International Journal of Neural Systems, Vol. 7, No. 1, pp. 29-43, 1996.
World Scientific Publisher Co.

Speaker Identification Using Time-Delay HMEs

KE CHEN[1], DAHONG XIE and HUISHENG CHI
National Laboratory of Machine Perception and Center for Information Science
Peking University, Beijing 100871, China

Abstract

In this paper, we extend the Hierarchical Mixtures of Experts (HME) to temporal processing and explore it for a substantial problem, that of text-dependent speaker identification. For a specific multiway classification, we propose a *generalized Bernoulli density* instead of the *multinomial logit density* to avoid the instability during training. Time-delay technique is applied for spatio-temporal processing in the HME and a combining scheme is presented for combining multiple time-delay HMEs in order to complete multi-scale analysis for the temporal data. Using the time-delay HME along with the EM algorithm as well as the combination of multiple time-delay HMEs, the speaker identification system has a good performance and yields significantly fast training. We have also addressed some issues about the time-delay techniques in the HME.

1 Introduction

Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. This technique will make it possible to verify the identity of persons accessing systems, that is, access control by voice, in various services. These services include banking transaction over a telephone network, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers [1]. From the viewpoint of technology, speaker recognition is a general term which refers to any task to discriminate people based upon their voice characteristics [2]. Within this general task description, there are two specific tasks that have been studied extensively. These are referred to as *speaker identification* and *speaker verification*. The objective of speaker identification is to determine which speaker is present based on the individual's utterance, whereas the speaker verification task is to verify the person's claimed identity. Speaker identification systems can be either *text-dependent* or *text-independent*. In this paper, only text-dependent speaker identification is considered. By text-dependent, we mean that the text in both training and testing 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 testing. Moreover, speaker identification can be subdivided into two further categories, *closed-set* and *open-set* problems [3]. The closed set problem is to identify a speaker from a group of N known speakers. Alternatively, one may want to decide whether the speaker of a test utterance belongs to a group of N known speakers. This is called the open-set problem since the speaker to be identified may not be one of the N known speakers. In this paper, the closed-set problem is simply considered. In general, the technique of speaker identification includes feature extraction and classification. There have been extensive studies in this field based upon conventional techniques of speech signal processing [4]. Employed as the classifier, recently, many kinds of neural networks have been adopted for speaker recognition [5], such as MLP [6], neural tree [7] and a hybrid model [8] etc. Unfortunately, the systems based on neural networks often suffer from a high computational burden during training and have only a limited improvement over conventional techniques in performance.

[1]Corresponding author. E-mail: chen@cis.pku.edu.cn

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) [9] is just a modular architecture in which the outputs of a number of modular nets called 'experts', each mapping a particular portion of the input space, are combined in a probabilistic way by 'gating' net which is modeling the probability that each portion of the input space generated the output. The HME has been successful in a number of regression and some classification problems [9]-[11], yielding significantly faster training through the use of the Expectation Maximization (EM) algorithm. In addition, it is also applied successfully to the non-linear prediction of acoustic vectors for speech processing [12]. In our previous work, we have already applied HME along with EM algorithm to cope with the text-dependent speaker identification [13]. In fact, many real-world applications require the processing of patterns that evolve over time and speech processing is a typical case. However, all aforementioned applications of HME architecture simply consider patterns as static ones. For the current task of speaker identification, the temporal characteristics of patterns play an important role in the final recognition results. Time-Delay Neural Network (TDNN) architecture was originally designed for speech recognition [14] and has strong abilities to spatio-temporal processing. Motivated by the TDNN architecture, in the paper we introduce the time-delay concept to HME architecture. For the time-delay, a time window with the fixed size should be chosen in advance [14]. In general, the larger window may capture more temporal information, but it suffers from an expensive computational cost. Due to the lack of understanding in features of speaker's identity, moreover, it is rather hard for us to choose an appropriate time window. This problem has already been encountered in our previous work [15]. To handle the problem, in this paper, we propose a combining scheme in which multiple time-delay HMEs with different short-term window sizes are combined under the framework of Bayesian formalism, while this idea was originally adopted in combining multiple classifiers to achieve the better performance of classification [16]. On the other hand, we also extend the HME model to a specific multiway classification base on a proposed probability density called *generalized Bernoulli density* instead of *multinomial logit density* to improve the performance of the HME and yield fast training. All aforementioned techniques have already been applied to the closed-set text-dependent speaker identification.

The remainder of this paper is organized as follows. Section 2 reviews the HME architecture with the use of EM algorithm and presents the *generalized Bernoulli density*. Section 3 introduces the time-delay concept to the HME architecture and propose a method to combine multiple time-delay HMEs. In section 4, an overview of the speaker identification system based the HME is addressed. Experimental results and discussions are proposed in section 5 and section 6, respectively. Conclusions are drawn in the final section.

2 Review of HME and Generalized Bernoulli Density

2.1 The HME Architecture

The HME 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. Unlike other systems with 'hard' decisions to partition the input space, the use of the gating net in the HME overcomes this limitation and allows adjacent clusters in the input space to overlap. A typical HME architecture is illustrated in Fig. 1. The architecture is a tree in which the *gating* nets sit at the nonterminals of the tree. These nets receive the vector $\mathbf{x}$ as input and produce scalar outputs that are a partition of unity at each point in the input space. The *expert* nets 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 net outputs [9]. In Fig. 1, there are only two levels in the architecture with

2-2[2]. In general, it is easy to extend this architecture to multiple levels.

In the HME, we denote the output of expert net (i, j) as μ_{ij}, the ith output of the top-level gating net as g_i and the jth output of the ith lower-level gating net as $g_{j|i}$, respectively. Accordingly, expert net (i, j) produces an output as a generalized linear function, hereafter called *link function* [17], of the input $\mathbf{x}$: $\mu_{ij} = f(\mathbf{W}_{ij}\mathbf{x})$, where $\mathbf{W}_{ij}$ is a weight matrix and f is a fixed continuous nonlinearity. Gating nets produce outputs of the input $\mathbf{x}$ based upon the *softmax* function[3] as follows: $g_i = \frac{\exp(\mathbf{v}_i^T\mathbf{x})}{\sum_k \exp(\mathbf{v}_k^T\mathbf{x})}$ and $g_{j|i} = \frac{\exp(\mathbf{v}_{ij}^T\mathbf{x})}{\sum_k \exp(\mathbf{v}_{ik}^T\mathbf{x})}$ where $\mathbf{v}_i$ and $\mathbf{v}_{ij}$ are weight vectors, respectively. For the two levels architecture, a probabilistic description of the HME is as follows

$$P(\mathbf{y}|\mathbf{x}, \theta) = \sum_i g_i(\mathbf{x}, \mathbf{v}_i) \sum_j g_{j|i}(\mathbf{x}, \mathbf{v}_{ij}) P(\mathbf{y}|\mathbf{x}, \theta_{ij}) \tag{1}$$

where θ_{ij} and θ are free parameters in expert net (i, j) and the set of all free parameters in the model, respectively. $P(\mathbf{y}|\mathbf{x}, \theta_{ij})$ is the probabilistic component of the HME model and can be generally modeled as a density in the exponential family [17] with the following form:

$$P(\mathbf{y}, \eta, \phi) = \exp\{\frac{\eta\mathbf{y} - b(\eta)}{\phi} + c(\mathbf{y}, \phi)\} \tag{2}$$

where η is called the *natural parameter* and ϕ is the *dispersion parameter* in Generalized Linear Model (GLIM) [17]. Accordingly, the link function $f(\eta) = b'(\eta)$ is known as the *cononical link* and $Var(\mathbf{y}) = b''(\eta)\phi$ is known as the *variance function* in the GLIM theory, respectively.

2.2 Expectation-Maximization (EM) Algorithm

The EM algorithm is a general technique for maximum likelihood estimation [20]. Assume that X_{obs} denotes all observable data and a likelihood function based on the data is $l(\theta; X_{obs})$. An application of the EM algorithm is to simplify the optimization of the likelihood function $l(\theta; X_{obs})$ by introducing some additional unobservable data called *missing* data, X_{mis}, to the original data in order to generate a new likelihood function $l_c(\theta; X_{obs}, X_{mis})$. In this context, the original likelihood, $l(\theta; X_{obs})$, is referred to as *incomplete-data likelihood* and the new likelihood, $l_c(\theta; X_{obs}, X_{mis})$, is referred to as *complete-data likelihood*. Obviously, the *missing* data is fictive and in fact unknown so that the *complete-data likelihood* is a random variable. As a result, for maximum likelihood estimation, the EM algorithm consists of two alternate steps, i.e. E-step and M-step, as follows,

E-step

$$Q(\theta, \theta^{(p)}) = E[l_c(\theta; X_{obs}, X_{mis})|X_{obs}]$$

where $\theta^{(p)}$ is the value of the parameters at the pth iteration and the expectation is taken with respect to $\theta^{(p)}$. This step yields a deterministic function Q.

M-step

$$\theta^{(p+1)} = \arg\max_\theta Q(\theta, \theta^{(p)})$$

This step maximizes the function with respect to θ to find the new parameter estimates $\theta^{(p+1)}$. It has already been shown that an iterative step of EM chooses a parameter value which increases

[2]This means that the HME consists of two modules of mixture-of-experts (ME) illustrated in the blocks in the Fig. 1 and there are two experts in each ME module.

[3]Some modified gating nets have already been presented and readers may be referred to the work [18][19].

the value of Q, the expectation of the *complete-data likelihood*. Accordingly, an increase in Q implies an increase in the *incomplete-data likelihood* [20]:

$$l(\theta^{(p+1)}; X_{obs}) \geq l(\theta^{(p)}; X_{obs})$$

For training the HME, there are two methods [9], i.e. gradient-based and EM algorithms, while the performance of the EM algorithm is usually much better than one of the gradient-based algorithm [9]-[13][15][21]. Jordan and Jacobs have already applied the EM algorithm to the HME architecture [9] by introducing indicator variables as *missing* data to the original observable data for simplifying the original likelihood function only with the observable data. The indicators may be explained as the labels that correspond to the decisions or specify the expert in the probability model [9]. Therefore, if we simplify the notation in Eq.(1) as $P(\mathbf{y}|\mathbf{x},\theta) = \sum_i g_i \sum_j g_{j|i} P_{ij}(\mathbf{y})$, the E-step of the EM algorithm by taking the expectation of the *complete-data likelihood* as follows:

$$Q(\theta, \theta^{(p)}) = \sum_t \sum_i \sum_j h_{ij} \{\ln g_i^{(t)} + \ln g_{j|i}^{(t)} + \ln P_{ij}(\mathbf{y}^{(t)})\} \tag{3}$$

where $h_i = \frac{g_i \sum_j g_{j|i} P_{ij}(\mathbf{y})}{\sum_i g_i \sum_j g_{j|i} P_{ij}(\mathbf{y})}$, $h_{j|i} = \frac{g_{j|i} P_{ij}(\mathbf{y})}{\sum_j g_{j|i} P_{ij}(\mathbf{y})}$ and $h_{ij} = h_i \cdot h_{j|i}$.

The M-step of the EM algorithm requires maximizing $Q(\theta, \theta^{(p)})$ with respect to all parameters. According to Eq.(3), thus, the M-step consists of the following separate maximization problems:

$$\theta_{ij}^{(p+1)} = \arg\max_{\theta_{ij}} \sum_t h_{ij}^{(t)} \ln P_{ij}(\mathbf{y}^{(t)}) \tag{4}$$

$$\mathbf{v}_i^{(p+1)} = \arg\max_{\mathbf{v}_i} \sum_t \sum_k h_k^{(t)} \ln g_k^{(t)} \tag{5}$$

and

$$\mathbf{v}_{ij}^{(p+1)} = \arg\max_{\mathbf{v}_{ij}} \sum_t \sum_k h_k^{(t)} \sum_l h_{l|k}^{(t)} \ln g_{l|k}^{(t)} \tag{6}$$

Each of these maximization problems is itself a maximum likelihood problem. All of these optimal problems belong to *Iteratively Reweighted Least Squares*(IRLS) problems. Since all components of the HME architecture are based upon the GLIM theory which provides the basic statistic structure for the HME, the likelihood is a product of densities from the exponential family of distributions. It may be solved by using the *Fisher scoring* algorithm [17]. Using the Fisher scoring algorithm, Jordan and Jacobs derive a particular iterative algorithm called IRLS algorithm for computing a maximum likelihood estimate of the parameters of the HME.[4]. Here, we summarize the algorithm as a general updated formula for IRLS problems as follows:

$$\beta_{r+1} = \beta_r + \eta(\mathbf{X}^T \mathbf{U} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{U} \mathbf{e} \tag{7}$$

where η is the learning rate ($0 < \eta \leq 1$). β_r is the value of the parameter β at the rth iteration in the M-step. $\mathbf{U}$ is a diagonal matrix whose diagonal elements are $\frac{c^{(i)}[f'(\beta^T \mathbf{x}^{(i)})]^2}{Var(\mathbf{y}^{(i)})}$ and $\mathbf{e} = [\frac{y^{(1)} - f(\beta^T \mathbf{x}^{(1)})}{f'(\beta^T \mathbf{x}^{(1)})}, \ \ldots \ , \frac{y^{(N)} - f(\beta^T \mathbf{x}^{(N)})}{f'(\beta^T \mathbf{x}^{(N)})}]$ if we assume $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$ is the ith observation data pair of the training set with N samples and $c^{(i)}$ is an observation weight accordingly. As a result, we may summarize the EM algorithm for the HME architecture as follows:

Algorithm: (Expectation-Maximization for the HME)

[4]For details, readers are referred to Appendix A in [9] for the complete derivation of the algorithm.

1. **E-step**
   For all data pairs $\{(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\}_1^N$, compute the posterior probabilities $h_i^{(t)}$ and $h_{j|i}^{(t)}$.

2. **M-step**

   (a) For each expert (i, j), solve the IRLS problem in Eq.(4) using Eq.(7) with observations $\{(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\}_1^N$ and observation weights $\{h_{ij}^{(t)}\}_1^N$, where $h_{ij}^{(t)} = h_i^{(t)} h_{j|i}^{(t)}$.

   (b) For each top-level gating network, solve the IRLS problem in Eq.(5) using Eq.(7) with observations $\{(\mathbf{x}^{(t)}, h_i^{(t)})\}_1^N$. And observation weights are $\{h_k^{(t)}\}_1^N$.

   (c) For each lower-level gating network, solve the weighted IRLS problem in Eq.(6) using Eq.(7) with observations $\{(\mathbf{x}^{(t)}, h_{j|i}^{(t)})\}_1^N$ and observation weights $\{h_k^{(t)}\}_1^N$, $\{h_{l|k}^{(t)}\}_1^N$.

In this algorithm, each complete update including both the E-step and the M-step. The training is not completed until a pre-specified 'stop' condition is satisfied.

2.3 Generalized Bernoulli Density

Consider a special multiway classification problem in which the output is binary($y_i \in \{0, 1\}$) with a single non-zero component, we may define a *generalized Bernoulli density* as follows

$$P(y_1, y_2, ..., y_K) = \prod_{k=1}^{K} p_k^{y_k} (1 - p_k)^{1 - y_k} \tag{8}$$

We show the *generalized Bernoulli density* is also a member of the exponential family as follows,

$$\begin{aligned} P(y_1, y_2, ..., y_K) &= \exp\{\ln \prod_{k=1}^{K} p_k^{y_k} (1 - p_k)^{1 - y_k}\} \\ &= \exp\{\sum_{k=1}^{K} \ln(p_k^{y_k} (1 - p_k)^{1 - y_k})\} \\ &= \exp\{\sum_{k=1}^{K} (y_k \ln p_k + (1 - y_k) \ln(1 - p_k))\} \\ &= \exp\{\sum_{k=1}^{K} y_k \ln \frac{p_k}{1 - p_k} + \sum_{k=1}^{K} \ln(1 - p_k)\} \end{aligned} \tag{9}$$

Accordingly, we may also derive its *link* and *variance* functions [17] from Eq.(2) and Eq.(4) as $f(t) = \frac{1}{1+\exp(-t)}$ and $Var(t) = f(t)(1 - f(t))$.

For a general multiway classification, the *multinomial logit density* has already been chosen in the original HME model [9]. From the viewpoint of neural computing, when both the *multinomial logit density* and the *generalized Bernoulli density* are used as the probabilistic model of an expert net, the difference between them lies in that they define different activation functions for the expert net. Accordingly, the activation function of each output neuron of the expert net is the *sigmoid* function if the *generalized Bernoulli density* is employed, while the activation function of each output neuron of the expert net is the *softmax* function if the *multinomial logit density* is employed. Obviously, the

aforementioned specific multiway classification definition provides a very convenient representation for an identification problem. For such a multiway classification, unfortunately, the HME with the *multinomial logit density* suffers from instability during training in our experiments so that the training procedure either takes quite a long time or often cannot reach a steady state probably due to the exponential operation and accumulated errors in the numerical calculations. As a result, we adopt the proposed *generalize Bernoulli density* as the probabilistic model of expert nets for speaker identification.

3 Time-Delay HMEs And Their Combination

In this section, we shall introduce the time-delay concept to the HME and propose a combining scheme for integrating multiple time-delay HMEs with different time windows to complete multiscale analysis for the temporal data.

3.1 Time-Delay in The HME Architecture

he basic principle of spatio-temporal processing with time-delay is to make the decision of the neural network at time *t* based on the inputs at time $(t-n), (t-n+1), \cdots, t$. This way, a well defined history of the temporal sequence is considered. For the use of time-delay techniques, a fixed size window, hereafter called *input-window*, is first chosen for catching the temporal features from the input patterns over time. The primary goal is to design an architecture that provides non-linear classification invariant under translation in time. This can be realized by sliding the input-window along the temporal input patterns. For speech processing, the temporal input patterns are some successive frames of acoustic data corresponding to an utterance after preprocessing. In this paper, we introduce this technique to HME architecture and Fig. 2 illustrates the input of component nets, i.e. *gating* and *expert* nets, in the time-delay HME architecture. For an utterance, after preprocessing and feature extraction, it can be denoted as $\mathbf{x} = \{\mathbf{f}^{(1)}, \cdots, \mathbf{f}^{(T)}\}$ where $\mathbf{f}^{(t)}$ is the frame at time *t*. If the time-delay processing occurs at time *t*, the new feature vector $\mathbf{x}_t^{(n)}$ will be composed of $n+1$ successive feature vectors $\mathbf{f}^{(t)}, \mathbf{f}^{(t-1)}, \cdots, \mathbf{f}^{(t-n)}$ subject to $\mathbf{x}_t^{(n)} = \{\mathbf{f}^{(t)}, \mathbf{f}^{(t-1)}, \cdots, \mathbf{f}^{(t-n)}\}$ instead of the previous feature vector $\mathbf{f}^{(t)}$ if we choose the input-window size as n.

It is worth pointing out that the time-delay used in the HME is slightly different from the original one in the time-delay neural network (TDNN) architecture [14] due to the different architectures. For the TDNN architecture, it employs a multilayer perceptron structure. Accordingly, the time-delay processing occurs between any two adjacent layers. For the HME, each component net can be modeled as a specific probability density belonging to the exponential family based upon GLIM theory [9][17]. Thus, it results in that each component net is simply a net without any hidden layer and the activation function of a neuron in the output layer is uniquely determined by the link function corresponding to the specific probability density. In addition, the HME is a modular neural network and the time-delay processing merely occurs between input and output layers of each component net in the HME.

3.2 The Scheme of Combining Multiple Time-Delay HMEs

For the time-delay HME architecture, there is a problem of the input-window size to be open for solution. As mentioned above, an input window must be chosen in advance for the time-delay processing. In general, the larger window may capture more temporal information, but it suffers from an expensive computational cost. The problem involves in a dilemma on the window size and the computational cost. Actually, it is also quite hard for us to determine which one is more

efficient for two short input-windows, such as $n = 1$ and $n = 2$. In fact, the efficiency of a short time window strongly depends upon the intrinsic statistical characteristics of the processed temporal input data. Usually, the knowledge of the intrinsic statistical characteristics is also hard to be available. In particular, the exact understanding of dependence among several successive frames for identifying a speaker is still not available. Here, we propose a method to attack the the aforementioned dilemma. The basic idea is as follows, first, we choose several short input-windows. Then we train several time-delay HMEs with these chosen different input-windows, respectively. After finishing the training, indeed, we may achieve several results with these trained time-delay HMEs for a given test pattern during test. Using an elaborate combining scheme, we can draw the integrated result based upon individual results. This idea has already been extensively used in combining multiple classifiers to improve the performance of individual classifiers for the development of highly reliable character recognition systems [16][22]-[25]. Here, we adopt the combining scheme in Bayesian formalism proposed in [16] to combine multiple time-delay HMEs with different short time windows. Fig. 3 illustrates such a scheme in which $n + 1$ time-delay HMEs($n > 0$), labeled as TD-HME[i] ($i = 0, 1, \cdots, n$), are combined. Here, TD-HME[i] represents the time-delay HME which uses an input-window with the size of i for temporal processing. In particular, TD-HME[0] denotes the HME without time-delay when $i = 0$.

In the sequel, we describe a combining scheme based upon the Bayesian formalism for combining multiple time-delay HMEs. This scheme was originally proposed for combining multiple classifiers with the static input [16]. Here, we extend it to sequence processing. In the context of combining multiple classifiers, all combined classifiers must be used on the same static input [16]. For sequence processing, here, we have to relax the condition so that all combined time-delay HMEs can be used on the different inputs which belong to the same sequence (utterance). For the purpose of combining different time-delay HMEs, we stipulate that $\{\mathbf{f}^{(t)}\}, \{\mathbf{f}^{(t)}, \mathbf{f}^{(t-1)}\}, \cdots, \{\mathbf{f}^{(t)}, \cdots, \mathbf{f}^{(t-n)}\}$ are valid inputs at time t if $\mathbf{f}^{(t-i)}$ denotes the speech frame of an utterance $\mathbf{x}$ at time $t-i$ and all outputs of time-delay HMEs based upon these inputs are permitted to be combined. Based upon this extension with respect to inputs, time-delay HMEs with different input-windows may be viewed as several different classifiers.

For the convenience to description, we relabel TD-HME[k] as CL_k which is still referred to as the time-delay HME which uses an input-window with the size of k. For the purpose of combination, it is necessary to acquire some prior knowledge about all combined time-delay HMEs. It leads us to take errors of all time-delay HMEs into consideration. Given a pattern space C consisting of N mutually exclusive sets $C = C_1 \bigcup \cdots \bigcup C_N$ with each of $C_i, \forall i \in \Phi = \{1, 2, \cdots, N\}$ representing a set of specified patterns called a class (e.g. the population of speakers in the problem of speaker identification). The errors of each combined time-delay HME CL_k are usually described by its *confusion matrix* [16] as follows,

$$PT_k = \begin{bmatrix} n_{11}^{(k)} & n_{12}^{(k)} & \dots & n_{1N}^{(k)} \\ n_{21}^{(k)} & n_{22}^{(k)} & \dots & n_{2N}^{(k)} \\ \dots\dots\dots\dots\dots\dots \\ n_{N1}^{(k)} & n_{N2}^{(k)} & \dots & n_{NN}^{(k)} \end{bmatrix} \tag{10}$$

for $k = 0, 1, \cdots, K-1$ if there are K time-delay HMEs to be combined; where each row i corresponds to class i and each column j corresponds to the event $CL_k(\mathbf{x}_t^{(k)}) = j$ where $\mathbf{x}_t^{(k)}$ is a time-delay feature vector at time t for the time-delay HME CL_k. All confusion matrices can be derived from outputs that each combined time-delay HME works on a subset of its test data. For an event $CL_k(\mathbf{x}_t^{(k)}) = j$ of an error-bearing time-delay HME CL_k, its truth has uncertainty. With the knowledge of its confusion matrix PT_k, such an uncertainty could be described by the conditional

probabilities that propositions $\mathbf{x}_t^{(k)} \in C_i (i = 1, \cdots, N)$ are true under the occurrence of the e vent $CL_k(\mathbf{x}_t^{(k)}) = j$, that is,

$$P(\mathbf{x}_t^{(k)} \in C_i | CL_k(\mathbf{x}_t^{(k)}) = j) = \frac{n_{ij}^{(k)}}{\sum_{i=1}^{N} n_{ij}^{(k)}}, \;\; i = 1, \cdots, N. \tag{11}$$

By means of the conditional probabilities, we may define a *belief value* $bel(\cdot)$. For each of N mutually exclusive propositions $\mathbf{x}_t^{(k)} \in C_i, \forall i \in \Phi$, the higher the $bel(\cdot)$ which gives to a proposition, the more likely it is true. Such a $bel(\cdot)$ is defined as follows,

$$bel(\mathbf{x}_t^{(k)} \in C_i | CL_k(\mathbf{x}_t^{(k)}), EN) = P(\mathbf{x}_t^{(k)} \imath n C_i | CL_k(\mathbf{x}_t^{(k)}) = j_k), \;\; i = 1, \cdots, N. \tag{12}$$

That is, $bel(\cdot)$ is defined as the probability under the condition of $CL_k(\mathbf{x}_t^{(k)}) = j_k$ and the environment EN where EN means the common classification environment that consists of any event which is independent of any of events $CL_k(\mathbf{x}_t^{(k)}) = j_k$ at time t, $k = 1, \cdots, K$. For instance, the environment at least contains the occurrence of a specific input pattern $\mathbf{x}_t^{(k)}$ at time t for the time-delay HME CL_k.

With K time-delay HMEs,[5] $CL_0, CL_1, \cdots, CL_{K-1}$, there are K matrices $PK_0, \cdots, PK_{K-1}$. When these time-delay HMEs are used on the valid inputs, K events $CL_k(\mathbf{x}_t^{(k)}) = j_k, (k = 0, 1, \cdots, K-1)$ will occur at time t. As discussed previously, each $CL_k(\mathbf{x}^{(k)}) = j_k$ and its corresponding PT_k could supply a set of $bel(\mathbf{x}_t^{(k)} \in C_i | CL_k(\mathbf{x}_t^{(k)}), EN), i = 1, \cdots, N$, each of which supports one of the N propositions. Next, the task is to integrate these individual supports to give the combined belief values. For an utterance **x**, at time t, we define the combined belief values as follows,

$$\begin{aligned} bel(i) &= bel(\mathbf{x} \in C_i | CL_0(\mathbf{x}_t^{(0)}), \cdots, CL_{K-1}(\mathbf{x}_t^{(K-1)}), EN) \\ &= P(\mathbf{x} \in C_i | CL_0(\mathbf{x}_t^{(0)}) = j_0, \cdots, CL_{K-1}(\mathbf{x}_t^{(K-1)}) = j_{K-1}, EN), \;\; i = 1, \cdots, N. \end{aligned} \tag{13}$$

where $\mathbf{x}_t^{(k)} = \{\mathbf{f}^{(t)}, \cdots, \mathbf{f}^{(t-k)}\}$ is the time-de lay feature vector consisting of $k+1$ successive frames from time $t-k$ to time t. Since time-delay HMEs $CL_0, CL_1, \cdots, CL_{K-1}$ perform independent of each other, the events $CL_0(\mathbf{x}_t^{(0)}) = j_0, \cdots, CL_{K-1}(\mathbf{x}_t^{(K-1)}) = j_{K-1}$ are independent of each other either under the condition of $\mathbf{x}_t^{(k)} \in C_i$ as well as EN or the condition of merely EN. For an utterance **x**, according to the Bayesian formula, we have

$$\begin{aligned} bel(i) &= P(\mathbf{x} \in C_i | CL_0(\mathbf{x}_t^{(0)}) = j_0, \cdots, CL_{K-1}(\mathbf{x}_t^{(K-1)}) = j_{K-1}, EN) \\ &= \frac{P(CL_0(\mathbf{x}_t^{(0)}) = j_0, \cdots, CL_{K-1}(\mathbf{x}_t^{(K-1)}) = j_{K-1} | \mathbf{x} \in C_i, EN) P(\mathbf{x} \in C_i | EN)}{P(CL_0(\mathbf{x}_t^{(0)}) = j_0, \cdots, CL_{K-1}(\mathbf{x}_t^{(K-1)}) = j_{K-1} | EN)} \\ &= \frac{\prod_{k=0}^{K-1} P(CL_k(\mathbf{x}_t^{(k)}) = j_k | \mathbf{x} \in C_i, EN) P(\mathbf{x} \in C_i | EN)}{\prod_{k=0}^{K-1} P(CL_k(\mathbf{x}_t^{(k)}) = j_k | EN)} \\ &= P(\mathbf{x} \in C_i | EN) \frac{\prod_{k=0}^{K-1} P(\mathbf{x} \in C_i | CL_k(\mathbf{x}_t^{(k)}) = j_k)}{\prod_{k=0}^{K-1} P(\mathbf{x} \in C_i | EN)} \end{aligned} \tag{14}$$

[5] Here, the original HME is regarded as the time-delay HME in which the size of input-window is 0.

Since

$$\frac{P(CL_k(\mathbf{x}_t^{(k)}) = j_k | \mathbf{x} \in C_i, EN)}{P(CL_k(\mathbf{x}_t^{(k)}) = j_k | EN)} = \frac{P(\mathbf{x} \in C_i | CL_k(\mathbf{x}_t^{(k)}) = j_k)}{P(\mathbf{x} \in C_i | EN)} \tag{15}$$

where $P(\mathbf{x} \in C_i | CL_k(\mathbf{x}_t^{(k)}) = j_k)$ could be estimated by Eq.(11) and $P(\mathbf{x} \in C_i | EN)$ represents the probability that $\mathbf{x} \in C_i$ is true under occurrence of $\mathbf{x}$ and the common environment. Although there may exist a better estimation of $P(\mathbf{x} \in C_i | EN)$, here, we still follow the original estimation in [16] so that we may achieve the integrated belief values as follows,

$$bel(i) = \beta \prod_{k=0}^{K-1} P(\mathbf{x} \in C_i | CL_k(\mathbf{x}_t^{(k)}) = j_k) \tag{16}$$

where $\beta = [\sum_{i=1}^{N} \prod_{k=0}^{K-1} P(\mathbf{x} \in C_i | CL_k(\mathbf{x}_t^{(k)}) = j_k)]^{-1}$. Based upon these integrated belief values, we may classify an unknown utterance $\mathbf{x}$ into a class $s \in \Phi$ at time t according to the decision rule as follows,

$$CL(\mathbf{x}) = s, \quad \text{if } bel(s) = \max_{i \in \Phi} bel(i) \tag{17}$$

It is necessary to note that for an utterance $\mathbf{x} = \{\mathbf{f}^{(1)}, \cdots, \mathbf{f}^{(T)}\}$, the different numbers of time-delay feature vectors on $\mathbf{x}$ will be achieved if we choose different input-windows for time delay. That is, the time-delay feature vector cannot be available when time t is shorter the size of the chosen input-window. For the $\mathbf{x}$, only $T - n$ feature vectors will be available if the size of time window is $n (n \geq 0)$. Accordingly, $T - n$ results are merely produced with the TD-HME[n]. For the purpose of combination, we shall have to add the same result(s) for the TD-HME[n] through averaging n previous $T - n$ results. The new added result(s) will be used as output(s) of the TD-HME[n] when time $t < n$. Using this trick, the current combination of time-delay HMEs is as same as the one of multiple classifiers in [16][22-25].

4 System Overview

We have already developed a closed-set text-dependent speaker identification system based upon the time-delay HME (TD-HME) architecture in Sun Sparc II workstation. The scheme of the system is illustrated in Fig. 4.

The preprocessing of the acoustic data consists of several steps. First, an utterance is sampled with 11.025 KHz sampling frequency and 16 bits digitization. Then it is processed by a silence-removing algorithm followed by the application of a pre-emphasis fileter $H(z) = 1 - 0.95z^{-1}$. In the current system, we adopt the linear predictive coding (LPC) power spectrum. With respect to speaker features used for training neural networks, there were an investigation and a comparison among several common LPC-based features [26], such as *power spectrum, cepstral coefficient* and *autocorrelation coefficient* etc. The experiment shows that the LPC power spectrum has the best performance for text-dependent speaker identification. In the phase of feature extraction, the processed acoustic data is segmented with 25.6 ms per frame and there is 12.8 ms overlapping between two successive frames. A 16th-order linear predictive analysis is first performed for each speech frame, then LPC power spectrum is computed with a 256 points FFT for each frame. Moreover, a *critical bandwidth filter* is employed for processing the LPC power spectrum further [26], which is considered as a simulation of the processing mechanism in the human peripheral auditory system. For this purpose, the power spectrum is divided from low frequency to high one ($0 \sim 5.512$ KHz) into 24 channels. In each channel, the energy is accumulated and denoted as E_i $(i = 1, 2, ..., 24)$. For simulating the firing of inner hair cells, an entropy is defined over each

channel for producing a 24-order feature vector $I = (I_1, I_2, \cdots, I_{24})$ for each frame where

$$I_i = -P_i \log P_i \quad \text{and } P_i = \frac{E_i}{\sum_{j=1}^{24} E_j}, \quad i = 1, 2, ...24.$$

Using feature vectors of the acoustic data, we can train the time-delay HME with the EM algorithm in the supervised manner for classification so that the speaker identity can be achieved.

For an utterance, we can create a set of 24-order features after preprocessing and feature extraction. For a supervised task, training patterns include input feature vectors and their corresponding target vectors. Here, we adopt an encoding scheme to form a target vector as follows. For class i, only the ith component of a target vector is *one* and other components are *zero*. So feature vectors and their corresponding target vectors are lumped together to form training pairs. In the current system, we choose 10 isolated digits from '0' to '9' as the fixed text. Depending upon the fixed text, 10 time-delay HMEs or combination of time-delay HMEs are established so that 10 classifiers correspond to 10 digits from '0' to '9' respectively. The EM algorithm in section 2 is used for training these classifiers. Furthermore, using the combining scheme, the system may produce the integrated result if there are multiple time-delay HMEs. During test, for an unknown utterance, the system may produce several results. Using the principle of *majority*, the system can decide who is the speaker.

5 Experiments

In this section, we shall describe the database and experimental results relevant to the comparison between the proposed *generalized Bernoulli density* and *multinomial logit density* for *multiway classification*, the individual time-delay HME and combination of multiple time-delay HMEs in detail.

5.1 Database and Results on Multiway Classification

We have created an acoustic database for the experiments of speaker identification. The database consists of 10 isolated digits from '0' to '9' uttered in Chinese and 10 male speakers are registered currently. For training and evaluating the performance of the system, we record the utterances at three different sessions, in which each digit is uttered 6 times in the first session and 10 times in the additional two sessions. There is an interval of one month between any two successive sessions. As a result, all utterances are divided naturally into three sets chronologically. The first set consists of some utterances recorded in the first session which are the first 5 times of utterances of each digit as the training set called Set-1. Other two sets are composed of all utterances recorded in the second and the third sessions; these as test sets are called Set-2 and Set-3 respectively. After preprocessing and feature extraction, we achieve three sets consisting of 24-order features. TABLE I lists the number of feature vectors in all three sets.

TABLE I
The number of feature vectors in training and test sets

Text	'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'
Set-1	1162	1163	1159	1125	1141	1174	1204	1173	1171	1198
Set-2	2399	2379	2335	2215	2164	2424	2403	2261	2318	2408
Set-3	2204	2175	2210	2149	2192	2348	2270	2148	2194	2298

We have already applied both the proposed *generalized Bernoulli density* (GBD) and *multinomial logit density* (MLD) to the HME without time delay in speaker identification in order to compare

them for the multiway classification. An HME with the 2-8 structure has been chosen and the GBD and the MLD are employed to model the probabilistic description of expert nets for multiway classification, respectively. Here, we define an epoch as the one consisting of E-step and M-step in the EM algorithm. Thus, the training time with the GBD and the MLD is shown in TABLE II. Using the trained HME with the GBD and the MLD, we may test data in Set-2 and Set-3 and the tests on Set-2 and Set-3 are hereafter called Test-1 and Test-2. In order to evaluate the performance of the system, we use a way called *digit-based* method to test the system. In this method, we simply use the utterance of single digit to determine the speaker's identity. The identifying accuracies of the HMEs with the GBD and the MLD are shown in TABLE III. According to TABLE II and TABLE III, obviously, the HME with the GBD yields faster training than the one with the MLD, though their identifying accuracies are quite similar. Actually, in our other experiments, the HME with the MLD often suffers from the instability so that it cannot reach a steady state.

TABLE II
The training time of the HME with the GBD and the MLD

Text	'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'	mean
Epoches (GBD)	4	3	5	4	4	3	3	4	5	3	3.8
Time(*min*) (GBD)	16	12	20	16	16	12	12	16	20	12	15.2
Epochs (MLD)	5	4	5	5	6	4	4	5	6	4	4.8
Time(min) (MLD)	20	16	20	20	24	16	16	20	24	16	19.2

TABLE III
The identifying accuracies of the HME with the GBD and the MLD

Text	'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'	mean
Test-1 (GBD)	91	96	87	86	83	90	85	89	86	91	88.4
Test-2 (GBD)	89	95	88	85	80	90	84	91	82	90	87.4
Test-1 (MLD)	90	95	89	85	84	89	86	87	86	90	88.1
Test-2 (MLD)	89	93	89	84	82	91	83	92	81	88	87.2

5.2 Results Using Individual Time-Delay HME

Like the problem appearing in the multilayer perception, there is also a pre-determined architecture problem before training for the HME. For solving this problem, we employ the 2-fold *cross-validation* method to remedy it. In the stage, we first divide the training set into two subsets; one for training and the other for test. Using this method, we have investigated 7 architectures covering from one level to three levels. Using a time-delay HME architecture, TD-HME[2], in which the size of input-window is $2(n = 2)$ for instance, we show the results in TABLE IV.

TABLE IV
The results of cross-validation for the TD-HME[2]

Architecture	1-16	2-2	2-8	2-16	2-2-8	2-4-8	2-4-16
Epoches	11	21	4	5	5	5	6
Time(min)	48	82	23	38	54	69	92
Identifying Accuracy(%)	97.0	96.0	99.0	96.0	98.0	98.0	98.0

According to the performance and training time, finally, we choose the two levels HME with 2-8 as the classifier. Using such an architecture, usually, 4 or 5 epoches are merely needed to

reach the steady state for a given error threshold. The mean-square-error threshold is *0.05* and the learning rate η is 0.4 in Eq.(9) in all our experiments.

In experiments, we have already applied three different time-delay HMEs, i.e. TD-HME[1], TD-HME[2] and TD-HME[3], to the problem of speaker identification. In order to evaluate the performance of the system, we also adopt the digit-based method to test the system. The experimental results are summarized in TABLE V and TABLE VI, respectively.

TABLE V
The identifying accuracies of three time-delay HMEs in Test-1

Text	'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'	mean
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
TD-HME[3]	90.0	97.0	87.0	85.0	86.0	87.0	88.0	92.0	86.0	95.0	89.3

TABLE VI
The identifying accuracies of three time-delay HMEs in Test-2

Text	'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'	mean
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
TD-HME[3]	90.0	95.0	92.0	87.0	85.0	87.0	85.0	91.0	84.0	92.0	88.8

According to the above experimental results in TABLE II, TABLE V and TABLE VI, we may make a conclusion that the performance of time-delay HMEs is really better than one of the HME without time-delay, but the improvement is limited. Furthermore, the use of different time-delay HMEs with a short input-window for speaker identification does not bring about the quite different performance of the system. Since the voice of a speaker changes over time, usually the longer the time interval between training and test is, the more difficult it is for the system to identify the speaker. With the consideration, based on the above experimental results in TABLE VI, we may draw an outcome that the longer input-window seems to be better in capturing a speaker's individual features. However, the computational cost is quite expensive; in comparison with the TD-HME[1], the training time of the TD-HME[3] is about twice as much as one of TD-HME[1] for the same problem.

We also adopt another method called *sequence-based method* to test the systems supported by time-delay HMEs in order to investigate the robustness of these systems. In this method, we first produce a sequence consisting of 5 digits at random (it may be viewed as a password.), then ask a speaker to utter digits in the sequence one by one. For each digit in the sequence, there is an identified result by using the aforementioned digit-based method. After obtaining all results, the system tolls a vote with the principle of majority that a speaker can be identified only if there are at least three same identification results for the speaker; otherwise, the system refuses to identify the unknown speaker. Using the method, all systems with TD-HME[1], TD-HME[2] and TD-HME[3] are robust with 100% identification accuracies over 2000 tests up to now.

5.3 Results by Combining Time-Delay HMEs

The experimental results on an individual time-delay HME with the short input-window indicate that it is difficult to find an appropriate size of the short input-window for the time-delay HME in order to achieve significant improvement in performance. On the other hand, the use of a long input-window for the time-delay HME will suffer from a very high computational burden. To attack these problems, in section 3.2, we have described a scheme of combining multiple time-delay

HMEs with different input-windows. In this method, first of all, we must acquire the prior knowledge about all combined time-delay HMEs through achieving their confusion matrices in Eq.(10). As a result, we employ the sixth time utterances of each digit recorded in the first session and the first four times utterances of each digit recorded in the second session as a data set to derive all confusion matrices. In our current experiments, we consider combining the HME without time delay, TD-HME[0], and the aforementioned three time-delay HMEs, i.e. TD-HME[1], TD-HME[2] and TD-HME[3]. Applying these trained HMEs on the aforementioned data set, we may achieve four confusion matrices corresponding to TD-HME[0], TD-HME[1], TD-HME[2] and TD-HME[3], respectively. During test, we complete three tests, i.e. Test-1, Test-2 and Test-1A. Test-1A refers to that the test data set consists of all data in Set-2 except those four times utterances of each digit used to achieve confusion matrices. Based upon confusion matrices and the decision rule described in Eq.(17) accordingly, the digit-based results of combining these time-delay HMEs for speaker identification are shown in TABLE VII.

TABLE VII
The identifying accuracies of combining time-delay HMEs.

Text	'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'	mean
Test-1	95.0	98.0	94.0	92.0	90.0	93.0	95.0	97.0	91.0	97.0	94.2
Test-2	91.0	96.0	92.0	88.0	87.0	93.0	87.0	94.0	88.0	93.0	90.9
Test-1A	93.0	97.0	92.0	89.0	90.0	91.0	95.0	95.0	90.0	95.0	92.7

In comparison with results shown in TABLE V and TABLE VI, the performance of the system using the combination of time-delay HMEs has been significantly improved. Furthermore, the sequence-based method has also been used for test and the identifying accuracies are still 100% over 2000 tests.

6 Discussions

In this section, we intend to discuss some problems about the HME and the time-delay in the HME.

For solving the IRLS problem, we have introduced a changeable learning rate to the updated formula in Eq.(7) instead of that the value is always one in the previous formula [9]. As a result, the modified updated formula may alleviate the instability and accelerate training during training. In addition, we have also proposed the *generalized Bernoulli density* for the specific multiway classification instead of the *multinomial logit density*. Indeed, the *Bernoulli density* can also be used for the multiway classification if we consider the multiway classification as N two-category classifications for N classes. Unfortunately, it will suffer from the trouble of time-consuming. For instance, the current task will need *100* HMEs with the Bernoulli density since there are 10 speakers registered in the system and 10 digits as the fixed text for text-dependent speaker identification. But the model with the generalized Bernoulli density avoids the problem successfully by using the appropriate architecture for the multiway classification and keeping the characteristics of the Bernoulli density. In addition, the model with the generalized Bernoulli density is more efficient than the one with the multinomial logit density since it keeps the same architecture as the one with the multinomial logit density but uses a more appropriate distributing density to model the specific multiway classification problem.

We have introduced the time delay concept to the HME architecture. In order to attack the dilemma on the size of input-window for time-delay and the computational cost, moreover, we use the trick of combining multiple classifiers to the combination of multiple time-delay HMEs.

As mentioned in section 3.1, currently, all component nets in the HME are two-layer architectures without any hidden layer. The reason is that such architectures of component nets in HME are based upon the GLIM theory which provides the basic statistical structure for the components of the HME. There seems to be another way to capture more temporal information in the HME if we do not employ the GLIM theory to model the basic statistical structure of expert nets in the HME. Instead we model the probabilistic description of expert net (i, j) in the HME as

$$P(\mathbf{y}|\mathbf{x}, \theta_{ij}) = \exp[-(\mathbf{y} - \mu_{ij})^T(\mathbf{y} - \mu_{ij})] \tag{18}$$

where $\mathbf{y}$ is chosen from the probability density $P_{ij}(\mathbf{y}|\mathbf{x}, \theta_{ij})$ and θ_{ij} is all free parameters in the probability model. Unlike the one in the GLIM theory, μ_{ij} is the mean of $\mathbf{y}$ which may be achieved using any probability approximator instead of the one in the exponential family. It is well known that a neural network can be viewed as a probability approximator [27]-[31]. Thus, we may use a classic time-delay neural network with at least one hidden layer [14] or an adaptive time-delay neural network [32] as expert nets in the HME. Hence, the EM algorithm is still employed to deal with the problem of parameter estimation in the HME. There is no change in the E-step and the optimization of gating net(s) in the M-step. The unique change is in the optimization of expert nets in the M-step. As a result, this problem becomes general unconstrained optimization due to $h_{ij}^{(t)} \geq 0$ as follows,

$$\begin{aligned}
\theta_{ij}^{(p+1)} &= \arg\max_{ij} \sum_t h_{ij}^{(t)} \ln P(\mathbf{y}|\mathbf{x}, \theta_{ij}) \\
&= \arg\max_{ij} \sum_t h_{ij}^{(t)} [-(\mathbf{y}^{(t)} - \mu_{ij}^{(t)})^T(\mathbf{y}^{(t)} - \mu_{ij}^{(t)})] \\
&= \arg\min_{ij} \sum_t h_{ij}^{(t)} [(\mathbf{y}^{(t)} - \mu_{ij}^{(t)})^T(\mathbf{y}^{(t)} - \mu_{ij}^{(t)})]
\end{aligned} \tag{19}$$

Instead of the IRLS algorithm, thus, there are many optimization algorithms for the problem such as the gradient-based method and Newton method etc. [33]. For the model, moreover, we shall adopt smooth delay functions over time instead of the current rectangular hat-shaped function for improving the performance of time-delay [34]-[38].

7 Conclusions

We have described the application of the time-delay HME with EM algorithm to speaker identification based on the proposed *generalized Bernoulli density*. It has been shown that the speaker identification system based on the time-delay HME can achieve the satisfactory performance. Moreover, we introduce a combining scheme to attack the dilemma on the size of input-window and the computational cost by completing multi-scale analysis for the temporal data. Its usefulness has been demonstrated in experiments. The merit of fast training of the HME along with the EM algorithm has also been verified in our experiments. Our near-future work is to develop a time-delay HME with stronger capabilities of temporal processing using the method discussed in this paper. In addition, the direction of our future work also includes developing the HME with self-architecture to attack the problem of pre-determined structure.

Acknowledgements

Authors would like to thank anonymous referees for their useful comments and suggestions which help improve this paper. This work was partially supported by China National Science Foundation under the Grant 69571002 and 69475007 as well as the Climbing Program – National Key Project for Fundamental Research in China with Grant NSC 92097.

References

[1] T. Matsui and S. Furui, "*Speaker recognition technology,*" NTT Review, Vol.7, No.2, March, 1995, pp. 40-48.

[2] G.R. Doddington, "*Speaker recognition–identifying people by their voices,*" Proceedings of IEEE, Vol.73, No.11, 1986, pp. 1651-1664.

[3] D. O'Shaughnessy, "*Speaker recognition*," IEEE ASSP Magazine, Vol. 3, No. 4, Oct. 1986, pp. 4-17.

[4] S. Furui, "*An overview of speaker recognition technology,*" Proceeding of ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, 1994, pp.1-9.

[5] Y. Bennani and P. Gallinari, "*Connectionist approaches for automatic speaker recognition,*" Proceedings of ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, 1994, pp. 95-102.

[6] J. He, L. Liu and G. Palm, "*A text-independent speaker identification system based on neural networks,*" Proceedings of International Conference on Spoken Language Processing, Yokohama, Japan, Oct. 1994.

[7] K.R. Farrell and R.J. Mammone, "*Speaker identification using neural tree networks,*" Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 1994, pp. I165-I168.

[8] X. Jiang, Z. Gong, F. Sun and H. Chi, "*A speaker recognition system based on auditory model,*" Proceedings of World Congress on Neural Networks, San Diego, 1994, pp. IV595-IV600.

[9] M.I. Jordan and R.A. Jacobs, "*Hierarchical mixture of experts and EM algorithm,*" Neural Computation, Vol. 6, 1994, pp. 181-214.

[10] S.R. Waterhouse, *Personal communication*, 1994.

[11] S.R. Waterhouse and A.J. Robinson, "*Classification using hierarchical mixtures of experts,*" Proceedings of IEEE Conference on Neural Networks and Signal Processing, 1994.

[12] S.R. Waterhouse and A.J. Robinson, "*Prediction of acoustic vectors using hierarchical mixture of experts,*" Advance in Neural Information Processing Systems 7, J.D. Cowan, G. Tesauro and J. Alspector eds., MIT Press, Cambridge, MA, 1995.

[13] K. Chen, D. Xie and H. Chi, "*Speaker identification based on hierarchical mixture of experts,*" Proceedings of World Congress on Neural Networks, Washington D. C., July, 1995, pp. I493-I496.

[14] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. Lang, "*Phoneme recognition using time-delay neural networks*", IEEE Transaction on Acoustics, Speech and Signal Processing, ASSP-37, March, 1989, pp. 328-339.

[15] K. Chen, D. Xie and H. Chi, "*Speaker identification based on a time-delay modular neural network,*" Journal of Advanced Software Research, Allerton Press, Vol.4, No.1, 1996. (in press)

[16] L. Xu, A. Krzyzak and C.Y. Suen, "*Methods of combining multiple classifiers and their applications to handwriting recognition,*" IEEE Transactions on Systems, Man and Cybernetics, Vol.23, No.3, 1992, pp. 418-435.

[17] P. McCullagh and J.A. Nelder, *Generalized linear models*, Chapman and Hall, London, 1989.

[18] L. Xu, M.I. Jordan and G.E. Hinton, "*A modified gating network for the mixture of experts architecture,*" Proceedings of World Congress on Neural Networks, San Diego, 1994, pp. II405-II410.

[19] L. Xu, M.I. Jordan and G.E. Hinton, "*An alternative model for mixtures of experts,*" Advance in Neural Information Processing Systems 7, J.D. Cowan, G. Tesauro and J. Alspector eds., MIT Press, Cambridge, MA, 1995.

[20] A.P. Dempster, N.M. Laird and D.B. Rubin, "*Maximum likelihood from incomplete data via the EM algorithm,*" J. R. Statist. Soc. B-39, 1977, pp. 1-38.

[21] M.I. Jordan and L. Xu, "*Convergence results for the EM approach to mixture-of-experts architectures,*" TR-9302, Computational Cognitive Science Lab, Massachusetts Institute of Technology, 1993.

[22] C.Y. Suen, C. Nadal, T.A. Mai, R. Legault and L. Lam, "*Recognition of totally unconstrained handwritten neurals based on the concept of multiple experts,*" Proceedings of International Workshop on Frontiers in Handwriting Recognition, Montreal, April, 1990, pp. 131-143.

[23] T.K. Ho, J.J. Hull and S.N. Srihari, "*Combination of structural classifiers,*" Proceedings of IAPR Workshop on Syntactic and Structural Pattern Recognition, June, 1990, pp.123-137.

[24] R. Battiti and A.M. Colla, "*Democracy in neural nets: voting schemes for classification,*" Neural Networks, Vol. 7, No. 4, 1994, pp.691-708.

[25] G. Rogova, "*Combining the results of several neural network classifiers,*" Neural Networks, Vol. 7, No. 5, 1994, pp.771-781.

[26] Xin Jiang, *Text-dependent speaker identification based on artificial neural networks*, Master Thesis, Center for Information Science, Peking University, 1994. (in Chinese)

[27] E.B. Baum and F. Wilczek, "*Supervised learning of probability distributions by neural networks,*" Advance in Neural Information Systems, D.Z. Anderson ed., New York: American Institute of Physics, 1988, pp. 52-61.

[28] H. Gish, "*A probabilistic approach to the understanding and training of neural network classifiers,*" Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Albuquerque, NM, 1990, pp. 1361-1364.

[29] H. Bourlard and C.J. Wellekens, "*Links between Markov models and multilayer perceptrons,*" IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 12, 1990, pp. 1167-1178.

[30] M.D. Richard and R.P. Lippmann, "*Neural network classifiers estimate Bayesian a posteriori probabilities,*" Neural Computation 3, 1991, pp. 461-483.

[31] H. Bourlard and C.J. Wellekens, *Connectionist speech recognition – a hybrid approach,* Amsterdam: Kluwer, 1994.

[32] S.P. Day and M.R. Davenport, "*Continuous-time temporal back-propagation with adaptable time delays,*" IEEE Transactions on Neural Networks, Vol.4, No.2, 1993, pp. 348-354.

[33] R. Fletcher, *Practical methods of optimization*, John Wiley&Sons, New York, 1987.

[34] D.W. Tank and J.J. Hopfield, "*Neural computation by concentrating information in time,*" Proceedings National Academy of Sciences, April, 1987, pp. 1896-1900.

[35] K.P. Unnikrsihnan, J.J. Hopfield and D.W. Tank, "*Connected digit speaker dependent speech recognition using a network with time-delay connections,*" IEEE Transactions on Signal Processing 39, 1991, pp. 698-713.

[36] K.P. Unnikrsihnan, J.J. Hopfield and D.W. Tank, "*Speaker-independent digit recognition using a neural network with time-delayed connections,* " Neural Computation 4, 1992, pp. 108-119.

[37] B. de Vires and J.C. Principe, "*A theory for neural networks with time delays,*" Advances in Neural Information Systems 3, Lippmann, Moody and Touretzky eds., Morgan Kaufmann, 1991, pp. 162-168.

[38] B. de Vires and J.C. Principe, "*The Gamma model – a new neural net model for temporal processing neural networks,*" Neural Networks 5, 1992, pp. 565-576.

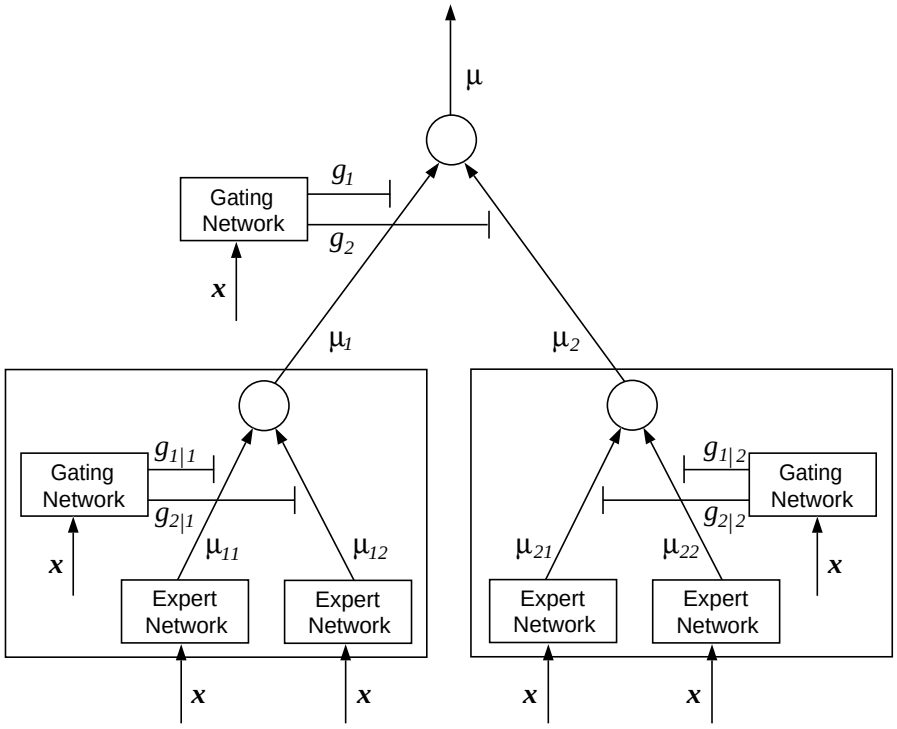

Fig. 1. The architecture of hierarchical mixture of experts.

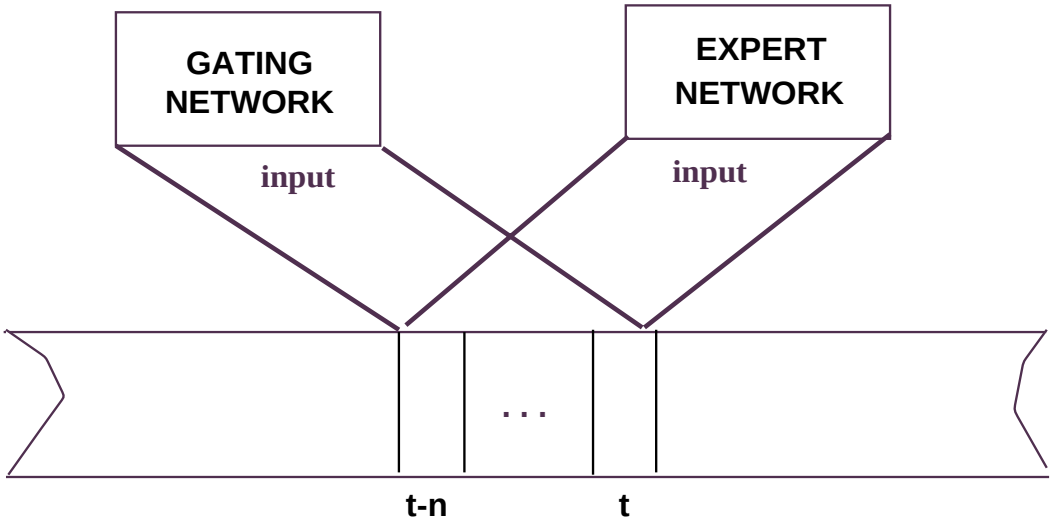

Fig. 2. The input of expert and gating nets in the time-delay HME.

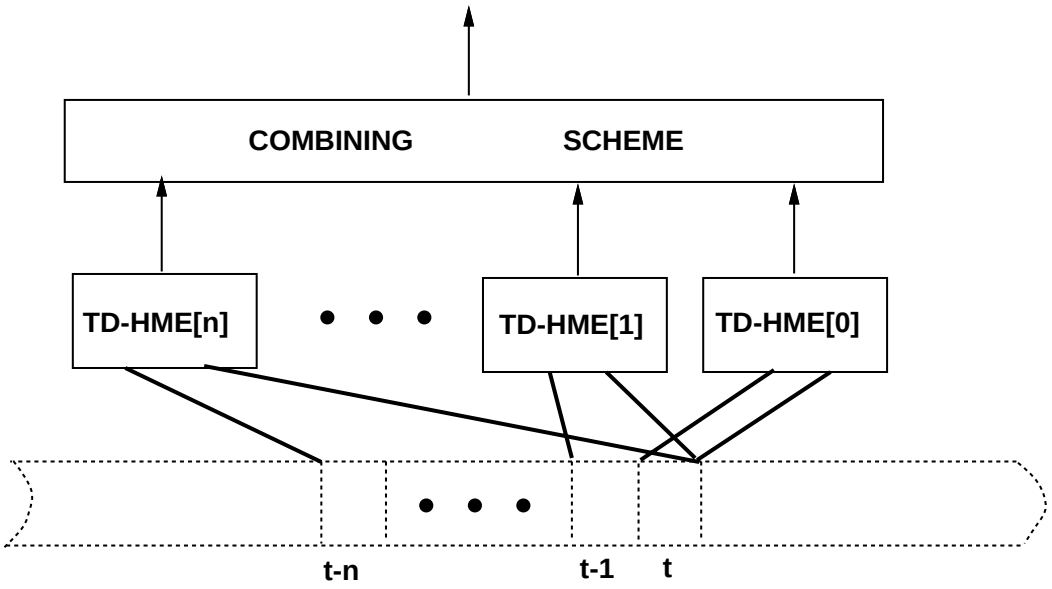

Fig. 3. The scheme of combining multiple time-delay HMEs.

Fig. 4. The scheme of speaker identification system based on the time-delay HME.