J. of Computer Science and Technology, 13(5): 426-437, Sept., 1998

Center-Distance Continuous Probability Models And the Distance Measure

Zheng Fang (郑 方), Wu Wenhu (吴文虎), and Fang Ditang (方棣棠)

Speech Lab., Department of Computer Science and Technology, Tsinghua University
Beijing, 100084, P.R.China
E-mail: fzheng@sp.cs.tsinghua.edu.cn
Received March 3, 1997; revised June 11, 1997

Abstract: In this paper, a new statistic model named Center-Distance Continuous Probability Model (CDCPM) for speech recognition is described, which is based on Center-Distance Normal (CDN) distribution. In a CDCPM, the probability transition matrix is omitted, and the observation probability density functions (PDF) in each state are in form of embedded multiple-model (EMM) based on the Nearest Neighbour rule. The experimental results across two giant real-world Chinese speech databases and a real-world continuous-manner 2000 phrase system show that this model is a powerful one. Also, a distance measure for CDCPMs is proposed which is based on the Bayesian minimum classification error (MCE) discrimination.

Keywords: Center-Distance Continuous Probability Model (CDCPM), Center-Distance Normal (CDN) Distribution, Embedded Multiple-Model (EMM) scheme, Minimum Classification Error (MCE)

1. Introduction

Dominant models in speech recognition (SR) are HMMs, including continuous mixture density HMMs [1-3] with full covariance matrices or diagonal covariance matrices, semi-continuous HMMs [4], and VQ-based discrete HMMs [5].

A Continuous HMM (CHMM) is represented by the state transition probability matrix A, the observation probability density function matrix B and the initial probability distribution vector π. Many algorithms are developed to estimate these HMM parameters, such as Baum-Welch [6], EM (Expectation and Maximization) [7], MMIE (Maximum Mutual Information Estimation) [8], and MAP (Maximum a Posterior) [9]. Also many algorithms are developed for recognition, such as Viterbi algorithm [10] and Frame Synchronous Search algorithm [11].

According to our researches, we found that the state transition probability matrix in a HMM is not very important [12-17], studies on the distance measure between acoustic models also support this conclusion [18]. This motivates us to propose a new model without the state transition probability matrix. In this paper, such a model named center-distance continuous probability model (CDCPM) is presented.

Acoustic model distance measure plays a very important role because of the following reasons:

(1) Discriminating ability: establishing acoustic models using different features and calculating the model distance matrix for each kind of features, we can find the best features according to the average model distance for each kind of features. Obviously, the bigger the average model distance is, the bigger discriminating ability the features have.

(2) Grouping the models: according to the model distance matrix and a pre-defined distance threshold, we can group the models. In each group, the model distance between every two models is very small, and the models may be difficult to distinct, which guides us to pay more attention to these intra-group models.

(3) Verifying the model generalization ability: calculating the model distance matrix for both the training and testing database, if the average distance in the distance matrix for testing database does not reduce very much, we can conclude that this kind of models for such features has a relatively good generalization ability and will be robust to other testing data.

So meanwhile a distance measure for CDCPMs based on the Bayesian minimal classification error (MCE) discrimination is proposed.

2. The CDN Distribution and Its Distance Measure

2.1 The center-distance normal distribution

The PDF of a normal random variable ξ with mean value μ_x and standard deviation σ_x is

$$N(x;\mu_x,\sigma_x)=\frac{1}{\sqrt{2\pi}\,\sigma_x}\exp\left\{-(x-\mu_x)^2/2\sigma_x^2\right\},\ x\in(-\infty,\infty), \tag{1}$$

Define a new random variable $\eta=\left|\xi-\mu_x\right|$, then the PDF of η is

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

In fact, η is the distance between a normal variable ξ and its mean value μ_x. That's why the defined distribution is referred to as a Center-Distance Normal (CDN) distribution.

By calculating the mean value μ_y of CDN variable η in Equ. (2), we have

$$\mu_y=\sqrt{\frac{2}{\pi}}\,\sigma_x \quad \text{or} \quad \sigma_x=\sqrt{\frac{\pi}{2}}\,\mu_y \tag{3}$$

Substituting Equ. (3) into (2), the PDF can be rewritten as

$$p(y;\mu_y)=\frac{2}{\pi\mu_y}\exp(-y^2/\pi\mu_y^2),\ y\geq 0. \tag{4}$$

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 $\eta=y(\vec{\xi},\vec{\mu}_x)$, where $y(\cdot,\cdot)$ is a kind of (weighted) Euclidean distance measure. Assume η is a CDN variable, then its CDN pseudo-PDF (**PPDF**) is

$$N_{CD}(\vec{x};\vec{\mu}_x,\mu_y)=\frac{2}{\pi\mu_y}\exp\left\{-y^2(\vec{x},\vec{\mu}_x)/\pi\mu_y^2\right\}. \tag{5}$$

Strictly speaking, $N_{CD}(\vec{x};\vec{\mu}_x,\mu_y)$ is the PDF of $y(\vec{\xi},\vec{\mu}_x)$ instead of that of $\vec{\xi}$, it is just for convenience and comparison. For simplification, Equ. (5) is called a CDN PPDF while Equ. (4) a CDN PDF.

2.2 The distance measure for two CDN distributions

Consider the CDN PDF described in Equ. (4), by shifting this function by *d* along *y* axis we have

$$p^{(d)}(y;\mu_y)\overset{def}{=}p(y-d;\mu_y)=\frac{2}{\pi\mu_y}\exp\left\{-(y-d)^2/\pi\mu_y^2\right\},\ y\in[d,\infty) \tag{6}$$

By unfolding the CDN PDF function, we get the corresponding normal PDF as follows

$$p_0^{(d)}(y;\mu_y)\overset{def}{=}p_0(y-d,\mu_y)=\frac{1}{\pi\mu_y}\exp\left\{-(y-d)^2/\pi\mu_y^2\right\},\ y\in(-\infty,\infty) \tag{7}$$

where

$$p_0(y,\mu_y)=\frac{1}{\pi\mu_y}\exp\left(-y^2/\pi\mu_y^2\right),\ y\in(-\infty,\infty) \tag{1'}$$

is the normal PDF corresponding to Equ. (4). And also we have

$$\int_{-\infty}^{y_0}p_0^{(d)}(y;\mu_y)dy=\Phi\left(\sqrt{\frac{2}{\pi}}\frac{y_0-d}{\mu_y}\right), \tag{8}$$

where

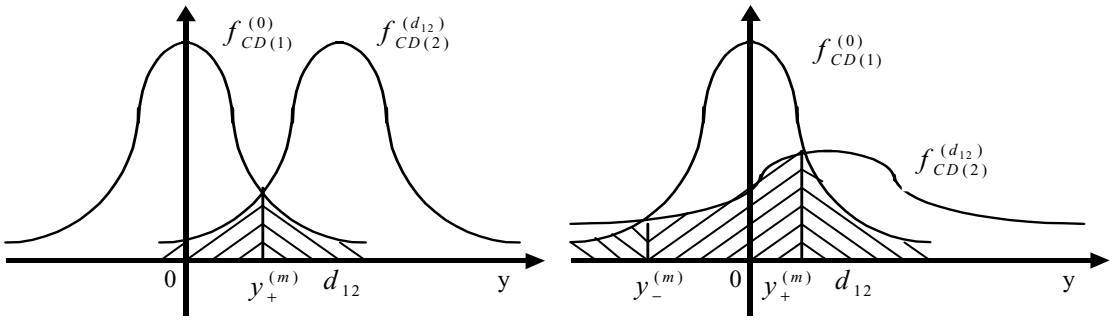
$$\Phi(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\exp(-x^2/2)dx \tag{9}$$

is the probability distribution function of a standard normal distribution. See Appendix for some properties of function $\Phi(x)$. Hereafter we simply denote

$$N_{CD(k)}(\vec{x})=N_{CD}(\vec{x};\vec{\mu}_{xk},\mu_{yk}) \tag{10}$$

$$p_{0(k)}^{(d)}(y)=p_0^{(d)}(y;\mu_{yk})=p_0(y-d;\mu_{yk}) \tag{11}$$

Now we are ready to define the distance measure between two CDN PPDFs $N_{CD(1)}$ and $N_{CD(2)}$. Denote the distance between the mean vectors of two CDN PPDFs by $d_{12}=y(\vec{\mu}_{x1},\vec{\mu}_{x2})$ and denote the ratio of the mean center-distances of two CDN PPDFs by $R_{21}=\mu_{y2}/\mu_{y1}$, and also assume $R_{21}\geq 1$. In order to derive the distance measure, we change the CDN distributions to their corresponding shifted normal PDFs $p_{0(1)}^{(0)}(y)$ and $p_{0(2)}^{(d_{12})}(y)$ (same result can be got by shifting them to $p_{0(1)}^{(d_{12})}(y)$ and $p_{0(2)}^{(0)}(y)$ respectively).



a. $\mu_{y1}=\mu_{y2}$ b. $\mu_{y1}<\mu_{y2}$

Fig.1 The (shifted) normal PDFs corresponding to the CDN PPDFs

By solving the equation

$$p_{0(1)}^{(0)}(y)=p_{0(2)}^{(d_{12})}(y) \tag{12}$$

we can get the root(s)

$$y_{12}^{(m)}=\begin{cases}d_{12}/2, & \mu_{y1}=\mu_{y2}\\ \dfrac{-d_{12}\pm\sqrt{d_{12}^2+\left(R_{21}^2-1\right)\left(d_{12}^2+\pi\mu_{y2}^2\ln R_{21}\right)}}{R_{21}^2-1}, & \mu_{y1}<\mu_{y2}\end{cases} \tag{13}$$

if $\mu_{y2}=\mu_{y1}$, Equ. (12) has only one root, denoted by $y_+^{(m)}=d_{12}/2$; if $\mu_{y2}>\mu_{y1}$, Equ. (12) will have two roots, one positive and one negative, denoted by $y_+^{(m)}$ and $y_-^{(m)}$, respectively, as shown in Fig. 1.

Because CDN distribution is derived from the normal distribution, it is reasonable to change one CDN PPDF back to one-dimensional normal distribution with zero mean and change another one back to one-dimensional normal distribution with mean value d_{12}. The distance between two mean vectors of the original CDN PPDFs is exactly d_{12}. The shadowed area showed in Fig. 1 stands for the classification error according the Bayesian minimum classification error (MCE) criterion. So the distance between two CDN PPDFs should be related to the shadowed area. Hence we define the distance $D(N_{CD(1)},N_{CD(2)})$ between $N_{CD(1)}$ and $N_{CD(2)}$ as "1 – Area (Samples of Class 1 classified to Class 2) – Area (Samples of Class 2 classified to Class 1)".

When Equ. (12) has one root, as in Fig. 1 (a), the distance between two CDN pseudo-distributions is

$$\begin{aligned} D(N_{CD(1)},N_{CD(2)}) &\overset{def}{=} 1-\left[\int_{-\infty}^{y_+^{(m)}} f_{CD(2)}^{(d_{12})}(y)dy+\int_{y_+^{(m)}}^{\infty} f_{CD(1)}^{(0)}(y)dy\right] \\ &= \Phi\left(\sqrt{\frac{2}{\pi}}\frac{y_+^{(m)}}{\mu_{y1}}\right)+\Phi\left(\sqrt{\frac{2}{\pi}}\frac{d_{12}-y_+^{(m)}}{\mu_{y2}}\right)-1 \\ &= \Phi\left(\sqrt{\frac{2}{\pi}}\frac{d_{12}/2}{\mu_{y1}}\right)+\Phi\left(\sqrt{\frac{2}{\pi}}\frac{d_{12}/2}{\mu_{y2}}\right)-1 \end{aligned} \tag{14-1}$$

and when Equ. (12) has two roots, as in Fig. 1 (b), according to the $\Phi(x)$ property $\Phi(x)+\Phi(-x)=1$, the distance between two CDN pseudo-distributions can be written for both $y_+^{(m)} \geq d_{12}$ and $y_+^{(m)} < d_{12}$ as follows

$$\begin{aligned} D(N_{CD(1)},N_{CD(2)}) = &\Phi\left(\sqrt{\frac{2}{\pi}}\frac{y_+^{(m)}-0}{\mu_{y1}}\right)+\Phi\left(\sqrt{\frac{2}{\pi}}\frac{0-y_-^{(m)}}{\mu_{y1}}\right) \\ &-\Phi\left(\sqrt{\frac{2}{\pi}}\frac{y_+^{(m)}-d_{12}}{\mu_{y2}}\right)-\Phi\left(\sqrt{\frac{2}{\pi}}\frac{d_{12}-y_-^{(m)}}{\mu_{y2}}\right) \end{aligned} \tag{14-2}$$

As a matter of fact, if $y_-^{(m)} << 0$ the shadowed area to the left of $y_-^{(m)}$ can be ignored. In this situation, the distance measure can be simplified to

$$D(N_{CD(1)},N_{CD(2)}) = \Phi\left(\sqrt{\frac{2}{\pi}}\frac{y_+^{(m)}}{\mu_{y1}}\right)+\Phi\left(\sqrt{\frac{2}{\pi}}\frac{d_{12}-y_+^{(m)}}{\mu_{y2}}\right)-1 \tag{14-3}$$

The CDN distance measure has the following properties [17]:

1. Non-negative: $D(N_{CD(1)},N_{CD(2)}) \in [0,1), \quad and \; D(N_{CD(1)},N_{CD(1)}) = 0$ (15)
2. Symmetric: $D(N_{CD(1)},N_{CD(2)}) = D(N_{CD(2)},N_{CD(1)})$ (16)
3. Triangular: $D(N_{CD(1)},N_{CD(2)}) + D(N_{CD(2)},N_{CD(3)}) \geq D(N_{CD(1)},N_{CD(3)})$, if $\mu_{y1}=\mu_{y2}=\mu_{y3}$ (17)
4. Marginal: $D(N_{CD(1)},N_{CD(2)}) \to 1, \quad if \quad d_{12} \to \infty$ (18)

Practically, with a look-up table for $\Phi(x)$, the distance between two CDN PPDFs can be calculated easily using Equ.s (13) and (14).

3. The CDCPM and Its Distance Measure

3.1 The center-distance continuous probability model

In a continuous hidden Markov model (CHMM) with mixed Gaussian densities (MGD) [2] in each state, Baum-Welch [6], Viterbi [10, 19] algorithms and many better improvement versions have been available for CHMM 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 $B = (b_j(\vec{\boldsymbol{x}}))_N$, the probability density of the CHMM Λ generating $\mathbf{O}$ is

$$f\{\mathbf{O}|\Lambda\} = \sum_S f\{\mathbf{O}, S|\Lambda\} = \sum_S \Pr\{S|\Lambda\} \cdot f\{\mathbf{O}|\Lambda, S\} = \sum_S \left(\pi_{s_1} \prod_{t=2}^{T} a_{s_{t-1}s_t} \right) \cdot \left(\prod_{t=1}^{T} b_{s_t}(\vec{\mathbf{o}}_t) \right)$$
$$= \sum_S \left(\pi_{s_1} \cdot b_{s_1}(\vec{\mathbf{o}}_1) \prod_{t=2}^{T} a_{s_{t-1}s_t} \cdot b_{s_t}(\vec{\mathbf{o}}_t) \right) \tag{19}$$

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. (19)

$$Score\{\mathbf{O}|\Lambda\} = f\{\mathbf{O}, S^{(ML)}|\Lambda\} = f\{\mathbf{O}|\Lambda, S^{(ML)}\} \cdot \Pr\{S^{(ML)}|\Lambda\} \tag{20}$$

CHMMs can describe signals very well, but the estimation of model parameters will cost too much time. Researches on model distance measures show that the A matrix contributes not too much as B does to the recognition performance [18], so we ignore the A matrix and then Equ. (20) is changed to

$$Score\{\mathbf{O}|\Lambda\} = f\{\mathbf{O}|\Lambda, S^{(ML)}\} = \prod_{t=1}^{T} b_{s_t^{(ML)}}(\vec{\mathbf{o}}_t) \tag{21}$$

According to the above discussion, a model named center-distance continuous probability model (CDCPM) based on CDN distributions is proposed [15].

In a CDCPM with embedded multiple-model (EMM) scheme [16], the following parameter should be determined: (1) N: number of states; (2) M: number of CDN densities each state; (3) D: number of dimensions each feature vector; (4) $\vec{\mu}_{xnm} = (\mu_{xd}^{(nm)})$: mean vector of the m-th density component in n-th state; and (5) μ_{ynm}: mean center-distance of the m-th density component in n-th state. Where $1 \le n \le N$, $1 \le m \le M$, $1 \le d \le D$, and the observation scoring function has the following form (EMM scheme)

$$b_n(\vec{\boldsymbol{x}}) = \max_{1 \le m \le M} N_{CD}(\vec{\boldsymbol{x}}; \vec{\mu}_{xnm}, \mu_{ynm}). \tag{22}$$

Thus a CDCPM can be denoted by $\Lambda = \{N_{CD(nm)} | 1 \le n \le N, 1 \le m \le M\}$, where $N_{CD(k)}$ is as in (10) and (5).

3.2 Training CDCPMs

Once N, M, and D have been determined, the training procedure is as simple as in the following:

(1) Each observation feature sequence **O** from the training database is first segmented into N segments (corresponding to N states) using some segmentation method such as the Non - Linear Segmentation (NLS) method [12] (see Appendix for more details).

(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 LBG algorithm [20].

(3) Estimation of $\vec{\mu}_{xnm}$ and μ_{ynm} is very easy for each density, namely each class, for the specified segment *n*.

3.3 The modified Viterbi decoding algorithm for CDCPMs

There is no state transition probability matrix, so what controls the state transition in a CDCPM during continuous speech recognition procedure ? Experiments have shown that NLS algorithm is efficient for providing a sequence segmentation both for training procedure and isolated word recognizing procedure [12]. As for the continuous speech recognition, we should modify the Viterbi decoding algorithm.

By assigning a constant to all the elements of the *A* matrix in a CHMM, we have the following modified Viterbi decoding algorithm for the CDCPM.

Step 1: Initialization

$$\Phi_1(j) = b_j(\vec{\mathbf{o}}_1), \quad 1 \le j \le N \tag{23}$$

Step 2: Forward-Searching (recursively for $2 \le t \le T, 1 \le j \le N$)

$$\Phi_t(j) = \max_{1 \le i \le N} \Phi_{t-1}(i) \cdot b_j(\vec{\mathbf{o}}_t) \tag{24}$$

$$\Psi_t(j) = \arg\max_{1 \le i \le N} \Phi_{t-1}(i) \cdot b_j(\vec{\mathbf{o}}_t) \tag{25}$$

Step 3: Back-tracking

$$s_T^{(ML)} = \arg\max_{1 \le i \le N} \Phi_T(i) \cdot 1 \tag{26}$$

$$s_t^{(ML)} = \Psi_{t+1}(s_{t+1}^{(ML)}), \quad 1 \le t \le T-1 \tag{27}$$

After getting the ML state sequence, the matching score can be got easily using Equ. (21).

Furthermore, the frame synchronous search algorithm [11] can also be used to decode the state sequence.

3.4 The distance measure for two CDCPMs

The distance measure between two CDCPMs is based on the distance measure for the CDN distribution. For CDCPMs with EMM scheme, i.e., using Equ. (22), let

$$\Lambda_k = \left\{ N_{CD(knm)} \middle| 1 \le n \le N, 1 \le m \le M \right\}, \quad k = 1,2 \tag{28}$$

denote two CDCPMs, and define

$$D(\Lambda_1, \Lambda_2) = \frac{1}{N} \sum_{n=1}^{N} \left[\min_{1 \le p,q \le M} D(N_{CD(1np)}, N_{CD(2nq)}) \right] \tag{29}$$

as the CDCPM model distance measure.

Obviously, the distance measure for CDCPMs has the following properties:

1. Non-negative: $D(\Lambda_1, \Lambda_2) \in [0,1), \quad and \ D(\Lambda_1, \Lambda_1) = 0$ (30)
2. Symmetric: $D(\Lambda_1, \Lambda_2) = D(\Lambda_2, \Lambda_1)$ (31)

4. Experimental Data

In order to evaluate the efficiency of the CDCPM, several groups of experiments are done across two real-world giant Chinese speech database and a continuous-manner 2000-phrase system is built up.

4.1 The database descriptions

There are two different databases used here: DBI and DBII.

DBI is a giant Chinese database, uttered by 80 people aged from 16 to 25 from all over China. Speakers consist of 40 males and 40 females. The sub-vocabulary for each speaker includes a mono-syllable word set (11groups by 100 Chinese words), a bi-syllable word set (63 groups by 100 words), a tri-syllable word set (11 groups by 100 words), a quad-syllable word set (10 groups by 100 words), a penta-syllable word set (1 group by 76 words), a hexa-syllable word set (1 group by 23 words) and a 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. Moreover, each speaker utters ten sentences different from those uttered by any other speakers.

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

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

DBII is another giant real-world Chinese continuous speech database of about 5GB, the data of which are sampled from the telephone network (bandwidth 3.4 kHz and sampling rate 8 kHz) and uttered by 200 males and females, aged from about 20 to 50, and with different accents from almost all over China. In such a database, people speak in a very spontaneous way, very fast and with different background noises. Fig.2 gives a histogram of the length of Chinese syllables labelled in this database. From the figure, we can see that, the average syllable dwell time is about 10 frames, i.e., 160 ms.

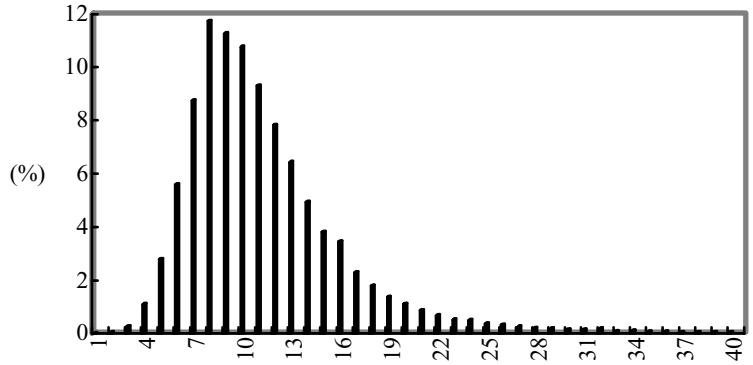


Fig. 2 A histogram of the lengths of Chinese syllables in DBII

4.2 Feature extraction

In our experiments, the linear predictive coding (LPC) analysis is performed on each windowed 32ms frame every 16ms using Levinson-Durbin recursive algorithm [21], and cepstral coefficients are computed from the Linear Predictive Coefficients [22]. Also regression analysis [23] is applied to each time function of the cepstral coefficients over 5 adjacent frames every 16 ms.

Every speech frame is then represented by a set of *D*-order cepstral coefficients and a set of regression coefficients, which are constructed as vectors in *D*-dimensional Euclidean spaces, and the vectors at *t*-th frame is denoted by $\vec{\mathbf{c}}(t)$ and $\vec{\mathbf{r}}(t)$ respectively.

Define the weighted Euclidean distance between two vectors $\vec{x}_1$ and $\vec{x}_2$ as

$$y(\vec{x}_1, \vec{x}_2) = \sqrt{\sum_{d=1}^{D} w_d (x_{1d} - x_{2d})^2}\ , \tag{32}$$

where $\vec{x}_1$ and $\vec{x}_2$ can be either cepstral vectors or regression vectors, and $\vec{w} = (w_1, w_2, \ldots, w_D)$ is the weight vector. In our experiments, the *d*-th component of the weight vector 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.

The two kinds of feature vectors are used in clustering and scoring procedures separately [15-17]. Both cepstral vectors and their corresponding regression vectors are described by their own PDFs. Let $b_n^{(c)}(\vec{\mathbf{c}})$ be the PDF of cepstral vectors in state *n*, and $b_n^{(r)}(\vec{\mathbf{r}})$ the PDF of regression vectors in state *n*, the score of an observation vector pair $(\vec{\mathbf{c}}(t), \vec{\mathbf{r}}(t))$ in state *n* is then computed by

$$b_n(\vec{\mathbf{c}}(t), \vec{\mathbf{r}}(t)) = b_n^{(c)}(\vec{\mathbf{c}}(t)) * b_n^{(r)}(\vec{\mathbf{r}}(t))\,. \tag{33}$$

4.3 Speech recognition units

There are about 419 toneless syllables and 5 tones in standard Chinese speech, totally about 1300 syllables with tone. Each syllable consists of two parts: an initial and a final. Totally 22 initials (including a null-initial) and 38 finals occur in Chinese speech. In order to supply more precise description, the initials and the finals can be divided into 22 consonant phonemes and 17 vowel phonemes [24]. Choosing Chinese speech recognition units is based on the described acoustics knowledge. Experiments show that Chinese syllables as speech recognition units are the best choice for continuous speech recognition [13-14].

5. Experimental Results

In the following experiments, the weighted Euclidean measure is chosen to be the Mahalanobis distance measure where the covariance matrix is simplified to a diagonal covariance matrix.

5.1 Comparison on CDCPM via CHMM

In this experiment [14], testing vocabulary include all 35 Chinese finals, the database is a 840 times' repetition of the vocabulary by a male speaker. The sampling rate is 8KHz, and feature is the 16^{th} order cepstrum derived from 12^{th} LPC coefficients.

The number of states for each final is 5, and the number of densities for each state is 2 and 3 for CDCPM and CHMM respectively. The error rates are 1.07% and 1.43% for CDCPM and CHMM respectively.

The results show that the CDCPM performs as well as the CHMM even it is simplified, because the ignored part is not so import and furthermore the CDCPM focuses on the more import part of the acoustic models.

The storage complexity CDCPM-to-CHMM ratio is $1/D$, and the time complexity ratio is $1/2D$, where *D* is the number of dimension of the features.

5.2 Comparison on EMM scoring via MGD scoring

This experiment is designed to test which form of scoring function is better, two kinds of scoring functions are compared, one is based on mixed CDN densities (MCDND) as

$$b_n(\vec{x}) = \sum_{m=1}^{M} g_{nm} N_{CD}(\vec{x}; \vec{\mu}_{xnm}, \mu_{ynm}) . \quad (34)$$

and another is the EMM scheme as defined in Equ. (22).

The training and testing data are taken from DBI, and cover about one fourth of DBI. The features are D=16th order cepstral coefficients, and each CDCPM has N=6 states and M=6 CDN densities each state.

The testing vocabulary here include all the Chinese finals. Experimental results are given in Tab. 1. From Tab. 1, we can see that the latter form, i.e., the EMM scheme, performs better.

Tab.1 Comparison on forms of scoring functions

Top *n*	1	2	3	4	5	6	7	8	9	10	12	18
EMM	78.22	91.48	95.52	97.54	98.53	99.12	99.40	99.59	99.70	99.77	99.9	n/a
MCDND	70.22	86.18	91.61	94.48	96.21	97.45	98.24	98.79	99.16	99.35	n/a	99.9

The EMM scheme 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 $\mathbf{O} = (\vec{\mathbf{o}}_1, \vec{\mathbf{o}}_2, \ldots, \vec{\mathbf{o}}_T)$. There exists a segmentation determining which state it belongs to for any $\vec{\mathbf{o}}_t$. For any segmentation, scoring using Equ. (22) 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 one. That's a good explanation why the EMM performs better.

5.3 A real-world continuous-manner speech recognition system

A 2000-phrase continuous-manner speech recognition system has been established based on CDCPMs [16]. The training data are taken from DBI, about one fourth of the database is used, and testing data are uttered by several speakers in a real-world environment. In this system, the vocabulary consists of 2000 Chinese phrases of 3 to 5 syllables. It is very flexible to manage the vocabulary, you can add a phrase to, modify a phrase in, or delete a phrase from the vocabulary without extra training. In Tab. 2, recognition rates for training and testing sets are listed.

Tab.2 Performance of a 2000-phrase real-world system

Training Set	1st candidate
M00	99.65%
M01	99.90%
M02	99.90%
M03	99.95%

Testing Set	1st candidate
M10	97.80%
M11	98.00%
M20	95.40%
M21	98.40%

5.4 Experiments in the very-bad environment

Experiments are also done across the whole DBII which is established in a very-bad environment. Some utterances are even not understandable by human. The sampling rate is 8 kHz. And 10th order LPC analysis is performed on the sampled data. The feature vector dimension is D=10.

Each CDCPM used here has N=6 states and maximal M=16 CDN densities each state. When training the CDCPM, the density number in each state is either fixed (FIX) (16 densities) or variant (VAR) (maximally 16 densities). In the VAR scheme, the number of CDN densities are

determined by the training data by some criterion, which will be discussed in another paper. Results show that using different number of densities performs better.

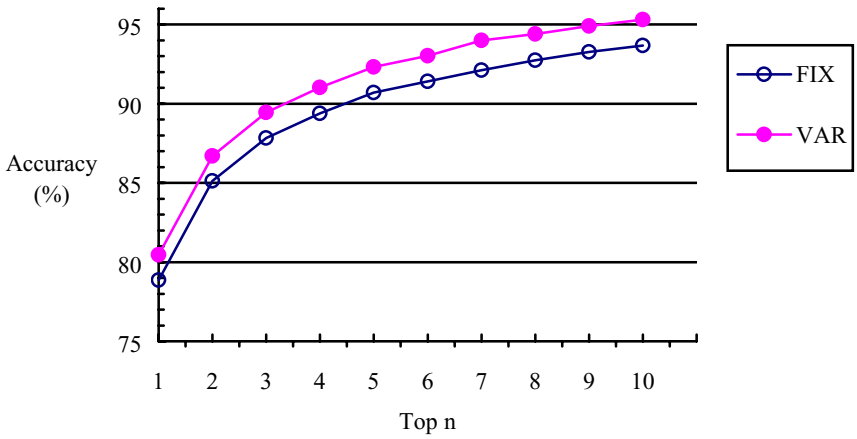


Fig. 3 Recognition Rate on DBII
("FIX" means fixed number of CDN densities each state while
"VAR" means various number of CDN densities each state)

6. Summary

In this paper, a new, powerful, and low-complexity model, named CDCPM, has been described, and the distance measure for CDCPMs is proposed. A CDCPM is a simplified HMM, all algorithms designed for HMM can be used for CDCPMs after modification. The experimental results show that it is a potential model.

We can draw the following conclusions:

(1) A CDCPM is a new simplified version of a CHMM, the observation probabilities in matrix B is simplified to be mono-dimensional distance-based PDFs.
(2) The Nearest Neighbour rule based EMM scoring scheme is better than MGD scoring scheme.
(3) Distance measure for acoustic models is important to theoretically studying and guiding.

ACKNOWLEDGMENT:

We'd like to thank Mr. Haixin Chai and Mr. Zhijie Shi of our lab, for their establishing a 2000-military-phrase Chinese speech recognition system using CDCPMs, and supplying us with the experimental platform and testing results based on CDCPMs. We would also offer our thanks to Prof. Yubo Ge of Dept. of Applied Mathematics, Tsinghua University for his valuable suggestions on this paper.

REFERENCES:

[1] **Bahl, L.R., Brown, P.F., de Souza, P.V., Mercer, K.L., (1990)** "Speech Recognition with Continuous-parameter Hidden Markov Models," In *Readings in Speech Recognition*, Alex Waibel & Kai-Fu Lee (eds.), 1990, pp.332-339

[2] **Rabiner, L.R., Juang, B.-H., Levinson, S.E., Sondhi, M.M., (1985)** "Recognition of Isolated Digits Using Hidden Markov Models With Continuous Mixture Densities," *AT&T Technical Journal*, July-August 1985, 64(6):1211-1234

[3] **Juang, B.-H., Rabiner, L.R., (1985b)** "Mixture autoregressive hidden Markov Models for speech signals," *IEEE Trans. on Acoust., Speech, and Signal Processing*, 1985, 33:1404-1413

[4] **Huang, X.D., Jack, M.A., (1989)** "Semi-Continuous Hidden Markov Models for Speech Signals," *Computer Speech and Language,* 1989, 3:239-251
[5] **Rabiner, L.R., Levinson, S.E., Sondhi, M.M., (1983)** "On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition," *Bell Syst. Tech. J.*, 1983, 62:1075-1105
[6] **Baum, L.E., (1972)** "An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of Markov Processes," *Inequalities*, 1972, 3:1-8
[7] **Dempster, A.P., Laird, N.M., Rubin, D.B., (1977)** "Maximum likelihood from incomplete data via the EM algorithm," In *Proc. R. Stat. Soc. B.*, 1977, 39(1):1-38
[8] **Bahl, L.R., Brown, P.F., de Souza, P.V., Mercer, R.L., (1986)** "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," *ICASSP-86*, Apr. 1986, pp. 49-52
[9] **Gauvain, J.-L., Lee, C.-H., (1992)** "Improved acoustic modeling with Bayesian learning," *ICASSP-92*, 1992, 1:481-485
[10] **Viterbi, A.J., (1967)** "Error Bounds for Convolutional Codes and An Asymptotically Optimum Decoding Algorithm," *IEEE Trans. on IT-13(2)*, Apr., 1967
[11] **Lee, C.-H., Rabiner, L.R., (1989)** "A Frame-Synchronous Network Search Algorithm for Connected Word Recognition," *IEEE Trans. on ASSP*, Nov.1989, 37(11):1649-1658
[12] **Jiang, L., Wu, W.H., Cai, L.H., Fang, D.T., (1990)** "A Real-time Speaker-independent Speech Recognition System Based on SPM for 208 Chinese Words," in Proc. of *ICSP'90*, 1990, pp.473-476
[13] **Zheng F., Wu, W.H., Fang, D.T., (1996a)** "The speech recognition unit in the Chinese Dictation Machine." In Proc. *4th National Conf. on Man-Machine Speech Commun. (NCMMSC-96)*, Oct. 1996, pp.32-35, (in Chinese)
[14] **Zheng F., Wu, W.H., Fang, D.T., (1996b)** "CDCPM with Its Applications to Speech Recognition," *J. of Software*, Oct. 1996, 7: 69-75
[15] **Zheng F., Wu, W.H., Fang, D.T., (1997a)** "A new model for speech recognition: center-distance continuous probability model," in Proc. of *the First China-Japan Workshop on Spoken Language Processing (CJSLP'97)* , Huangshan, P.R. China, March, 1997, pp.224-228
[16] **Zheng F., Chai, H.X., Shi, Z.J., Wu, W.H., Fang, D.T., (1997b)** "A real-world speech recognition system based on CDCPMs," in Proc. of *Int'l Conf. On Computer Processing of Oriental Languages (ICCPOL'97)*, Hong Kong, Apr. 2, 1997, 1: 204-207
[17] **Zheng F., (1997c)** Studies on Approaches of Keyword Spotting in Unconstrained Continuous Speech: PhD Dissertation. Beijing: Tsinghua University, May. 1997.
[18] **Juang, B.-H., Rabiner, L.R., (1985a)** "A probabilistic distance measure for hidden Markov models," *AT&T Tech. J.* , Feb. 1985*,* 64(2): 391-408
[19] **Forney, G.D., (1973)** "The Viterbi algorithm," in Proc. *IEEE*, March, 1973, 61:268-278
[20] **Linde, Y., Buzo, A., Gray, R.M., (1980)** "An Algorithm for Vector Quantization Design," *IEEE Trans. on COM-28(1)*, Jan., 1980, 28(1): 84-95
[21] **Makhoul, J., (1975)** "Linear Prediction: A Tutorial Review," in Proc. *IEEE*, Apr. 1975, 63:562-580
[22] **Gold B., Rader, C.M., (1969)** "Digital Processing of Signals," *New York: McGraw-Hill*, 1969, p.246
[23] **Furui, S., (1986)** "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," *IEEE Trans. on Acoust., Speech, and Signal Processing*, Feb., 1986, 34(1):52-59
[24] **Chen, Y.B., Wang, R.H., (1990)** *Speech Signal Processing*, China Science and Technology University Press, 1990

作者英文简介:

Dr. Zheng Fang currently is an associate professor with the Department of Computer Science & Technology, and also the executive director of the Analog Devices Inc.-Tsinghua DSP Technology Research Center, Tsinghua University, P.R.China. He was born in Jiangsu Province, P.R. China, in 1967. He received his B.S., M.S. and Ph.D. degrees from the Department of Computer Science & Technology of Tsinghua University, in 1990, 1992 and 1997 respectively, in Computer Science and Technology, Computer Science and Technology

and Computer Application respectively. Dr. Zheng has been working in Speech Recognition at Speech Lab., Dept. of Computer Science and Technology, Tsinghua, since 1988. His major research interests include acoustic modelling, language modelling, dictation, keyword spotting, language understanding and so on.

Prof. Wu Wenhu was born in Beijing, P.R.China, in 1936. He studied in the Department of Electrical Engineering, Tsinghua University, from 1955 to 1958, and then in the Department of Automation, Tsinghua University, from 1958 to 1961.
Since then, he has been teaching at Tsinghua University and now a Full Professor in the Department of Computer Science and Technology. He is the director of the Speech Lab now.
He is devoted in researching Chinese speech recognition and understanding, especially the speaker-independent Chinese speech recognition. As a result, he has been awarded several times.
He is also devoted in the computer spread education. He is the chairman of Computer Spread Education Commission of CCF (China Computer Federation). He has led the China Team to take part in the IOI'89 - IOI'95 (International Olympiad in Informatics) and won many golden medals.

Prof. Fang Ditang was born in Shanghai, P.R.China, in 1930. He received the B.S. degree from Jiaotong University and the M.S. degree from Tsinghua University, both in electrical engineering, in 1953 and 1956, respectively.
Since then, he has been teaching at Tsinghua University and now a Full Professor in the Department of Computer Science and Technology. In 1979, he founded the Laboratory for Human-Machine Speech Communications and has been its director from 1979 to 1990. The laboratory received the National Scientific Research and Technology Progress Award twice, in 1987 and 1989, respectively, the National Scientific Invention Award in 1990, and three other awards.
He is the Deputy Chief of the Artificial Intelligence and Pattern Recognition Committee of the Chinese Computer Science Society.

Appendices

A.1 Non-Linear Segmentation (NLS) Algorithm

Non - Linear Segmentation (NLS) algorithm used in this paper can be described as follows.

Denote the observation sequence of feature vectors by $\mathbf{O} = (\vec{o}_1, \vec{o}_{2,\cdots,} \vec{o}_T)$, which is to be segmented into N parts. Define

$$y_i \overset{def}{=} y(\vec{o}_{i+1}, \vec{o}_i), \quad 1 \le i \le T-1, \quad \Delta y \overset{def}{=} \frac{1}{N} \sum_{i=1}^{T-1} y_i, \tag{A1}$$

Then L_n, the length of the first n segments, can be calculated by the following rule

$$\begin{aligned} &if \quad \sum_{i=1}^{k-1} y_i < n\,\Delta y \le \sum_{i=1}^{k} y_i, \quad then \quad L_n = k, \qquad 1 \le n \le N-1 \\ &L_N \overset{def}{=} T \end{aligned} \tag{A2}$$

From Equ. (36), it is obvious that the sum of feature variations in each segment equals approximately to one another, so the NLS algorithm is efficient, robust, and useful for providing a initial segmentation for training procedure and for isolated word recognizing procedure.

A.2 The properties of function $\Phi(x)$

The normal probability distribution function $\Phi(x)$ has the following properties [17]:

1. $0 \le \Phi(x) \le 1, \quad \forall x \in R$ (A3)
2. $\Phi(x) \ge 0.5, \quad \forall x \ge 0 \quad and\ \Phi(0) = 0.5$ (A4)
3. $\Phi(x+\delta) \ge \Phi(x), \quad \forall \delta \ge 0, x \in R$ (A5)
4. $\Phi(x) + \Phi(-x) = 1, \quad \forall x \in R$ (A6)
5. $\big(\Phi(x) - \Phi(0)\big) + \big(\Phi(y) - \Phi(0)\big) \ge \big(\Phi(x+y) - \Phi(0)\big), \quad \forall x, y \ge 0$ (A7)

The following is the proof for Property 5.

[Proof]

$$\because \big(\Phi(x) - \Phi(0)\big) + \big(\Phi(y) - \Phi(0)\big) - \big(\Phi(x+y) - \Phi(0)\big)$$

$$= \int_0^x \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)dt + \int_0^y \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)dt - \int_0^{x+y} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)dt$$

$$= \int_0^y \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)dt - \int_x^{x+y} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)dt$$

$$= \int_0^y \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)dt - \int_0^y \frac{1}{\sqrt{2\pi}} \exp(-(t+x)^2/2)dt$$

$$= \int_0^y \frac{1}{\sqrt{2\pi}} \big\{\exp\big(-t^2/2\big) - \exp\big(-(t+x)^2/2\big)\big\}dt$$

$$\ge 0$$

$$\therefore \big(\Phi(x) - \Phi(0)\big) + \big(\Phi(y) - \Phi(0)\big) \ge \big(\Phi(x+y) - \Phi(0)\big)$$