ACKNOWLEDGMENT

The authors gratefully acknowledge the help of two anonymous referees and thank Prof. N. P. Bhatia, University of Maryland and Dr. G. Agrawal, Dayalbagh Educational Institute, for their constructive comments.

REFERENCES

[1] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," in *Proc. Nat. Acad. Sci. USA*, vol. 79, 1982, pp.

[2] 2554–2558.

[3] ——, "Neurons with graded response have collective computational properties like that of two-state neurons," in *Proc. Nat. Acad. Sci. USA*, vol. 81, 1984, pp. 3088–3092.

[4] J. J. Hopfield and D. W. Tank, "Neural computation of decisions in optimization problems," *Biol. Cybern.*, vol. 52, pp. 141–152, 1985.

[5] D. W. Tank and J. J. Hopfield, "Simple neural optimization networks: An A/D converter, signal decision circuit, and a linear programming circuit," *IEEE Trans. Circuits Syst.*, vol. CAS-33, pp. 533–541, 1986.

[6] B. Kosko, *Neural Networks and Fuzzy Systems: A Dynamical Systems Approach to Machine Intelligence*. Englewood Cliffs, NJ: Prentice-Hall, 1992.

[7] J. Bruck, "On the convergence properties of the Hopfield model," *Proc. IEEE*, vol. 78, no. 10, pp. 1579–1585, Oct. 1990.

[8] E. Goles, F. Fogelman, and D. Pellegrin, "Decreasing energy functions as a tool for studying threshold networks," *Discrete Appl. Math.*, vol. 12, pp. 261–277, 1985.

[9] H. H. Szu, *Neural Networks: Theory, Applications and Computing*, Lecture Notes for UCLA Eng. Short Course, March 20–23, 1989.

[10] S. Aiyer, M. Niranjan, and F. Fallside, "A theoretical investigation into the performance of the Hopfield model," *IEEE Trans. Neural Networks*, vol. 1, pp. 204–216, 1990.

[11] S. R. Das, "On the synthesis of nonlinear continuous neural networks," *IEEE Trans. Syst., Man., Cybern.*, vol. 21, pp. 413–418, 1991.

[12] A. Dembo, "On the capacity of associative memories with linear threshold functions," *IEEE Trans. Inform. Theory*, vol. 35, pp. 709–720, 1989.

[13] R. J. McEliece, C. E. Posner, R. R. Rodemich, and S. S. Venkatesh, "The capacity of the Hopfield associative memory," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 461–482, 1987.

[14] A. Michael, J. Si, and G. Yen, "Analysis and synthesis of a class of discrete-time neural networks described on hypercubes," *IEEE Trans. Neural Networks*, vol. 2, pp. 32–47, 1991.

[15] S. S. Venkatesh and D. Psaltis, "Linear and logarithmic capacities in associative neural networks," *IEEE Trans. Inform. Theory*, vol. 35, pp. 558–568, 1989.

[16] D. J. Amit, *Modelling Brain Function: The World of Attractor Neural Networks*. Cambridge, U.K.: Cambridge Univ. Press, 1989.

[17] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation*. Reading, MA: Addison-Wesley, 1991.

A Modified HME Architecture for Text-Dependent Speaker Identification

Ke Chen, Dahong Xie, and Huisheng Chi

***Abstract*—A modified hierarchical mixtures of experts (HME) architecture is presented for text-dependent speaker identification. A new gating network is introduced to the original HME architecture for the use of instantaneous and transitional spectral information in text-dependent speaker identification. The statistical model underlying the proposed architecture is presented and learning is treated as a maximum likelihood problem; in particular, an expectation-maximization (EM) algorithm is also proposed for adjusting the parameters of the proposed architecture. An evaluation has been carried out using a database of isolated digit utterances by 10 male speakers. Experimental results demonstrate that the proposed architecture outperforms the original HME architecture in text-dependent speaker identification.**

I. INTRODUCTION

The speaker identification task is to classify an unlabeled voice token as belonging to one of a set of N reference speakers. Speaker identification systems can be either *text-dependent* or *text-independent*. By text-dependent, we mean that the text in both training and test is the same or is known. This is a different problem in comparison with text-independent identification, where the text should be any text in either training or test. In this paper, only text-dependent speaker identification is considered. There have been extensive studies on speaker identification [1]–[5]. Recently, the connectionist approaches have been introduced to speaker identification systems [6]. Neural-network classifiers may lead to good performance because they allow to take into account interspeaker information and to build complex decision regions for classification. There has recently been widespread interest in the use of multiple models for classification and regression in the statistics and neural networks communities. The hierarchical mixtures of experts (HME) is a typical modular neural network architecture in which multiple subnetworks cooperate with each other based upon the principle of divide-and-conquer for dealing with a given problem. In particular, learning in the HME is treated as a maximum likelihood problem and expectation-maximization (EM) algorithm is employed for adjusting the parameters of the architecture. Both theoretical and empirical studies [7], [8] have shown that the HME yields significantly fast training and has been successful in a number of regression and classification problems [7]. Recently, the HME has been applied to text-dependent speaker identification and outperforms classic neural networks (e.g., MLP and RBF) in both identifying accuracy and training speed [9]–[11].

In general, the process of automatic speaker identification consists of three phases, i.e., preprocessing, feature extraction and classification. For text-dependent speaker identification, the text in both training and test is the same or is known. Thus, the utterance of a fixed text naturally becomes a sequence consisting of successive feature frames after preprocessing and feature extraction and the problem of text-dependent speaker identification may be viewed as a

Manuscript received May 31, 1995; revised May 15, 1996. This work was supported in part by the Chinese National Science Foundation Grants 69571002 and 69475007 as well as the Climbing Program–National Key Project for Fundamental Research in China Grant NSC 92097.

The authors are with the National Laboratory of Machine Perception and Center for Information Science, Peking University, Beijing 100871, China.

Publisher Item Identifier S 1045-9227(96)06616-7.

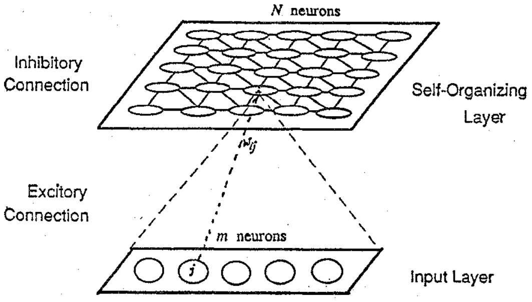

Fig. 1. The modified HME architecture.

specific problem of *sequence recognition.* As for a feature frame, it conveys the instantaneous spectral information which carries speaker related information correlated with talking behavior and physiological structure of the vocal tracts in additional to conveying phonetic information. On the other hand, successive feature frames convey the transitional information. Earlier studies have shown that both instantaneous and transitional spectral information is useful to speaker identification. Furthermore, the instantaneous and transitional spectral information is relatively uncorrelated, thus providing complementary information for speaker identification [12]. In most of connectionist approaches, however, feature frames belonging to an utterance are regarded as independent samples so that the instantaneous information is merely used for speaker identification. To use the transitional spectral information, some connectionist systems adopt the time-delay technique or recurrent neural models. Unlike the aforementioned connectionist approaches, we propose a modified HME architecture in this letter for text-dependent speaker identification. In the proposed architecture, the original HME architecture remains to deal with the instantaneous information based upon each feature frame of an utterance and a new gating network is added for use of both instantaneous and transitional spectral information. In the modified HME, the transitional information is utilized by performing a sequence recognition instead of identifying the unknown utterance merely by an individual feature frame in most of connectionist approaches [6]. Like the original HME [7], learning in the modified HME architecture is still treated as a maximum likelihood problem and we present an EM algorithm for adjusting the parameters of the proposed architecture. We have already applied the modified HME architecture to text-dependent speaker identification. The experimental results show that the system based upon the modified HME yields both satisfactory performance and significantly fast training.

The remainder of this letter is organized as follows. Section II describes the modified HME architecture and the EM algorithm. Section III reports experimental results. Conclusions are drawn in the final section.

II. The Modified HME Architecture and the EM Algorithm

A. The Modified HME Architecture

The modified HME architecture is based on the principle of divide-and-conquer in which a large, hard to solve problem is adaptively broken up into many, smaller, easier to solve problems. Illustrated in Fig. 1, the architecture is a tree in which the *gating networks* sit at the nonterminals of the tree. For the sample $\mathbf{X}$ consisting of $\{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$, these gating networks receive the vector $\mathbf{x}_t$ at time t as input and produce scalar outputs that are a partition of unity at each point in the input space. The *expert networks* sit at the leaves of the tree. Each expert produces an output vector for each input vector. These output vectors proceed up the tree, being blended by the gating network outputs. A new gating network called *S-gating network* is added and sits on the top of the original HME. It receives both $\mathbf{x}_t$ and $\mathbf{X}$ at time t and produces a weight for the output of the original HME. After all T results based upon all $\mathbf{x}_t$ in $\mathbf{X}$ are obtained, they are linearly combined to produce the final identifying result for $\mathbf{X}$. The basic idea underlying the S-gating network is that different feature frames belonging to an utterance may convey unequal speaker related information and the S-gating network is used to enhance the results produced based upon those feature frames reflecting more speaker related information. The use of the S-gating network may be also viewed as that the modified HME is capable of extracting speaker related feature furthermore during classification. In Fig. 1, there are only two levels in the architecture with 2-2; that is, there are two modules of mixtures of experts (ME) depicted in the blocks in Fig. 1 and two experts is in each ME module. To simplify the presentation, we restrict ourselves to a two-level hierarchy below. Obviously, the architecture can be generalized readily to hierarchies of arbitrary depth. For each vector $\mathbf{x}_t$ in $\mathbf{X}$, expert network (i, j) produces its output $\mathbf{O}_{ij}(\mathbf{x}_t)$ as a generalized linear function [7] of the input $\mathbf{x}_t$

$$\mathbf{O}_{ij}(\mathbf{x}_t) = f(W_{ij}\mathbf{x}_t) \qquad (1)$$

where W_{ij} is a weight matrix and $f(\cdot)$ is a fixed continuous nonlinearity. The ith output of the top-level gating network and outputs of the gating networks at lower levels are obtained by the "softmax" function, respectively,

$$g_i(\mathbf{x}_t) = \frac{e^{\mathbf{v}_i^T\mathbf{x}_t}}{\sum_k e^{\mathbf{v}_k^T\mathbf{x}_t}}, g_{j|i}(\mathbf{x}_t) = \frac{e^{\mathbf{v}_{ij}^T\mathbf{x}_t}}{\sum_k e^{\mathbf{v}_{ik}^T\mathbf{x}_t}} \qquad (2)$$

where $\mathbf{v}_i$ and $\mathbf{v}_{ij}$ are weight vectors, respectively. For the S-gating network, we use the Gaussian distribution to model the weighting factor of the result produced based on $\mathbf{x}_t$ since the conventional speech processing often makes an inaccurate assumption that successive observations (short-time frames of speech) are independent. The S-gating network produces its output $\lambda_{\mathbf{X}}(\mathbf{x}_t)$ for $\mathbf{x}_t$ in $\mathbf{X}$ as

$$\lambda_{\mathbf{X}}(\mathbf{x}_t) = \frac{P(\mathbf{x}_t, \Phi)}{\sum_{s=1}^{T} P(\mathbf{x}_s, \Phi)} \qquad (3)$$

where $P(\mathbf{x}_t, \Phi) = P(\mathbf{x}_t, \mathbf{m}, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp[-\frac{1}{2}(\mathbf{x}_t - \mathbf{m})^T\Sigma^{-1}(\mathbf{x}_t - \mathbf{m})]$. Note that the $\lambda_{\mathbf{X}}(\mathbf{x}_t)$ is positive and sum to one for all $\mathbf{x}_t$ in $\mathbf{X}$, i.e., $\sum_{t=1}^{T} \lambda_{\mathbf{X}}(\mathbf{x}_t) = 1$.

The modified hierarchy can be given a probabilistic interpretation. In the two levels architecture with M-N, for a paired observation $(\mathbf{X}, \mathbf{y})$ with $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$, the total probability of generating $\mathbf{y}$ from $\mathbf{X}$ can be described with a generalized finite mixture model as follows:

$$P(\mathbf{y}|\mathbf{X}, \Theta) = \sum_{t=1}^{T} \lambda_{\mathbf{X}}(\mathbf{x}_t, \Phi) \sum_{i=1}^{M} g_i(\mathbf{x}_t, \mathbf{v}_i) \sum_{j=1}^{N} g_{j|i}(\mathbf{x}_t, \mathbf{v}_{ij}) P(\mathbf{y}|\mathbf{x}_t, \theta_{ij}) \qquad (4)$$

where Θ is the set of all free parameters which include the expert network parameters θ_{ij}, the gating network parameters $\mathbf{v}_i$ and $\mathbf{v}_{ij}$ as well as the S-gating network parameters Φ. In the generalized finite mixture model, we interpret the value $\lambda_{\mathbf{X}}(\mathbf{x}_t, \Phi)$ as the probability that the result based on $\mathbf{x}_t$ is used for making the final decision in the sequence recognition for $\mathbf{X}$. The probabilistic interpretation of other components in the generalized finite mixture model is as same as ones of the original HME [7]. The classification in speaker identification is usually a specific multiway classification in which the output is a binary vector with a single nonzero component. As a result, instead of the *multinomial logit distribution* for the general multiway classification, we use a *generalized Bernoulli distribution* as the probabilistic model of expert networks in the general finite mixture model

$$P(y_1, y_2, \cdots, y_K) = \prod_{k=1}^{K} p_k^{y_k}(1-p_k)^{1-y_k}. \tag{5}$$

It can be shown that the generalized Bernoulli distribution is also a member of the exponential family and can be written in the following form:

$$P(y_1, y_2, \cdots, y_K) = \exp\{\sum_{k=1}^{K} y_k \ln \frac{p_k}{1-p_k} + \sum_{k=1}^{K} \ln(1-p_k)\}. \tag{6}$$

Accordingly, we may also give its *link function* $f(t)$ and *variance function* $Var(t)$ from (6) as

$$f(t) = \frac{1}{1+\exp(-t)}, Var(t) = f(t)[1-f(t)]. \tag{7}$$

In speaker identification, the major benefit of the generalized Bernoulli distribution is that it generates the same architecture of expert networks as the one generated by multinomial logit distribution but uses a more appropriate distribution to model the specific multiway classification. With respect to these two distributions used in speaker identification, the detailed analysis and the empirical investigation were presented in [11].

Suppose that all parameters have been already determined (the method of parameter estimation will be presented in the sequel.), the modified HME can be used for text-dependent speaker identification. For an unknown utterance, let us denote the set of its feature vectors as $\mathbf{X}_u = \{\mathbf{x}_1(u), \cdots, \mathbf{x}_{T_u}(u)\}$. For $\mathbf{X}_u$, the output of the modified HME as

$$\begin{aligned}&\mathbf{O}(\mathbf{X}_u)\\&= \sum_{t=1}^{T_u} \lambda_{\mathbf{X}_u}[\mathbf{x}_t(u)] \sum_{i=1}^{M} g_i[\mathbf{x}_t(u)] \sum_{j=1}^{N} g_{j|i}[\mathbf{x}_t(u)]\mathbf{O}_{ij}[\mathbf{x}_t(u)]\end{aligned} \tag{8}$$

where $\mathbf{O}_{ij}[\mathbf{x}_t(u)], g_i[\mathbf{x}_t(u)], g_{j|i}[\mathbf{x}_t(u)]$ and $\lambda_{\mathbf{X}_u}[\mathbf{x}_t(u)]$ are the outputs of expert networks, gating networks and the S-gating network, respectively, for $\mathbf{x}_t(u)$. Assume that the population in a speaker identification system is K, moreover, the output $\mathbf{O}(\mathbf{X}_u)$ is a K-dimensional vector, i.e., $\mathbf{O}(\mathbf{X}_u) = [O_1(\mathbf{X}_u), \cdots, O_K(\mathbf{X}_u)]$. Thus, we can identify the unknown speaker with $\mathbf{X}_u$ as speaker k^* according to the following decision rule:

$$k^* = \arg \max_{1 \le k \le K} O_k(\mathbf{X}_u). \tag{9}$$

B. The EM Algorithm

In text-dependent speaker identification, the data is assumed to form a countable set of P paired observations $\mathcal{X} = \{(\mathbf{X}(p), \mathbf{y}(p)); p = 1, \cdots, P\}$ where $\mathbf{X}(p) = \{\mathbf{x}_1(p), \cdots, \mathbf{x}_{T_p}(p)\}$. $\mathbf{X}(p)$ refers to the set of feature frames, say $\mathbf{x}_t(p)(t = 1, \cdots, T_p)$, corresponding to the pth utterance in the data set and T_p is the number of frames in $\mathbf{X}(p)$. It is worth noting that two utterances of the same text may have different numbers of frames due to changes in speaking rate. To develop an EM algorithm for the modified HME architecture, we must define appropriate "missing data" so as to simplify the likelihood function. We define indicator variables I_i, $I_{j|i}$, such that one and only one of the I_i is equal to one, and one and only one $I_{j|i}$ is equal to one. We also define the indicator variable I_{ij}, which is the product of I_i and $I_{j|i}$. These indicator variables have an interpretation as the labels that either correspond to decision of the probability model or specify the expert in the probability model. Moreover, we define indicator variables $I_t(\mathbf{X}_p)$, such that one and only one of $I_t(\mathbf{X}_p)$ is equal to one for $t = 1, \cdots, T_p$ in $\mathbf{X}_p$. We interpret these indicator variables as the labels that specify the frame $\mathbf{x}_t(p)$ in $\mathbf{X}_p$ in the probability model. Thus, the maximum likelihood estimation can be found iteratively using the EM algorithm as follows. Given the current estimate $\Theta^{(s)}$, each iteration consists of two steps, i.e., E-step and M-step. In order to simplify the maximization in the M-step, we first rewrite (4) using the trick in [13] into an equivalent form

$$\begin{aligned}P(\mathbf{y}, \mathbf{X}) &= P(\mathbf{y}|\mathbf{X}, \Theta)P(\mathbf{X}, \Phi)\\&= \sum_{t=1}^{T} P(\mathbf{x}_t, \Phi) \sum_{i=1}^{M} g_i(\mathbf{x}_t, \mathbf{v}_i) \sum_{j=1}^{N} g_{j|i}(\mathbf{x}_t, \mathbf{v}_{ij}) P(\mathbf{y}|\mathbf{x}_t, \theta_{ij})\end{aligned} \tag{10}$$

where $P(\mathbf{X}, \Phi) = \sum_{s=1}^{T} P(\mathbf{x}_s, \Phi)$. Instead of the probability model in (4), the probability model in (10) will be used for the maximimum likelihood estimation.

In the E-step, for each pair $(\mathbf{X}(p), \mathbf{y}_p)$ with $\mathbf{X}(p) = \{\mathbf{x}_1(p), \cdots, \mathbf{x}_{T_p}(p)\}(p = 1, \cdots, P)$, we compute posterior probabilities using Bayes' rule based upon the probability model (using the current estimate) incorporated with the "missing data" as follows:

$$\begin{aligned}&h_i^{(t)}(\mathbf{X}_p)\\&= \frac{g_i(\mathbf{x}_t(p), \mathbf{v}_i^{(s)}) \sum_{j=1}^{N} g_{j|i}(\mathbf{x}_t(p), \mathbf{v}_{ij}^{(s)}) P(\mathbf{y}_p|\mathbf{x}_t(p), \theta_{ij}^{(s)})}{\sum_{i=1}^{M} g_i(\mathbf{x}_t(p), \mathbf{v}_i^{(s)}) \sum_{j=1}^{N} g_{j|i}(\mathbf{x}_t(p), \mathbf{v}_{ij}^{(s)}) P(\mathbf{y}_p|\mathbf{x}_t(p), \theta_{ij}^{(s)})}\end{aligned} \tag{11}$$

$$h_{j|i}^{(t)}(\mathbf{X}_p) = \frac{g_{j|i}(\mathbf{x}_t(p), \mathbf{v}_{ij}^{(s)}) P(\mathbf{y}_p|\mathbf{x}_t(p), \theta_{ij}^{(s)})}{\sum_{j=1}^{N} g_{j|i}(\mathbf{x}_t(p), \mathbf{v}_i^{(s)}) P(\mathbf{y}_p|\mathbf{x}_t(p), \theta_{ij}^{(s)})} \tag{12}$$

$$\begin{aligned}&h_{ij}^{(t)}(\mathbf{X}_p)\\&= h_i^{(t)}(\mathbf{X}_p) \cdot h_{j|i}^{(t)}(\mathbf{X}_p)\\&= \frac{g_i(\mathbf{x}_t(p), \mathbf{v}_i^{(s)}) \sum_{j=1}^{N} g_{j|i}(\mathbf{x}_t(p), \mathbf{v}_{ij}^{(s)}) P(\mathbf{y}_p|\mathbf{x}_t(p), \theta_{ij}^{(s)})}{\sum_{i=1}^{M} g_i(\mathbf{x}_t(p), \mathbf{v}_i^{(s)}) \sum_{j=1}^{N} g_{j|i}(\mathbf{x}_t(p), \mathbf{v}_{ij}^{(s)}) P(\mathbf{y}_p|\mathbf{x}_t(p), \theta_{ij}^{(s)})}\end{aligned} \tag{13}$$

and (14), shown at the bottom of the page.

$$h_t(\mathbf{X}_p) = \frac{\lambda_{\mathbf{X}_p}(\mathbf{x}_t(p), \Phi^{(s)}) \sum_{i=1}^{M} g_i(\mathbf{x}_t(p), \mathbf{v}_i^{(s)}) \sum_{j=1}^{N} g_{j|i}(\mathbf{x}_t(p), \mathbf{v}_{ij}^{(s)}) P(\mathbf{y}_p|\mathbf{x}_t(p), \theta_{ij}^{(s)})}{\sum_{t=1}^{T} \lambda_{\mathbf{X}_p}(\mathbf{x}_t(p), \Phi^{(s)}) \sum_{i=1}^{M} g_i(\mathbf{x}_t(p), \mathbf{v}_i^{(s)}) \sum_{j=1}^{N} g_{j|i}(\mathbf{x}_t(p), \mathbf{v}_{ij}^{(s)}) P(\mathbf{y}_p|\mathbf{x}_t(p), \theta_{ij}^{(s)})} \tag{14}$$

TABLE I
THE IDENTIFYING ACCURACIES (%) IN TEST-1

Text	'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'	mean
MLP	97.14	97.62	98.10	97.14	93.33	96.19	96.19	96.67	93.81	97.14	96.33
HME	99.05	99.52	98.57	95.24	93.80	98.57	98.10	98.57	96.67	98.10	97.62
TD-HME[1]	99.52	99.52	98.57	99.05	96.67	98.57	99.05	98.57	95.24	98.57	98.33
TD-HME[2]	100.0	100.0	98.57	97.62	96.67	99.05	98.10	99.05	96.67	99.05	98.78
Modified HME	99.52	99.52	99.05	99.05	95.24	98.57	98.57	99.05	97.62	98.10	98.43

TABLE II
THE IDENTIFYING ACCURACIES (%) IN TEST-2

Text	'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'	mean
MLP	88.0	90.0	87.0	85.0	82.0	83.0	84.0	83.0	77.0	89.0	84.8
HME	91.0	96.0	87.0	86.0	83.0	90.0	85.0	89.0	86.0	91.0	88.4
TD-HME[1]	92.0	96.0	89.0	88.0	84.0	91.0	90.0	88.0	85.0	92.0	89.5
TD-HME[2]	91.0	95.0	91.0	85.0	86.0	88.0	93.0	93.0	85.0	92.0	89.9
Modified HME	92.0	96.0	90.0	87.0	86.0	93.0	92.0	90.0	86.0	92.0	90.4

In the M-step, the following separate maximization problems need solving:

$$\theta_{ij}^{(s+1)} = \arg\max_{\theta_{ij}} \sum_{p=1}^{P}\sum_{t=1}^{T_p} h_{ij}^{(t)}(\mathbf{X}_p)\log P(\mathbf{y}_p|\mathbf{x}_t(p),\theta_{ij}) \qquad (15)$$

$$\mathbf{v}_i^{(s+1)} = \arg\max_{\mathbf{v}_i} \sum_{p=1}^{P}\sum_{t=1}^{T_p}\sum_{k=1}^{M} h_{k}^{(t)}(\mathbf{X}_p)\log g_k(\mathbf{x}_t(p),\mathbf{v}_k) \qquad (16)$$

$$\mathbf{v}_{ij}^{(s+1)} = \arg\max_{\mathbf{v}_{ij}} \sum_{p=1}^{P}\sum_{t=1}^{T_p}\sum_{k=1}^{M} h_{k}^{(t)}(\mathbf{X}_p)\sum_{l=1}^{N} h_{l|k}(\mathbf{X}_p)\log g_{l|k}(\mathbf{x}_t(p),\mathbf{v}_{lk}) \qquad (17)$$

and

$$\Phi^{(s+1)} = \arg\max_{\Phi} \sum_{p=1}^{P}\sum_{t=1}^{T_p} h_t(\mathbf{X}_p)\log P(\mathbf{x}_t(p),\Phi). \qquad (18)$$

Problems in (15)–(17) belong to *iteratively reweighted least squares* (IRLS) problems. They can be solved by using the IRLS algorithm[1] in [7]. Thanks to the use of (10), the problem in (18) is analytically solvable as follows:

$$\mathbf{m}^{(s+1)} = \frac{1}{\sum_{p=1}^{P}\sum_{t=1}^{T_p} h_t(\mathbf{X}_p)} \sum_{p=1}^{P}\sum_{t=1}^{T_p} h_t(\mathbf{X}_p)\mathbf{x}_t(p) \qquad (19)$$

$$\mathbf{\Sigma}^{(s+1)} = \frac{1}{\sum_{p=1}^{P}\sum_{t=1}^{T_p} h_t(\mathbf{X}_p)} \sum_{p=1}^{P}\sum_{t=1}^{T_p} h_t(\mathbf{X}_p)\left[\mathbf{x}_t(p)-\mathbf{m}^{(s+1)}\right]\left[\mathbf{x}_t(p)-\mathbf{m}^{(s+1)}\right]^T. \qquad (20)$$

III. EXPERIMENTAL RESULTS

We used a 10-speaker (10 male) isolated digit database in this study. The database consists of 10 isolated digits from zero to nine uttered in Chinese. For each speaker, 300 utterances were recorded (30 utterances/digit). The 300 utterances were equally divided and recorded in three different sessions over about a two-month period. The technical details of the acoustic preprocessing and feature extraction are briefly described as follows: 1) 16-bit A/D-converter with 11.025 KHz sampling rate; 2) processing the data with a preemphasis filter $H(z) = 1 - 0.95z^{-1}$; 3) 16-order linear predictive coding (LPC) analysis; 4) 256-point LPC-based fast Fourier transform (FFT) formed every 12.8 ms using a Hamming window; 5) combination of spectral channels from 0 Hz to 5.0125 KHz into a 24-component feature vector; 6) subtraction of the average from the components; and 7) normalization of the feature vectors. Depending upon the fixed text (10 isolated digits), 10 modified HME's are used so that 10 modified HME classifiers correspond to 10 digits from zero to nine, respectively. The *predetermined structure* problem refers to that for a given task an appropriate structure of neural network must be determined before training. The use of the modified HME also encounters the problem. We have applied the two-fold *cross-validation* method to remedy the problem. Using the method, we have investigated seven structures covering from one level to three levels. According to the performance and training time, we have finally chosen a three levels modified HME with 2-2-10 as the classifier.

In order to evaluate the effectiveness of the modified HME in text-dependent speaker identification, we adopted the same experimental method described in [12]. As a result, we divided all utterances corresponding to every digit in the database into two sets; 90 utterances (30 utterances/session and three utterances/speaker per session) were randomly chosen as the training set and the other 210 utterances were used as the testing set. For test, we used a so-called *digit-based* method in which we merely used an utterance of a single digit to identifying the unknown speaker. We denote the test using the data in just mentioned testing set as TEST-1 and the experimental results in TEST-1 are shown in Table I. For the purpose of comparison, some results using the MLP, the original HME [9] and time-delay HME's [10], [11] are also listed in Table I. In Table I, TD-HME[i] ($i = 1, 2$) denotes a time-delay HME in which the size of the input window is $i + 1$ frames. According to results shown in Table I, both the modified HME and time-delay HME's outperform the MLP and the original HME. It is worth noting that both the original HME and the modified HME took almost the same time for training (four or five epochs) but the training of TD-HME [2] took much longer time than one of the modified HME (about three times).

In the practical application of speaker identification, only acoustic

[1] Due to a limit of space, the IRLS algorithm is not summarized here and readers are referred to the detailed description of the IRLS algorithm in [7, Appendix A].

TABLE III
THE IDENTIFYING ACCURACIES (%) IN TEST-3

Text	'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'	mean
MLP	85.0	81.0	87.0	79.0	85.0	77.0	88.0	90.0	77.0	88.0	83.7
HME	89.0	95.0	88.0	85.0	80.0	90.0	84.0	91.0	82.0	90.0	87.4
TD-HME[1]	87.0	95.0	91.0	84.0	82.0	88.0	87.0	88.0	86.0	89.0	87.7
TD-HME[2]	89.0	95.0	88.0	86.0	85.0	92.0	86.0	87.0	87.0	89.0	88.4
Modified HME	90.0	97.0	89.0	85.0	82.0	90.0	86.0	91.0	85.0	91.0	88.6

TABLE IV
EXPERIMENTAL RESULTS BASED ON THE MODIFIED HME USING THE SEQUENCE-BASED METHOD

Test No.	500	1000	2000	3000	4000	5000
Recognition No.	500	999	1997	2995	3993	4989
Substitution No.	0	0	1	1	2	4
Rejection No.	0	1	2	4	5	7

data recorded in finite sessions are available for training a system. In fact, a speaker's voices largely depend upon the characteristics of the speaker's vocal tracts, his/her mood, and the environment where the speaker stays. Although acoustic data recorded in several sessions can be used for training, an utterance recorded beyond those sessions may often make the performance of the system degrade. For robustness, the performance of a speaker identification system is often evaluated by using utterances recorded beyond the sessions for training [1], [3], [4]. For evaluating robustness of the system based upon the modified HME, we have already done another experiment. In the experiment, for a digit, five utterances of each speaker recorded in the first session were merely used as the training data and all utterances of the digit recorded in other two sessions were used as the testing data. Tests using utterances recorded in the second and the third session are called TEST-2 and TEST-3, respectively. Accordingly, these experimental results are, respectively, shown in Table II and Table III. For the purpose of comparison, we also list results of the MLP, the original HME [9] and time-delay HME's [11]. According to results in Table II and Table III, it is evident that the system using the modified HME outperforms ones using the MLP, the original HME and time-delay HME's. Moreover, we also used a so-called *sequence-based* method to evaluate the performance of the modified HME's trained using the just mentioned training set. In the method, we first produced a sequence consisting of five digits at random (it may be viewed as a password), then asked a speaker to utter the digit sequence. For each digit in the sequence, obviously, an identifying result was available based upon the digit-based method. After obtaining all five individual results, the system tolled a vote with the principle of majority that an unknown speaker can be identified only when there are at least three same identification results for the speaker; otherwise, the system rejects the unknown utterance. In the experiment, for each test, we randomly selected five utterances (belonging to the same speaker and, respectively, corresponding to five digits in the prompted digit sequence) from the data in the testing set. For the modified HME, experimental results using the sequence-based method are shown in Table IV.

IV. CONCLUSIONS

We have described a modified HME architecture for text-dependent speaker identification. In the proposed architecture, a new gating network is added for weighting results produced by the original HME based on each feature frame of an utterance and performing the identification with the linear combination of weighted results. A generalized finite mixture model has been proposed for the architecture and an EM algorithm has also been presented for adjusting parameters of the proposed architecture. Experimental results demonstrate the effectiveness of the modified HME in text-dependent speaker identification.

ACKNOWLEDGMENT

The authors are very grateful to two anonymous referees for their useful comments, particularly the referee who pointed out a mistake in an earlier version of this paper.

REFERENCES

[1] G. R. Doddington, "Speaker recognition— Identifying people by their voices," *Proc. IEEE*, vol. 74, no. 11, pp. 1651–1663, Nov. 1986.

[2] B. Atal, "Automatic recognition of speakers from their voices," *Proc. IEEE*, vol. 64, no. 4, pp. 460–475, Apr. 1976.

[3] D. O'Shaugenessy, "Speaker recognition," *IEEE Acoust., Speech, Signal Processing Mag.*, vol. 3, pp. 4–17, 1986.

[4] S. Furui, "An overview of speaker recognition technology," in *Proc. ESCA Wkshp. Automat. Speaker Recognition, Identification, and Verification*, Martigny, Switzerland, 1994, pp. 1–9.

[5] T. Matsui and S. Furui, "Speaker recognition technology," *NTT Rev.*, vol. 7, no. 2, pp. 42–48, 1995.

[6] Y. Bennani and P. Gallinari, "Connectionist approaches for automatic speaker recognition," in *Proc. ESCA Wkshp. Automat. Speaker Recognition, Identification, and Verification,* Martigny, Switzerland, 1994, pp. 95–102.

[7] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and EM algorithm," *Neural Computa.*, vol. 6, no. 2, pp. 181–214, 1994.

[8] M. I. Jordan and L. Xu, "Convergence results for the EM approach to mixtures of experts," *Neural Networks*, vol. 8, no. 9, pp. 1409–1431, 1995.

[9] K. Chen, D. Xie, and H. Chi, "Speaker identification based on hierarchical mixtures of experts," in *Proc. WCNN,* Washington, D.C., July 1995, pp. I493–I496.

[10] ——, "Speaker identification based on time-delay HME," in *Proc. IEEE ICNN*, Perth, Australia, Dec. 1995, pp. 2062–2066.

[11] ——, "Speaker identification using time-delay HME's," *Int. J. Neural Syst.*, vol. 7, no. 1, pp. 29–43, Mar. 1996.

[12] F. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 36, pp. 871–879, June 1988.

[13] L. Xu, M. I. Jordan, and G. E. Hinton, "A modified gating network for the mixtures of experts architecture," in *Proc. WCNN*, San Diego, CA, June 1994, pp. II405–II409.