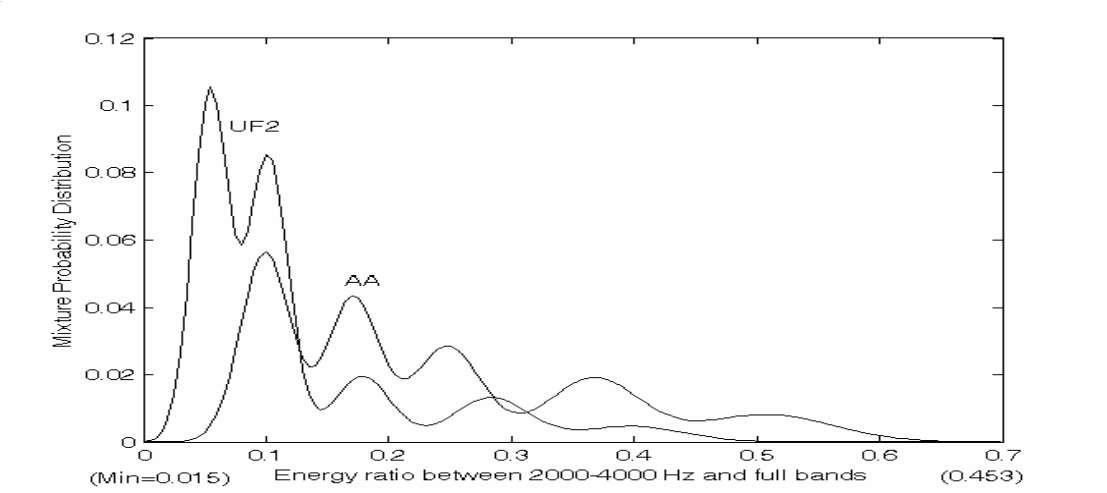

Fig. 6 The mixture Gaussian probability distribution of energy ratio for UF2 and AA classes between 2000-4000 Hz and the full band.

2.3 Decision making

The decision-making has two steps—namely, training stage of consonant speech patterns, and testing stage for classification of patterns via design rule. Fig. 7 shows the block diagram of decision rule for INITIAL classification. In the training stage, the training data set was used to construct the reference patterns by using the probability density function (pdf) for each selected feature and each class in the form of multivariate mixture Gaussian (MMG) function [6]. In the testing stage, for each consonant class, the minimum distance between input vector values and the features of trained stage in the simplified MMG was calculated, and then total distance was determined according a distance measure rule. Finally, the input data was assigned to the consonant class. The details for decision rule were described as follow:

In equation (2), it could be formed the MMG function [6, 8] as

$$p(x_i^j) = \sum_{m=1}^{M} a_{ijm} p_m(x_i^j) = \sum_{m=1}^{M} a_{ijm} \frac{1}{\sqrt{2\pi \cdot \upsilon_{ijm}}} \exp\left(\frac{-(x_i^j - \mu_{ijm})^2}{2\upsilon_{ijm}} \right). \tag{12}$$

or

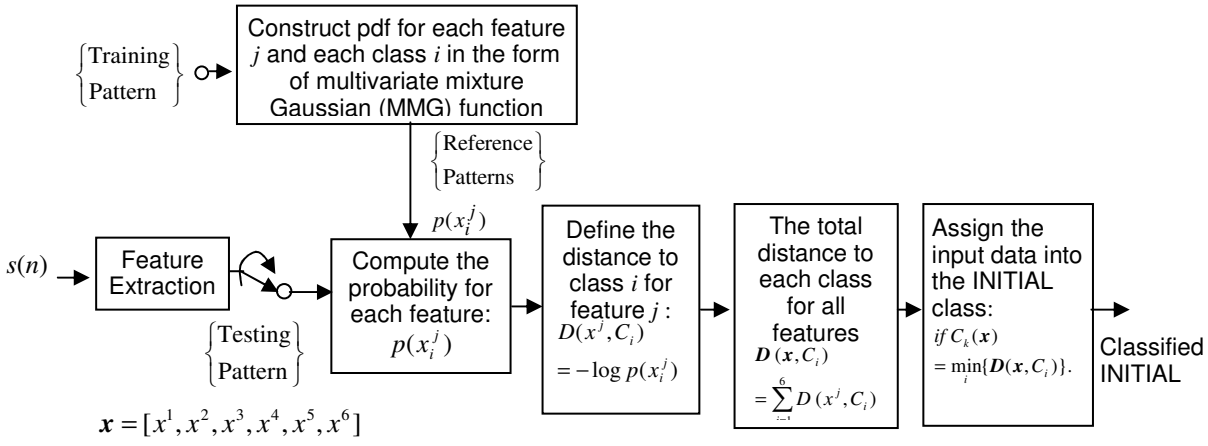

Fig 7 Block diagram of decision rule for INITIAL classification.

$$p(x_i^j) = \max_{1\le m\le M}[a_{ijm}\, p_m(x_i^j)] = \max_{1\le m\le M}\left[a_{ijm}\cdot\frac{1}{\sqrt{2\pi\cdot\upsilon_{ijm}}}\exp\left(\frac{-(x_i^j-\mu_{ijm})^2}{2\upsilon_{ijm}}\right)\right]. \quad (13)$$

where μ_{ijm} is the mean of prototype vector. υ_{ijm} is a scale parameter that is the variance in class *i* feature *j* for the *m*-th mixture Gaussian density function. In our experiment results show that the equation (13) is superior to the equation (12) by about 0.1-0.2% consonant classification rate increasing. Therefore, we use the equation (13) in our experiments.

Large numbers of products among probability coefficients may cause underflow problem. A better solution consists of using logarithmic probabilities that decrease computational load since multiplications are replaced by sums. Another reason to use logarithm causes the nonlinear mapping between small numbers to aggravate the distances of un-consistency features to eliminate the candidates. We define the distance to class *i* for feature *j* and get following the equation for the negative logarithm of $p(x_i^j)$ as:

$$D(x^j, C_i) = -\log p(x_i^j) = \min_m\left[-\log(a_{ijm}) - \log\left(\frac{1}{\sqrt{2\pi\cdot\upsilon_{ijm}}}\right) + \frac{(x_i^j-\mu_{ijm})^2}{2\upsilon_{ijm}}\right] \quad m = 1,...,M. \quad (14)$$

The total distance to each class for all features between input vector $\boldsymbol{x}$ is

$$\boldsymbol{D}(\boldsymbol{x}, C_i) = \sum_{j=1}^{6} D(x^j, C_i) = \sum_{j=1}^{6}\ \min_m\left[-\log(a_{ijm}) - \log\left(\frac{1}{\sqrt{2\pi\cdot\upsilon_{ijm}}}\right) + \frac{(x_i^j-\mu_{ijm})^2}{2\upsilon_{ijm}}\right] \quad m = 1,...,M. \quad (15)$$

Finally, the input data is assigned to the consonant class *k*, C_k:

$$if\ C_k(\boldsymbol{x}) = \min_i\{\boldsymbol{D}(\boldsymbol{x}, C_i)\}. \quad (16)$$

In the equation (14), if the variance υ becomes small, the parameters a_m do not contribute much to the overall distance. For large values of υ, the contribution of vector distance $(x_i^j - \mu_{ijm})$ can be neglected, and the mixture weights a_m become dominant.

3. Consonant Recognition in the Second Stage

In the second stage, with the selected top two consonant classes as candidates, detailed consonant recognition is performed by feature extraction of cepstral coefficients followed by the pattern comparison using the continuous hidden Markov model (CHMM) [9, 10] or the segmental probability model (SPM) [11, 12].

Shen [12] developed an approach specially for isolated Mandarin base-syllable recognition, referred to as the segmental probability model (SPM). Using SPM, comparable recognition accuracy but significantly reduce computational complexity can be obtained as compared with HMM based approach for Mandarin syllable recognition.

In SPM, each utterance of a given syllable *m* is equally divided into *N* segments (similar to the "state" in HMM), and the feature vectors in each segment *j* are modeled by an observation probability distribution function $b_j(\cdot)$, which is composed of several mixtures of Gaussian distributions. Thus the SPM for a given syllable *m* with *N* segment is denoted as

$$\lambda_m : \{b_j(\cdot),\ 1 \le j \le N\} \tag{17}$$

where $b_j(\cdot)$ is the observation probability density function (pdf) of a segment *j*.

The SPM model λ_m defines the likelihood of observing a feature vector sequence with length T, denoted as $O = o_1, o_2, ..., o_T$, for a given syllable by

$$P(O \mid \lambda_m) = \prod_{1 \le i \le N} \left\{ \prod_{o_t \in segment\ i} b_i(o_t) \right\} \tag{18}$$

where the feature vectors sequence *O* is equally divided into *N* segments.

In the recognition phase, after calculating the likelihood functions of observing an unknown feature vector sequence for all possible syllable models, the syllable k^* with maximal likelihood was chosen as the recognition output, i.e.,

$$k^* = \arg \max_{1 \le m \le L} P(O \mid \lambda_m) \tag{19}$$

where *L* is the total number of candidate of all possible syllables.

The estimation or training of the parameters for SPM is also simple. All the training utterances for the syllable *m* are first individually divided into *N* segments of equal length. The feature vectors collected from all the training utterance belonging to the same segment are then clustered into *M* mixtures using vector quantization technique with Euclidean distance measure. Each of these clusters is then modeled by a Gaussian distribution with the sample mean vector μ_{ik}, and covariance matrix Σ_{ik}. These sample mean vectors and covariance matrix, $\{\mu_{ik}, \Sigma_{ik}; 1 \le k \le M, 1 \le i \le N\}$, are then used as the SPM parameters for the syllable *m*.

SPM can be understood from an analogy to HMM. In HMM, if the state transition probabilities $[a_{ij}]$ are abandoned because of its insignificant effects on the recognition, the likelihood of observing a feature vector sequence *O* for a model λ is as follows:

$$P(O \mid \lambda) = \max_s \prod_{t=1}^{T} b_{st}(o_t) \tag{20}$$

where $b_{st}(o_t)$ is the output probability density function for the feature vector o_t at segment s_t. The maximization is performed over all possible state sequence *S*. In SPM, instead of searching for the optimal state sequence as in HMM, a state sequence was simply specified deterministically by the segments of equal length.

The feature selection in this stage is consideration. The Mel-frequency cepstrum coefficients (MFCC) are employed for the feature selection in this stage. MFCC obtained by applying a discrete cosine transform on the log subband energies of the spectrum, obtained from a Mel-frequency scale. The MFCC were computed as Davis and mermelstein [13]:

$$MFCC(m) = \sum_{k=1}^{K} Y_k \cos\left[m(k - 1/2)\frac{\pi}{K} \right],\ m = 1,2,...L, \tag{21}$$

Here, the notation Y_k denotes the sum of the log energy within the k-th critical band, and K is the number of critical band and L is the number of MFCC coefficients. In this paper, we used 24 (K=24) triangular filters and 12 (L=12) MFCC coefficients. The procedure has great advantages. Meanwhile, its use has been show empirically to improve recognition accuracy [14].

4. Experiments

The Computer & Communication Research Laboratories of the Industrial Technology Research Institute, ROC, provide the speech database to use in this research. All the recorded materials are obtained in an office-like laboratory environment without special soundproof treatment. These speech samples are chosen at 16 kHz at 16-bit precision. The database was recorded by four male speakers. Each speaker uttered in the 1345 Mandarin tone-syllables (the 408 base syllables and all possible tone variations give 1345 tone-syllables). There are totally 4186 consonants in our experiment database. The consonant utterances are obtained by the INITIAL/FINAL automatic segmentation algorithm proposed in this paper from the previous work [15].

In the consonant classification of first stage, the experiment compared classification results are listed in Table 2. The correct percentage of consonant classification is achieved 93.1% for the first selection and 97.9% for the top two selections. According Table 2, we find that the classified error happened at the classes of similar pronounce place or manner. For example, UP incorrectly classified into AP, AP into UP, UA into AA or UF1, AA into UA or UF1, UF1 into AA or UF2, UF2 into UF1, and S into UP. Our multi-speaker experiments showed that a high correct percentage (97.9% for the top two selections) could be achieved in our first-stage consonant classification. Therefore, the selected top two consonant classes are employed as candidates for consonant recognition of the second stage, and the search procedure of SPM/CHMM reduced to 5-7 models instead of 21 models of full search.

In second stage in speech signal processing, the analysis frame size is 20 ms (320 samples) and is overlapped by a 10-ms duration. In our recognizer, each utterance is preemphasized with a filter $H(z)=1-0.95z^{-1}$. A Hamming window of 320 points is applied to each frame. Then, 12 Mel-frequency cepstrum coefficients (MFCC) of each frame are calculated and retained to form a basic element acoustic vector. By using the K-mean algorithm, MFCC vectors of each segment are clustered into appropriate partitions, and each is corresponding to a mixture in the CHMM/SPM.

The first experiment is to compare the recognition rate with the two-stage algorithm and one-stage algorithm (omitted first stage). The four set utterances of four male speakers are taken for training the models and testing the results.

Table 3 shows the comparison of consonant classification accuracy rate (first stage) and recognition rate (first stage + second stage by using either the SPM or the CHMM) for top n candidate selection. We find that the top 2 candidates of consonant class have better performance than other selections. Thus, we select top 2 candidates of consonant class in the next experiments.

Table 4 shows the comparison of consonant recognition rate and the relative processing time. The experimental results show that both SPM and CHMM can achieve comparable accuracy percentage of 89.3%, but the SPM requires only about 1/8 processing time. In addition, the proposed two-stage algorithm is superior to the one-stage algorithm (omitting the first stage) by about 3% accuracy rate increasing and 12% processing time saving (since the 21 recognition units is reduced 5-7 units in 2nd stage). These two-stage results strongly supported the concept and consideration in the previous section is highly acceptable for our proposed system.

The results of our investigation show that using only a single pattern classifier with limited features is nearly impossible to achieve a recognition rate as high as desired in our Mandarin consonant recognizer. Fig. 8 illustrated the classification rates of the unaspirate plosives using various

feature sets; we find that six features can achieve the high classified rate than only one feature. Based on the phonetic knowledge of Mandarin consonant, the sonorant (S) class is the voiced speech that can be detected and classified by the pitch period feature. Fig. 9 illustrated the classification rates of the sonorants using various feature sets. We can find that only use pitch period feature for sonorant class can achieve higher classification rate than other consonant classes.

Table 2. Confusion matrix for the accuracy rate of consonant classification.

		Detected as							Accuracy Rate (Top1/Top2) %
		UP (626)	AP (666)	UA (540)	AA (622)	UF1 (588)	UF2 (279)	S (865)	
ㄅ/b/,ㄉ/d/,ㄍ/g/	UP	**595**	20	1	2	-	-	8	95.2/98.4
ㄆ/p/,ㄊ/t/,ㄎ/k/	AP	29	**631**	3	-	1	2	-	94.8/98.1
ㄓ/j/,ㄗ/tz/,ㄐ/j(i)/	UA	1	1	**498**	22	13	5	-	92.4/98.3
ㄔ/ch/,ㄘ/ts/,ㄑ/ch(i)/	AA	-	1	31	**572**	16	1	1	92.0/97.7
ㄕ/sh/,ㄙ/s/,ㄒ/sh(i)	UF1	-	-	5	15	**553**	15	-	94.2/98.5
ㄈ/f/, ㄏ/h/	UF2	2	3	3	5	10	**254**	2	91.2/94.7
ㄇ/m/,ㄋ/n/,ㄌ/l/,ㄖ/r/	S	17	13	8	7	7	6	**807**	93.4/98.7
Average									93.1/97.9

* (•) denoted the number of consonant class.

Table 3. The comparison of accuracy rate of consonant classification (first stage) and recognition rate (first stage + second stage by using either the SPM or the CHMM) for top n candidate selection.

Ton n consonant class		1	2	3	4	5	6
Consonant classification rate % (1st stage)		93.1	97.9	99.8	100	100	100
Consonant recognition rate %	(1st stage + 2nd stage by using SPM)	85.1	**89.2**	89.0	88.3	87.5	86.4
	(1st stage + 2nd stage by using CHMM)	85.0	**89.3**	89.1	88.3	87.6	86.5

Table 4. Consonant recognition rates and relative processing time.

Methods	Two-stage algorithm (Use Top 2 class candidates)		One-stage algorithm	
	SPM in the 2nd stage	CHMM in the 2nd stage	SPM	CHMM
Recognition Rates % (Top1/Top2)	**89.2/96.4**	**89.3/96.5**	86.3/94.2	86.2/94.1
Relative Processing Time	**1**	**8.20**	1.12	9.57

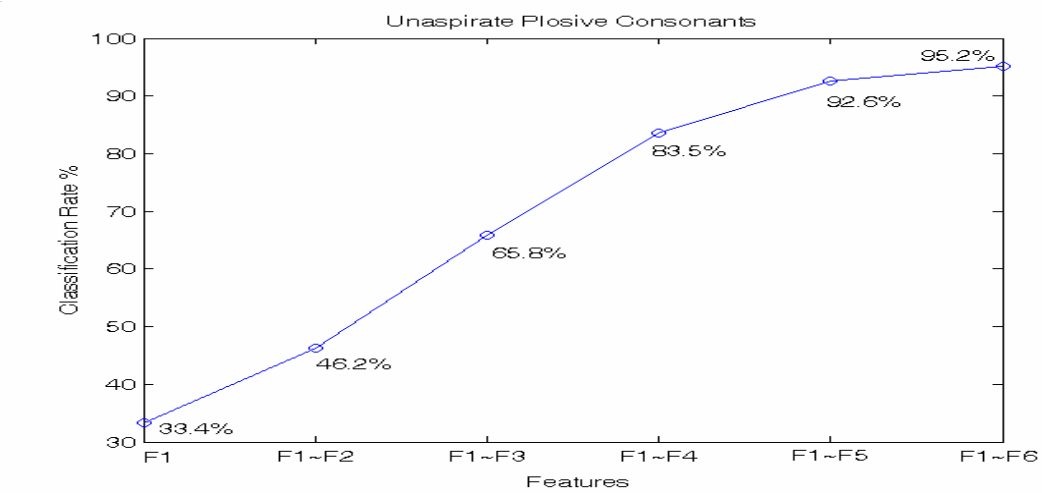

Fig. 8 The classification rates of the unaspirate plosives using various feature sets.

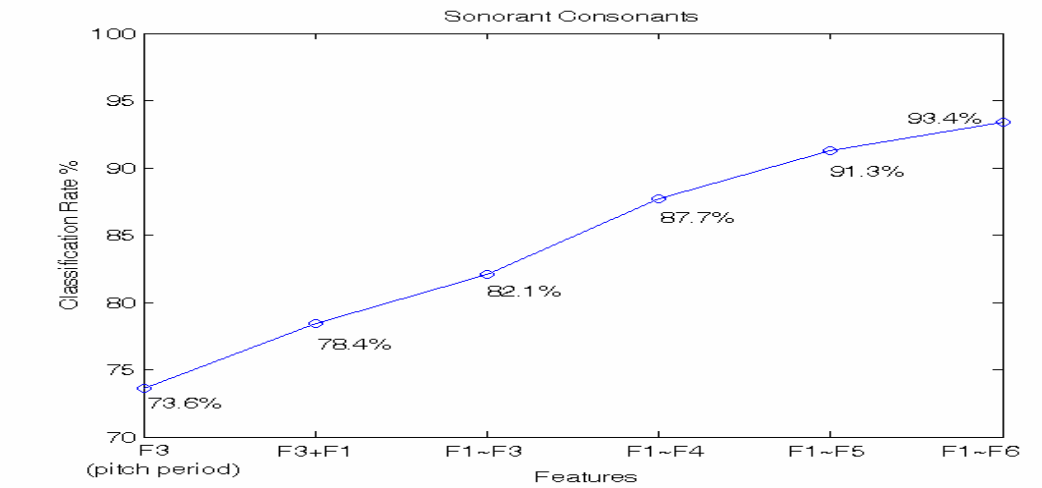

Fig. 9 The classification rates of the sonorants using various feature sets.

Table 5 shows the confusion matrix of INITIALs for the two-stage scheme with SPM for the 2nd stage. We find that the error recognition of the INITIALs always happens at the same cluster. For example, the ㄓ/j/ incorrectly recognized to ㄗ/tz/, ㄔ/ch/ to ㄗ/tz/ or ㄘ/ts/, ㄉ/d/ to ㄅ/b/, ㄆ/p/ to ㄊ/t/ or ㄅ/b/, ㄍ/g/ to ㄅ/b/, ㄒ/sh(i)/ to ㄕ/sh/ or ㄙ/s/, and ㄌ/l/ to ㄇ/m/ or ㄋ/n/. Because those INITIALs of the same cluster have similar specialties and features in speech production, therefore, the error recognition is always occurred at the same cluster.

5. Conclusions

For consonant recognition, a two-stage algorithm with multiple features was proposed. We classified Mandarin consonant into seven classes, each with similar phonetic and acoustic characteristics. In the first stage of the proposed algorithm, consonant classification was conducted by computing features to construct mixture density models for each feature and for each consonant class, and then we defined for the decision making of consonant classification. Our multi-speaker experiments showed that a high correct percentage (97.9% for the top two selections) could be achieved in our first-stage consonant classification. In the second stage, with the selected top two consonant classes as candidates, detailed consonant recognition is performed by feature extraction of

cepstral coefficients followed by the pattern comparison using the continuous hidden Markov model (CHMM) or the segmental probability model (SPM). The experimental results showed that both SPM and CHMM can achieve comparable accuracy percentage of 89.2%, but the SPM requires only about 1/8 processing time. Furthermore, the proposed two-stage algorithm is superior to the one-stage algorithm (omitting the first stage) by about 3% accuracy rate increasing and 12% processing time saving.

Table 5 Confusion matrix of INITIAL recognition by using two-stage scheme with SPM for the 2nd stage.

Result / Test			**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	%
			j	ch	sh	r	tz	ts	s	g	k	h	ji	chi	shi	d	t	n	l	b	p	m	f	
			ㄓ	ㄔ	ㄕ	ㄖ	ㄗ	ㄘ	ㄙ	ㄍ	ㄎ	ㄏ	ㄐ	ㄑ	ㄒ	ㄉ	ㄊ	ㄋ	ㄌ	ㄅ	ㄆ	ㄇ	ㄈ	
2	j	ㄓ	**201**	3	5		14	1	2				5	1	2								1	85.9
3	ch	ㄔ	17	**196**			9	25	5							1	1		1					76.9
4	sh	ㄕ	9	3	**165**		6	3	20				1	1	2									78.6
5	r	ㄖ		1		**99**				1								2	7	5	1	2		83.9
6	tz	ㄗ	10	3	4		**155**	2	8														1	84.7
7	ts	ㄘ	5	19			5	**145**	2															88.1
8	s	ㄙ	6	2	8		5	2	**123**															84.3
9	g	ㄍ								**189**	1					1				2	1		1	97.0
10	k	ㄎ								2	**180**	4								1	1		1	95.3
11	h	ㄏ									2	**150**					1		1		3	1	1	94.3
12	ji	ㄐ	1						1				**158**	8	1		1							92.9
13	chi	ㄑ							1				7	**182**	1									95.3
14	shi	ㄒ	1		3				1				13	6	**159**	1								86.4
15	d	ㄉ					1	1		22	1	2				**170**	3		1	12	1		12	75.3
16	t	ㄊ		1		1				3	2	3					**231**		1	4	5		1	91.7
17	n	ㄋ				1												**206**	5	3		14	1	89.6
18	l	ㄌ				3											1	7	**260**	4	2	18		88.1
19	b	ㄅ								8						1	1		1	**184**	1		9	89.8
20	p	ㄆ								1							5				**217**		2	96.5
21	m	ㄇ																2	1	6		**212**	1	95.5
22	f	ㄈ								2						1	1			13	1		**102**	85.0

Although the techniques are proposed specially for the consonant recognition of Mandarin syllable, it is definitely believed that the concepts are potentially applicable to solve similar problem in FINAL recognition of Mandarin syllables or other language. We believe that the proposed algorithm and the knowledge gained from this research not only can be used in Mandarin speech recognition systems, but also are valuable for other related applications.

References

[1] L. S. Lee, "Voice dictation of Mandarin Chinese," *IEEE Signal Processing Magazine,* vol. 14, no. 4, pp. 63-101, July, 1997.

[2] J. K. Chen and F. K. Soong, "An N-best candidates-based discriminative training for speech recognition application," *IEEE Trans. Speech and Audio Speech*, vol. 2, pp. 206-216, 1994.

[3] L. C. Liu and H. C. Wang, "A study on the recognition of Mandarin consonants," Ph. D. dissertation, Univ. of Ching-Hwa, 1990.

[4] G. S. Poo and Y. Ou, "A large phonemic time-delay neural network technique for all Mandarin consonants recognition," in *Proc. Int. Conf. Theme, Frontiers of Computer Technology*, 1994, pp. 521-525.

[5] Y. Lee, L. S. Lee and C. Y. Tseng, "Isolates Mandarin syllable recognition with limited training data specially considering the effect of tones," *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 1, pp. 75-80, Jan., 1997.

[6] B. H. Juang and L. R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Trans. Acoustic, Speech and Signal Processing,* vol. 33, no. 6, pp. 1404-1413, Dec., 1985.

[7] L. R. Rabiner, J. G. Wilpon, & B. H. Juang, "A segment K-means algorithm training procedure for connected word recognition," *AT&T Technique Journal*, vol. 65, no. 3, pp. 21-32, 1986.

[8] H. Ney and A. Noll, "Acoustic-phonetic modeling in the SPICOS system," *IEEE Trans. Speech and Audio Speech*, vol. 2, pp. 312-319, 1994.

[9] L. R. Rabiner and B. H. Juang, "An introduction to hidden Markov models," *IEEE ASSP Magazine*, pp. 4-15, Jan. 1986.

[10] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition,* NJ: Printice Hall, 1993.

[11] R. Y. Lyu, I. C. Hong, J. L. Shen, M. Y. Lee and L. S. Lee, "Isolated Mandarin base-syllable recognition based upon the segmental probability model," *IEEE Trans. Speech and Audio Speech*, vol. 6, no. 3, pp. 293-299, May, 1998.

[12] J. L. Shen, "Continuous Mandarin speech recognition for Chinese language with large vocabulary based on segmental probability model," *IEE Proc. Vision, Image Processing,* vol. 145, no. 5, pp. 309-315, Oct., 1998.

[13] S. D. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustic, Speech and Signal Processing*, vol. 28, no. 4, pp. 357-366, Aug., 1980.

[14] S. Young, "A review of large-vocabulary continuous-speech recognition," *IEEE Signal Processing Magazine*, vol. 13, no. 5, pp. 45-57, Sept., 1996.

[15] M. T. Lin, C. K. Lee and C. Y. Lin, "Consonant/vowel segmentation for Mandarin syllable recognition," *Computer Speech and Language*, vo1. 13, no. 3, pp. 207-222, July, 1999.