The First China-Japan Workshop on Spoken Language Processing (CJSLP'97), 187-192, Mar. 1997, Huangshan, P.R. China

A New Model for Speech Recognition: Center-Distance Continuous Probability Model

Fang ZHENG, Wenhu WU, and Ditang FANG
Speech Lab., Dept. of Computer Science and Technology,
Tsinghua Univ., Beijing, 100084, P.R.China
fzheng@sp.cs.tsinghua.edu.cn, +86-10-62784141

ABSTRACT

In this paper a new statistic model named Center-Distance Continuous Probability Model (CDCPM) is proposed for speech recognition, which is based on Center-Distance Normal (CDN) distribution. In a CDCPM, state parameters for each mixture include a mean vector and a CDN distribution parameter. Unlike the continuous Hidden Markov Model (CHMM), it preserves only the observation probability density function (PDF) matrix B, in which PDFs are always mono-dimensional ones and the scoring scheme is based on Embedded Multi-Model (EMM) scheme. The experimental results across a giant Chinese speech database and a real-world continuous-manner 2000 phrase system show that this model is a powerful one, which extremely reduces the space and time complexities, preserving good performance.

1. INTRODUCTION

Compared to the traditional HMMs [1-5], the center-distance continuous probability model (CDCPM) preserves only the B-matrix and the observation probability density function (PDF) is replaced by a one-dimensional (center-distance) PDF. This replacement will reduce the time and space complexities to a great extent, preserving good performance.

This paper will focus on the CDCPM and the forms of scoring functions of observation feature vectors. Scoring functions based on mixed CDN density and Nearest-neighbor (NN) rule are compared, the later is referred to as an Embedded Multi-Model (EMM) scheme and performs better.

2. FEATURE EXTRACTION

In our experiments, speech signal is digitized at 16KHz sampling rate with 8KHz cut-off , emphasized using a simple 1st-order digital filter with 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 *D*-order (where *D*=16) LPC cepstral coefficients [6] and denoted by a vector $\vec{\mathbf{c}}(t) = (c_1(t), c_2(t), \ldots, c_D(t))$. Regression analysis [3] is applied to each time function of the cepstral coefficients over several adjacent frames every 16 msec. The result is denoted by another vector $\vec{\mathbf{r}}(t) = (r_1(t), r_2(t), \ldots, r_D(t))$ .For convenience, define the weighted Euclidean distance measure between two vectors $\vec{\mathbf{x}}_1$ and $\vec{\mathbf{x}}_2$ as:

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

where $\vec{\mathbf{x}}$'s 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 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 cepstral vector sequence $\{\vec{\mathbf{c}}(t)\}$ and the regression vector sequence $\{\vec{\mathbf{r}}(t)\}$. There are two ways to combine the two kinds of features. The first method is to combine them into one large *D*2*-dimensional vector as $\vec{\mathbf{v}}(t) = (\vec{\mathbf{c}}(t), \alpha\, \vec{\mathbf{r}}(t))$ where α is a balance coefficient. And the distance measure between $\vec{\mathbf{v}}_1$ and $\vec{\mathbf{v}}_2$ in the *D*2*-dimensional Euclidean space is defined as

$$y(\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2) = \sqrt{[y(\vec{\mathbf{c}}_1, \vec{\mathbf{c}}_2)]^2 + [\alpha\, y(\vec{\mathbf{r}}_1, \vec{\mathbf{r}}_2)]^2}\ . \qquad (2)$$

The second method is to use the cepstral vector and its corresponding regression vector separately, in both clustering and scoring procedures. Both cepstral vectors and their corresponding regression vectors are described by their own probability density functions (PDFs), see Section 3 for details. Let $b_n^{(c)}(\vec{\mathbf{c}}(t))$ be the PDF of cepstral vectors in state *n*, and $b_n^{(r)}(\vec{\mathbf{r}}(t))$ the PDF of regression vectors in state *n*, the score of an observation vector $\vec{\mathbf{v}}(t)$ in state *n* is then computed by

$$b_n(\vec{\mathbf{v}}(t)) = b_n^{(c)}(\vec{\mathbf{c}}(t)) * b_n^{(r)}(\vec{\mathbf{r}}(t))\,, \qquad (3)$$

where α is of no use. The later method has been proved better [8-9].

3. THE CDCPM

3.1 The Center-Distance Normal Distribution

Let $p(x; \mu_x, \sigma_x)$ be the PDF of a normal variable ξ with mean value μ_x and standard deviation σ_x. Define $\eta = |\xi - \mu_x|$, we have the PDF of η as

$$p(y; \sigma_x) = \frac{2}{\sqrt{2\pi}\,\sigma_x} \exp(-y^2 / 2\sigma_x{}^2), \quad y \ge 0\,. \qquad (4)$$

By calculating the mean value of this distribution, we can change it to another form:

$$p(y; \mu_y) = \frac{2}{\pi \mu_y} \exp(-y^2 / \pi \mu_y^2)\,, \qquad (5)$$

where μ_y is the mean value of η. In fact, η is the distance between a normal variable ξ and its mean value μ_x, thus the derived distribution is referred to as Center-Distance Normal (CDN) distribution.

In D-dimensional case, denote the (weighted) Euclidean distance between a D-dimensional normal vector $\vec{\xi}$ and its mean value vector $\vec{\mu}_{\mathbf{x}}$ by another random variable η. Assume η is a CDN variable, then its CDN PDF is

$$p(\vec{x};\vec{\mu}_{\mathbf{x}},\mu_y) = \frac{2}{\pi\mu_y}\exp(-y^2(\vec{x},\vec{\mu}_{\mathbf{x}})/\pi\mu_y^2) \overset{def}{=} \mathcal{N}_{CD}(\vec{x};\vec{\mu}_{\mathbf{x}},\mu_y)\,. \qquad (6)$$

Strictly speaking, $p(\vec{x};\vec{\mu}_{\mathbf{x}},\mu_y)$ is the PDF of $y(\vec{\xi},\vec{\mu}_{\mathbf{x}})$ instead of that of $\vec{\xi}$, it is just for convenience and comparison. The distribution parameters $\vec{\mu}_{\mathbf{x}}$ and μ_y in Eq. (6) can be estimated easily.

3.2 The Center-Distance Continuous Probability Model

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

3.3 The Scoring Scheme

Given an observation (feature vector) sequence $\mathbf{O} = (\vec{o}_1,\vec{o}_2,\cdots,\vec{o}_T)$, where $\vec{o}_t$ is a cepstral vector or a combined feature vector, the matching score of the sequence with the model $\Lambda = \left\{\vec{\mu}_{xn},\mu_{yn},b_n(\vec{x}) \middle| 1 \le n \le N\right\}$ is calculated as follows:

$$score(\mathbf{O};\Lambda) = \prod_{t=1}^{T} b_n(\vec{o}_t \,|\, \vec{o}_t \in state\ n)\,. \qquad (7)$$

Similarly, mixed Gussian densities (MGD) [2], tied MGD [10], or other forms [11] can be used as observation PDF or scoring function. The question is how to determine the segmentation of the observation sequence, i.e., how to determine which state a feature vector belongs to, see Section 3.5 for more details.

Mixed CDN densities have the following equations for cepstral and regression representations, the Bayesian learning method [12] can be employed for a CDCPM to train $b_n^{(c)}(\vec{\mathbf{c}})$ and $b_n^{(r)}(\vec{\mathbf{r}})$:

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

where $1 \le n \le N$, $1 \le m \le M$, $1 \le d \le D$ and $b_n^{(c)}(\vec{\mathbf{c}})$ and $b_n^{(r)}(\vec{\mathbf{r}})$ are the PDFs of cepstral and regression features in state n respectively. The scoring function for vector $\vec{\mathbf{v}}$ is defined as in Eq. (3).

In this paper, we propose another form based on Nearest-Neighbor rule:

$$b_n^{(c)}(\vec{\mathbf{c}}) = \max_{1 \le m \le M} \mathcal{N}_{CD}(\vec{\mathbf{c}};\vec{\mu}_{xnm}^{(c)},\mu_{ynm}^{(c)}), b_n^{(r)}(\vec{\mathbf{r}}) = \max_{1 \le m \le M} \mathcal{N}_{CD}(\vec{\mathbf{r}};\vec{\mu}_{xnm}^{(r)},\mu_{ynm}^{(r)}) \qquad (9)$$

3.4 Training the CDCPM

Once N, M, and D have been determined, the training procedure is simple: 1) Each observation feature sequence $\mathbf{O}$ from the training set is segmented into N segments (states) using some segmentation method such as the Non-Linear Segmentation (NLS) method [13]. 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 [14]. 3) Estimate $\vec{\mu}_{xnm}$ and μ_{ynm} for each mixture (class) of the specified segment n.

3.5 The State Transition Rule

For an isolated system, the given observation sequence is first segmented before it is scored. The matching score with a CDCPM is calculated using Eq. (7). This scoring strategy is useful and proved efficient [8, 15].

In a continuous recognition system multi-state left-to-right CDCPMs can be adopted for meaningful speech recognition units, and mono-state CDCPMs for SILENCE or GARBAGE models. The frame synchronization algorithm [16] and its modified versions can be adopted to determine the state transition sequence in recognizing procedure.

4. DATABASE DESCRIPTION

The speech database used to train and test here is a giant real-world Chinese database [17-18] consisting of 25GB speech data about 230 hours' utterances.

5. EXPERIMENTAL RESULTS

Experiments across three different databases have shown that CDCPMs are good models [9]. In this paper, totally 46 experiments are done across the above database, only one group is given due to space considerations, but all experimental results support the conclusions to be given below.

5.1 Comparison on Forms of Scoring Functions

This group of experiments are designed to test which form of scoring function is better. See Table 1. The SR units used here are Chinese finals, and the features are cepstral coefficients. Two kinds of scoring functions are compared, one is based on mixed CDN densities (MCDND) using Eq. (8) and another one is NN-based, named Embedded Multi-Model Scheme (EMM), using Eq. (9).

In our experiments, the utterances by the first 20 males are used as the training set while those by the second 20 males as the testing set. Listed in Table 1 are the average rates over training and testing sets.

5.2 Experiments on a 2000-phrase Real-world System

A 2000-phrase continuous-manner speech recognition system has been established based on CDCPMs, the vocabulary consists of 2000 Chinese phrases of 3 to 5 syllables. Also, the constrained frame synchronization algorithm is applied to recognizing procedure. In Table 2, 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.

6. SUMMARY

In this paper a new model named CDCPM is proposed. Through the experiments, we have the following conclusions: (1) A CDCPM is a new simplified version of a CHMM, the observation probabilities in matrix B is simplified to be a mono-dimensional distance-based PDFs. The information that is contained in the $D \times D$ covariance matrix of a CHMM is partly included in the weight vector of the distance measure of the CDCPM. A simple rule instead of the transition matrix A is adopted to make transitions, and the rule is based on observation probabilities. Not only the time and space complexities are much smaller than those of the traditional CHMM, but also the performance is not reduced. (2) The EMM scoring functions are better than mixed CDN density scoring functions. Simply speaking, for N frames of unknown utterance, a CDCPM based on EMM can be regarded as a somewhat mixture of M^N mono-mixture CDCPMs, and the resulted matching score is the maximum of the unknown utterance among these mono-mixture CDCPMs. But actually this is more complicated. It is an embedded multi-model scheme so we name it an EMM scheme. The estimation of EMM scoring function parameters is much easier, some simple clustering algorithm is enough, but the performance is better.

REFERENCES:

[1] **L.R. Rabiner, S.E. Levinson, and M.M Sondhi**, "On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition," *Bell Syst. Tech. J.*, vol.62, pp.1075-1105

[2] **L.R. Rabiner, B.-H. Juang, S.E. Levinson, and 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

[3] **B.-H. Juang, L.R. Rabiner**, "Mixture autoregresive hidden Markov Models for speech signals," *IEEE Trans. on Acoust., Speech, and Signal Processing*, vol. ASSP-33, 1985, pp.1404-1413

[4] **X.-D. Huang & M.A. Jack**, "Semi-Continuous Hidden Markov Models for Speech Signals," *Computer Speech and Language (1989)*, 3:239-251

[5] **L.E. Bahl, P.F. Brown, P.V. de Souza, and K.L. Mercer**, "Speech Recognition with Continuous-parameter Hidden Markov Models," *Readings in Speech Recognition*, edited by Alex Waibel & Kai-Fu Lee, 1990, pp.332-339

[6] **J. Makhoul**, "Linear Prediction: A Tutorial Review," Proc. *IEEE*, vol. 63, Apr. 1975, pp.562-580

[7] **S. Furui**, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," *IEEE Trans. on Acoust., Speech, and Signal Processing*, vol. ASSP-34, No. 1. Feb., 1986, pp.52-59

[8] **F. Zheng, W.-H. Wu, D.-T. Fang**, "Speech Recognition Units in the Chinese Dictation Machines," Proc. in *4th National Conf. on Man-Machine Speech Commu. (NCMMSC-96)*, pp.32-35, Oct. 1996, Beijing, P.R. China

[9] **F. Zheng, W.-H. Wu, D.-T. Fang**, "CDCPM with Its Applications to Speech Recognition," *Chinese J. of Advanced Software Research*, Supplement, 1996 (Accepted)

[10] **J.R. Bellegarda, D. Nahamoo,** "Tied Mixture Continuous Parameter Modeling for Speech Recognition," *IEEE Trans. on ASSP*, vol.ASSP-38, No.12, Nov. 1990, pp.2033-2045

[11] **H. Ney**, "Modeling and Search in Continuous Speech Recognition," in Proceedings of *European Conf. On Speech Technology*, Vol.1, Berlin, 1993, pp.491-498

[12] **J.-L. Gauvain, C.-H. Lee,** "Bayesian Learning for Hidden Markov Model with Gaussian Mixture State Observation Densities," *Speech Comm.*, Vol. 11, 1992, pp.205-213

[13] **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 Proc. of *ICSP'90*, pp.473-476, 1990

[14] **Y. Linde, A. Buzo, R.M. Gray**, "An Algorithm for Vector Quantization Design," *IEEE Trans. On COM-28(1)*, Jan., 1980

[15] **F. Zheng, Q.-X. Hu, X. Deng, W.-H. Wu, D.-T. Fang**, "An Introduction to a Kind of Voice Diallers for Dummies," Proc. in *4th National Conf. on Man-Machine Speech Communications (NCMMSC-96)*, pp.165-168, Oct. 1996, Beijing, P.R.China

[16] **C.-H. Lee, 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

[17] **S.-Q. Li, D.-T. Fang, S. Qing**, "A Giant Chinese Speech Database," Proc. in *4th National Conf. on Man-Machine Speech Communications (NCMMSC-96)*, pp.342-345, Oct. 1996, Beijing, P.R.China

[18] **H.-X. Chai, F. Zheng, W.-H. Wu, D.-T. Fang,** "A Real-World Large Vocabulary Speaker-Independent Speech Recognition System", somewhere in this proceeding

Table 1. *Comparison on forms of scoring functions*

Top *n*	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**18**
EMM	78.2	91.5	95.5	97.5	98.5	99.1	99.4	99.6	99.7	99.8	99.9	n/a
MCDND	70.2	86.2	91.6	94.5	96.2	97.5	98.2	98.8	99.2	99.4	n/a	99.9

Table 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%