International Symposium on Chinese Spoken Language Processing (*ISCSLP'98*), ASR-A3, pp.49-53, Dec.7-9, 1998, Singapore

ON THE EMBEDDED MULTIPLE-MODEL SCORING SCHEME FOR SPEECH RECOGNITION

ZHENG Fang, MOU Xiaolong, WU Wenhu, and FANG Ditang
Speech Laboratory, Department of Computer Science & Technology
Tsinghua University, Beijing, 100084
Tel. /Fax: +86 10 62772001, E-mail: fzheng@sp.cs.tsinghua.edu.cn

ABSTRACT

A traditional hidden Markov model (HMM) consists of two stochastic processes, a hidden state transition process and a symbol-generating (output observation) process. Researches show that the description of the feature space or the scoring method in the second process is more important. In speech recognition, the feature space is often represented by the mixed Gaussian densities, which consumes a lot of time and storage. In this paper, a novel method is proposed, which is based on the nearest neighbour rule and is named as the embedded multiple-model (EMM) scheme. Taking both the time and space complexities, the EMM scheme has been proved efficient.

1. INTRODUCTION

In the recognition procedure of a continuous hidden Markov model (CHMM), given an *N*-state CHMM $\Lambda = \{\pi, A, B\}$ with the initial probability distribution $\pi = (\pi_i)_N$, the state transition matrix $A = (a_{ij})_{N\times N}$, and the output observation probability density function (*pdf*) matrix $B = (b_j(\cdot))_N$, the probability density of the CHMM Λ generating the specified *T*-frame observation feature sequence $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_T)$ is evaluated by Equation (1).

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

where $S = \{s_t | 1 \le t \le T\}$ is an arbitrary state transition sequence. The most widely used famous 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 Equation (1), i.e.,

$$\begin{aligned} Score\{\mathbf{O}|\Lambda\} &= f\{\mathbf{O}, S^{(ML)}|\Lambda\} \\ &= \Pr\{S^{(ML)}|\Lambda\} \cdot f\{\mathbf{O}|\Lambda, S^{(ML)}\} \end{aligned} \tag{2}$$

In this equation, the matching score is a product of two terms, (i) $\Pr\{S^{(ML)}|\Lambda\}$, the probability of Λ generating the state transition sequence on a basis of maximum likelihood estimation (MLE), (ii) $f\{\mathbf{O}|\Lambda, S^{(ML)}\}$, the *pdf* of the specified observation feature sequence given Λ and the maximum likelihood state transition sequence.

Researches on model distance measures have showed that in HMMs the probability transition matrix **A** contributes not too much as the observation function matrix B does to the recognition performance [8][11]. Hence among the two items of the scoring equation the second term plays an extremely more important role than the first term does, resulting in the elimination of the state transition matrix and the focus on the intra-state feature space description [15]. In this case, Equation (2) is changed to

$$\begin{aligned} Score\{\mathbf{O}|\Lambda\} &= f\{\mathbf{O}|\Lambda, S^{(ML)}\} \\ &= \prod_{t=1}^{T} b_{s_t^{(ML)}}(\mathbf{o}_t) \end{aligned} \tag{3}$$

where the maximum likelihood state transition sequence $S^{(ML)}$ can be decoded by modified Viterbi algorithms or modified frame synchronous search algorithms[14].

The varying of the scoring scheme, i.e., the form of the function $b_n(\cdot), 1 \le n \le N$, of the intra-state feature space description results in many better improved technologies, including vector quantization (VQ) [12], the famous mixed Gaussian densities (MGD) [1], tied mixed Gaussian densities [2], Semi-Continuous HMM [6], *etc.*.

In this paper, a novel scoring scheme, named embedded multiple-model (EMM) scheme, is introduced and studied, which is based on the nearest neighbor rule and the Gaussian density or the Center Distance Normal (CDN) density [13].

2. THE SCORING SCHEME IN ACOUSTIC MODELS

In this section, we will discuss our proposed scoring scheme in details.

2.1. The Traditional Method – Mixed Densities

As above mentioned, the most widely used form of function $b_n(\cdot)$ in Equation (3) is well known as the mixed Gaussian densities (MGD) as follows,

$$b_n(x) = \sum_{m=1}^{M} g_{nm}\, p(x|\theta_{nm}) \qquad (4)$$

where M is the number of density mixtures, $\Theta_n = \{g_{nm}, \theta_{nm} | 1 \le m \le M\}$ is the mixture parameter set of State n, $p(x|\theta_{nm})$ is a Gaussian *pdf* with parameters $\theta_{nm} = \{\mu_{nm}, \Sigma_{nm}\}$ and g_{nm} is the gain or weighting of corresponding Gaussian *pdf*. The covariance matrix Σ of a Gaussian *pdf* is often a diagonal one as $\Sigma = (\sigma_d^2)_{D\times D}$.

2.1.1. pdf: CDN vs. Gaussian density

The kind of function $p(x|\theta_{nm})$ in Equation (4) is often, but not limited to, the Gaussian *pdf*. We have tried another kind of function that performs not bad overall in the recognition accuracy and the time & space complexities. This is known as the center-distance normal (CDN) density [13]. Here is the detail.

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

$$p(y;\sigma) = \frac{2}{\sqrt{2\pi\sigma}} \exp(-y^2/2\sigma^2), y \ge 0, \qquad (5)$$

where the mean value ρ of η can be calculated to be $\rho = \frac{2\sigma}{\sqrt{2\pi}}$. In fact, η is the distance between the normal variable ξ and its mean value μ, thus the defined distribution is referred to as a CDN distribution. And the CDN pseudo-PDF can be

$$\mathrm{N}_{CD}(x;\mu,\rho) = \frac{2}{\pi\rho} \exp(-y^2(x,\mu)/\pi\rho^2) \qquad (6)$$

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

$$\mathrm{N}_{CD}(\boldsymbol{x};\boldsymbol{\mu},\rho) = \frac{2}{\pi\rho} \exp(-y^2(\boldsymbol{x},\boldsymbol{\mu})/\pi\rho^2) \qquad (7)$$

As a matter of fact, $\mathrm{N}_{CD}(\boldsymbol{x};\boldsymbol{\mu},\rho)$ is not the PDF of $\boldsymbol{\xi}$ but that of $y(\boldsymbol{\xi},\boldsymbol{\mu})$, i.e., the distance between a normal vector and its mean vector, it is just for convenience and comparison purpose, that is the reason why we name it the pseudo-pdf (*ppdf*).

2.2. The Embedded Multiple-Model (EMM) Scheme

The mixed-density as a kind of scoring scheme is not the unique one. Based on Nearest-Neighbour rule, we can change Equation (4) into an alternative form

$$b_n(x) = \max_{1\le m\le M} p(x|\theta_{nm}) \qquad (8)$$

In this situation, $p(x|\theta_{nm})$'s in Equation (8) can be the same as in Equation (4). This means the training method. Alternatively, $p(x|\theta_{nm})$'s in Equation (8) can be different from those in Equation (4) because of the different modelling methods.

No matter what kind of training method is adopted, the scoring scheme represented by Equation (8) 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 acoustic model with *N* states and *M* densities each state, and there is an unknown speech feature sequence $\mathbf{O} = (\vec{\mathbf{o}}_1, \vec{\mathbf{o}}_2, ..., \vec{\mathbf{o}}_T)$. There must in the sense of maximal likelihood exist a decoded state sequence determining which state it belongs to for any $\vec{\mathbf{o}}_t$. For any state sequence, scoring using Equation (8) leads to choosing a maximal matching score from M^T one-density models. These M^T one-density models can be regarded as embedded in the original *M*-density *N*-state model. Thus the original model is called an EMM.

The EMM scheme has been proved efficient and powerful in our previous work, especially for gender-dependent, accent-dependent, and context-dependent models and so on. If *M* is well chosen, it is enough for one model to represent several different cases for each vocabulary word [16].

The EMM is not just an approach to the mixed version, because there are two different ways available for the estimation of the parameters in Equation (8) as mentioned above. The first one is to use the same parameters as in Equation (4), which can be estimated using EM (Expecting and Maximisation) method [3]. The other training method is the simple clustering algorithm such as LBG algorithm [9]. They differ a lot

not only in the actual values obtained but also the target functions.

2.3. The Use of Context-Dependent Trajectory in EMM

When using the EMM scheme for recognition scoring, we will be able to model the context-dependent state transition trajectory as illustrated in Figure 1. Because the matching of the unknown utterance with a specific model results in a one-density acoustic model whose state number is the length of the unknown utterance, it traces the state transition trajectory. We have reasons to think that these trajectories are context-dependent, and also speaker-dependent. So we can model these trajectories for different contexts, but the model storage consumption remains the same. This idea is to be verified in later research.

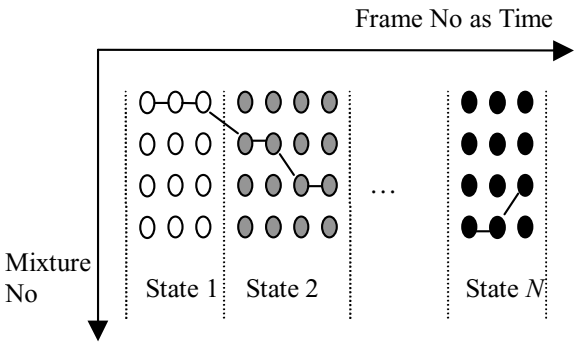

Figure 1. Context-Dependent Trajectory of state sequence in the EMM scheme

Table 1. Recognition rates (for Top 10 candidates) of different scoring schemes

Density Type	Training Method	Scoring Scheme	1	2	3	4	5	6	7	8	9	10
Gaussian	EM	Mix	62.60	77.15	83.37	86.75	89.02	90.62	91.69	92.54	93.27	93.88
		MaxW	61.86	76.66	83.15	86.65	88.91	90.53	91.73	92.53	93.29	93.90
		MaxNW	61.09	76.07	82.60	86.12	88.37	90.05	91.25	92.12	92.87	93.49
	LBG		60.21	75.81	82.63	86.42	88.85	90.47	91.64	92.52	93.27	93.81
CDN	EM	Mix	45.52	60.03	67.93	72.90	76.33	78.93	81.13	82.87	84.27	85.40
		MaxNW	57.14	73.33	80.83	85.04	87.66	89.44	90.87	91.88	92.71	93.41
	LBG		52.46	70.13	78.43	83.05	86.18	88.24	89.91	91.16	92.19	92.99

3. COMPARING SCORING SCHEMES

3.1. CDCPM and GMSM

Center-Distance Continuous Probability Model (CDCPM) [13] and Gaussian Mixture Segmentation Model (GMSM) [10] are two simplified left-to-right HMMs that ignore the probability transition matrix. During the training procedure, the state decoding is performed by the Non-linear Partition (NLP) [7] algorithms, and during the recognising procedure the modified Viterbi algorithm is adopted for state decoding. For the space description, CDCPM is based on CDN densities while GMSM based on Gaussian densities, either in mixed densities or in EMM scheme.

3.2. Comparison on Performance

According to the above discussion, once the model has been built, the scoring scheme adopted in the recognition procedure can be chosen to be any one of the following three types: (1) *Mix* as mixed densities, (2) *MaxW* as weighted maximum, and (3) *MaxNW* as non-weighted maximum.

$$\text{Mix:} \quad score(x|\Theta) = \sum_{m=1}^{M} g_m p(x|\theta_m) \tag{9}$$

$$\text{MaxW:} \quad score(x|\Theta) = \max_{1\le m\le M} g_m p(x|\theta_m) \tag{10}$$

$$\text{MaxNW:} \quad score(x|\Theta) = \max_{1\le m\le M} p(x|\theta_m) \tag{11}$$

where $p(x|\theta_m)$ can be either a Gaussian *pdf* or a CDN *ppdf*. Two methods can be used to estimate the *pdf* or *ppdf* parameters $\Theta = \{\theta_m, g_m | 1 \le m \le M\}$ in the above three equations, the EM algorithm or the cluster algorithm.

In order to test the efficiencies of the above scoring schemes, we have designed and done several experiments. The standard Mandarin database was established jointly by the University of Science and Technology of China (USTC), the Acoustic Institute of the Chinese Academy of Sciences, the Linguistic Institute of Chinese Academy of Society, as a task from the National 863 Hi-Tech project, hereafter is called the 863 Database.

Table 2. Comparison on space and time Complexities of different scoring schemes
(Number of dimensions of any feature vector is D=16.)

Scoring Function		Stored quantities			Computation in Recognition	
		K_m	μ_m Σ_m	Complexity (# of floats)	Scoring function form	Complexity (# of Add)*
Gaussian densities	Mix	g_m	μ_{md} σ^2_{md}	M(2D+1)	$\sum_{m=1}^{M} K_m \exp\left\{-ln\left(\prod_{d=1}^{D}\sigma^2_{md}\right)/2-\left(\sum_{d=1}^{D}\frac{(x_d-\mu_{md})^2}{\sigma^2_{md}}\right)/2\right\}$ M(2D+1) Add + M(2D+1) Mul + M(D+2) Div + M Ln + M Exp	M(32D+93) = 605 M
	MaxW	$ln\left(g_m/\prod_{d=1}^{D}\sigma_{md}\right)$			$\max_{1\le m\le M}\left\{K_m-\left(\sum_{d=1}^{D}\frac{(x_d-\mu_{md})^2}{\sigma^2_{md}}\right)/2\right\}$ M(2D+2) Add + MD Mul + M(D+1) Div	M(18D+4) = 292 M
	MaxNW	$ln\left(1/\prod_{d=1}^{D}\sigma_{md}\right)$				
CDN densities	Mix	$ln\left(\frac{g_m}{\rho_m}\right)$	μ_{md} $\pi\rho^2_m$	M(D+2)	$\sum_{m=1}^{M}\exp\left\{K_m-\left(\sum_{d=1}^{D}(x_d-\mu_{md})^2\right)/\left(\pi\rho^2_m\right)\right\}$ M(2D+2) Add + MD Mul + M Div + M Exp	M(16D+62) =318 M
	MaxW				$\max_{1\le m\le M}\left\{K_m-\left(\sum_{d=1}^{D}(x_d-\mu_{md})^2\right)/\left(\pi\rho^2_m\right)\right\}$ M(2D+2) Add + MD Mul + M Div	M(16D+4) = 260 M
	MaxNW	$ln\left(\frac{1}{\rho_m}\right)$				

* On an average, 1Mul= 14Add, 1Div= 2Add, 1Exp= 58Add, 1Ln= 16Add under Pentium MMX 233MHz platform.

In 863 Database, speech signal is sampled at 16 kHz sampling rate with 8 kHz cut-off through the SoundBlaster under the PC environment. Digitalized speech is emphasised using a simple first-order digital filter with the transfer function $H(z)=1-0.95z^{-1}$. The pre-emphasised 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 [5]. Regression analysis [4] is applied to each time function of the cepstral coefficients over several frames every 16 msec and the regression coefficients are obtained then.

Each of the two sets of coefficients is constructed as a vector in a D-dimensional Euclidean space and is modelled separately as if they are independent.

The 863-Database is divided into training and testing parts. The training set covers 180,063 Chinese syllable samples of 30 men's utterances while the testing set covers 70,462 Chinese syllable samples of 8 men's utterances. The experimental results presented in Table 1 are based on the testing set only and no result on the training set is given.

From Table 1, we can see that CDN-based CDCPM and MGD-based GMSM achieve almost the same performance if considering the accuracy of top 10 candidates which would be important for continuous speech recognition, and that the EMM scheme performs better than the normal mixed-density scheme. Using the EMM scheme, for CDN densities the space complexity is reduced by 18.2% and the performance is improved by 25.2% while for Gaussian densities the space complexity is reduced by 51.7% and performance is only reduced by 1.2%.

3.3. Comparison on Time and Space Complexities

For any kind of scoring scheme based on any kind of *pdf* form, for example Equations (9), (10) or (11) based on Gaussian density or CDN density, well-defined stored quantities can speed up the recognition procedure. In Table 2, the time complexity as well as the space complexity is given.

No matter what kind of space description method is used, CDNs or MGDs, the EMM scheme always shows higher advantages than the normal mixed densities in both accuracy and storage consumption.

4. CONCLUSIONS

Conclusions can be drawn upon the above described comparisons on both performance and complexities of different scoring scheme and different *pdf* forms as follows.

(1) The EMM scheme performs better than the normal mixed-density scheme if taking both the recognition rate and the time & space complexities into consideration. Using the EMM scheme, for CDN densities the space complexity is reduced and the performance is improved, and for Gaussian densities the space complexity is reduced a lot with a little reduction in performance.

(2) The EMM scheme is robust for gender-dependent, accent-dependent, and context-dependent models, because this kind of scoring scheme is based on the nearest neighbour rule which can interpret the model dynamically according to the unknown speech feature sequence to be matched.

(3) The EMM scheme can be used to model the state transition trajectories in different contexts, leaving the storage consumption not increased.

REFERENCES

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

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

[3] **Dempster A P, Laird N M and Rubin D B.** Maximum Likelihood from Incomplete Data via the EM Algorithm. *Proc. R. Stat. Soc. B*, 39(1): 1-38, 1977

[4] **Furui S.** 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

[5] **Gold B and Rader C M.** Digital Processing of Signals. *New York: McGraw-Hill*, 1969, p.246

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

[7] **Jiang L, Wu W H, Cai L H and Fang D T.** A Real-time Speaker-independent Speech Recognition System Based on SPM for 208 Chinese Words. in Proc. of *ICSP'90*, 1990, pp.473-476

[8] **Juang B H and Rabiner L R.** A probabilistic distance measure for hidden Markov models. *AT&T Technical Journal*, Feb., 1985, 64(2): 391-408

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

[10] **Mou X L.** Study and Implementation of Chinese Dictation Machine. Master Thesis. Department of Computer Science and Technology, Tsinghua University. May 1998.

[11] **Ney H.** Modeling and search in continuous speech recognition. *European Conf. On Speech Technology*, 1993, Berlin, 1: 491-498

[12] **Rabiner L R, Levinson S E and Sondhi M M.** On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition. *Bell Syst. Tech. J.*, 1983, 62:1075-1105

[13] **Zheng F, Wu W H and Fang D T.** CDCPM with its applications to speech recognition. *J. of Software*, 7:69-72, Oct. 1996 (In Chinese)

[14] **Zheng F.** Studies on Approaches to Keyword Spotting in Unconstrained Continuous Speech: PhD Dissertation. Beijing: Tsinghua University, May. 1997.

[15] **Zheng F, Xu M X, and Wu W H.** The Description of the Intra-State Feature Space in Speech Recognition. '97 Int'l Conf. Research on Computational Linguistics, 272-276, Aug. 22-24, 1997, Taiwan

[16] **Zheng F, Chai H X, Shi Z J, *et al.*** A real-world speech recognition system based on CDCPMs. *J. CPOL (Journal of Computer Processing of Oriental Languages)*, 11(3): 221-231, March. 1998