Capture Interspeaker Information With a Neural Network for Speaker Identification

Lan Wang, Ke Chen, *Senior Member, IEEE*, and Huisheng Chi, *Senior Member, IEEE*

***Abstract*—Model-based approach is one of methods widely used for speaker identification, where a statistical model is used to characterize a specific speaker's voice but no interspeaker information is involved in its parameter estimation. It is observed that interspeaker information is very helpful in discriminating between different speakers. In this paper, we propose a novel method for the use of interspeaker information to improve performance of a model-based speaker identification system. A neural network is employed to capture the interspeaker information from the output space of those statistical models. In order to sufficiently utilize interspeaker information, a rival penalized encoding rule is proposed to design supervised learning pairs. For better generalization, moreover, a query-based learning algorithm is presented to actively select the input data of interest during training of the neural network. Comparative results on the KING speech corpus show that our method leads to a considerable improvement for a model-based speaker identification system.**

***Index Terms*—Interspeaker information, KING speech corpus, model-based method, neural networks, query-based learning algorithm, rival penalized encoding scheme, speaker identification.**

I. Introduction

THE goal of speaker identification is to automatically determine a speaker's identity by his/her voice among a population. It is widely applied to many fields from confidential data access to audio indexing in multimedia [14]. In general, a speaker identification system may be either text-dependent, where the same text is required for both training and test, or text-independent, where arbitrary text is allowed to utter.

Speech is a dynamic acoustic signal with many variations, where both interspeaker and intraspeaker variabilities are highly correlated with speaker identification problem. For interspeaker variability, a primary source is physiological differences, e.g., vocal tract shape and length, between different speakers, which could be encoded in acoustic aspect of speech. On the other hand, person's manner of speech, e.g., word usage, is another source to induce some interspeaker variabilities. However, such a feature is too mysterious to extract from acoustic signals. For intraspeaker variability, it refers to the differences in speaking style [19], including speech rate, stress, voice quality, and temporal variation. In reality, there exist quite different spectral characteristics for the same words produced by the same speaker on different conditions. For speaker identification, the primary objective is to capture interspeaker variabilities, while accommodating intraspeaker variabilities.

Sophisticated methods for a speaker identification system, particularly in the text-independent style, include vector quantization (VQ) [4], [14], [26], hidden Markov model (HMM) [4], [6], and Gaussian mixture model (GMM) [22], [24]. These methods could be summarized as the so-called *model-based approach*; for each speaker a parametric statistical model is created to characterize the speaker's voice, but other speakers' information is not considered in building such a model. For a specific speaker, ideally, his/her statistical model should always yield the highest likelihood score for his/her own utterances. In reality, however, intraspeaker variabilities may prevent a statistical model from characterizing speaker's voice perfectly based on only a limited duration of voice. Since interspeaker variabilities are not emphasized effectively by those statistical models, the mismatch due to miscellaneous variabilities may lead to *misidentification* that utterances belonging to a specific speaker are wrongly identified as belonging to another speaker. As the population of speaker candidates increases, such misidentification is more severe such that the performance of a model-based speaker identification system can be highly degraded.

To improve the performance of such a system, some efforts have been made by considering interspeaker information in previous studies. As a consequence, some normalization methods have been proposed through the use of a cohort model to overcome intraspeaker variabilities [18], [16], [23], [25]. For a specific speaker, his/her cohort model is created based on the data belonging to a group of speakers who have similar characteristics in speech. To some extent, information on interspeaker variabilities is somehow conveyed in the cohort model and, therefore, is helpful in discriminating between different speakers in an indirect way. Since a cohort model is highly associated with a specific speaker, such a normalization method is mostly used for speaker verification where an identity claim is available. Therefore, it is inconvenient to speaker identification.

In previous studies, connectionist approaches have been applied to build speaker identification systems [2], [3], [7], [9], [13], [21] where a neural-network model is typically used to characterize all the speakers' voice in a given set. In this circumstance, the input space of a neural network is composed of feature vectors extracted from acoustic signals belonging to all the

Manuscript received November 29, 2000; revised August 30, 2001. A preliminary version of this paper was presented at IJCNN'2000 (see Proceedings of The IEEE-INNS-ENNS International Joint Conference on Neural Networks, vol. 5, pp. 247–252, IEEE Press, 2000). This work was supported in part by NSFC under Grant 60 075 017. The work of K. Chen was supported in part by an MSR Grant.

L. Wang is with the Speech Vision and Robotocs Group, Department of Engineering, University of Cambridge, Cambridge CA2 1PZ, U.K. (e-mail: lw256@cam.ac.uk).

K. Chen is with School of Computer Science, The University of Birmingham, Edgbaston, Birmingham B15 2TT, U.K. (e-mail: k.chen@cs.bham.ac.uk).

H. S. Chi is with National Laboratory of Machine Perception and The Center for Information Science, Peking University, Beijing 100871, P.R. China (e-mail: chi@cis.pku.edu.cn).

Publisher Item Identifier S 1045-9227(02)01804-0.

speakers, while the outputs are usually labels of corresponding speaker identities. Although interspeaker information is considered for modeling, phonetic features underlying speakers' voice are more difficult to be captured, due to the catastrophic or cross-talk effects of interaction among a huge number of synaptic weights during training in contrast to a model-based approach. On the other hand, there are a large amount of training data since a neural network usually works directly on the speech feature space, which results in a heavy computational burden, and thus, prevents the application of a neural-network based speaker identification system in practice. In general, speaker identification is a typical multicategory pattern classification task. Theoretically, any multicategory classification task can be decomposed into a set of binary classification subtasks, where each subtask is to discriminate between the data belonging to a specific class and all the others. By this fact, some connectionist methods have been proposed by constructing a set of neural networks with binary outputs for speaker identification [13], [21]. Indeed, those neural networks of binary outputs may work in a parallel way, which speeds up training. However, such methods have to encounter the training data *unbalance* problem that the available data belonging to a specific speaker are often much fewer than those of others during training. This bias could cause the performance of such a system to be degraded. On the other hand, modular neural-network methodologies [2], [7], [9] have been proposed to automatically decompose a complicated speaker identification task, which yields the better performance and fast training. However, the modular neural-network methods may suffer the catastrophic effects due to high dimensionality of the speech feature space such that interspeaker information still cannot be sufficiently utilized.

Our observation indicates that exploiting direct cross-relationship among different statistical speaker models, which we refer as to *interspeaker information*, provides a feasible way to lower misidentification [1]. Unlike previous approaches, we propose a novel connectionist method for the use of interspeaker information to improve the performance of a model-based speaker identification system. In our method, a neural network is employed to capture interspeaker information from the output space of those statistical speaker models, which forms a hybrid system consisting of statistical models and neural apparatus. The basic idea underlying our method is to build a mapping between the misidentification caused by statistical speaker models and their correction. To utilize interspeaker information sufficiently, a *rival penalized encoding rule* (RPER) is proposed to design supervised learning pairs for training the neural network. In contrast to traditional connectionist speaker identification systems, the neural network in our system works on the output space of statistical speaker models rather than the speech data space.

Active learning is an effective way to improve the generalization capability of a learner [10], [11], [15], [17], [20], [27]. Motivated by the idea of large-margin classifiers [28], we present an alternative query-based learning algorithm to actively select the input data of interest during training for better generalization. Unlike the previous query-based learning methods, our method always selects the data conveying maximal information in terms of a measure defined by ourselves. Thus, the neural network in our hybrid system is no longer a passive learner.

For evaluating the effectiveness of our method, we adopt the KING speech corpus consisting of wide-band and narrow-band sets, a benchmark for speaker recognition [5], in simulations. Comparative results show that our method leads to a considerable improvement, in particular, for the narrow-band set.

The remainder of this paper is organized as follows. Section II reviews the GMM-based speaker identification scheme. Section III presents our methodology. Section IV reports simulation results on the KING speech corpus. Conclusions are drawn in Section V.

II. GMM-Based Speaker Identification

As a typical model-based approach, GMM has been used to characterize speaker's voice in the form of probabilistic model. It has been reported that the GMM approach outperforms other classical methods for text-independent speaker identification [8], [22]. In this section, we briefly review the GMM-based speaker identification scheme that will be used in our simulations.

For a feature vector denoted as $\mathbf{x}_t$ belonging to a specific speaker s, the GMM is a linear combination of K Gaussian components as follows:

$$P(\mathbf{x}_t \,|\, \lambda_s) = \sum_{k=1}^{K} \omega_{s,k} P(\mathbf{x}_t \,|\, \mathbf{m}_{s,k}, \mathbf{\Sigma}_{s,k}) \tag{1}$$

where $P(\mathbf{x}_t \,|\, \mathbf{m}_{s,k}, \mathbf{\Sigma}_{s,k})$ is a Gaussian component parameterized by a mean vector, $\mathbf{m}_{s,k}$, and covariance matrix, $\mathbf{\Sigma}_{s,k}$, and $\omega_{s,k}$ is a linear combination coefficient for speaker s $(s = 1, 2, \ldots, S)$. Usually, a diagonal covariance matrix is used in (1). Given a sequence of feature vectors, $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots\}$, from a specific speaker's utterances, parameters estimation for $\lambda_s = (\omega_{s,k}, \mathbf{m}_{s,k}, \mathbf{\Sigma}_{s,k})$ $(k = 1, \ldots, K, s = 1, \ldots, S)$ is performed by the expectation-maximization (EM) algorithm. Thus, a specific speaker model is built through finding proper parameters in the GMM based on the speaker's own feature vectors.

To evaluate the performance, a sequence of feature vectors is divided into overlapping segments of T feature vectors for identification [22]

$$\overbrace{\mathbf{x}_l, \mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+T-1}}^{\text{segment}\,l}, \mathbf{x}_{l+T}, \ldots$$

$$\mathbf{x}_l, \overbrace{\mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+T-1}, \mathbf{x}_{l+T}}^{\text{segment}\,l+1}, \mathbf{x}_{l+T+1}, \ldots.$$

For a test segment $X^{(l)} = \{\mathbf{x}_l, \mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+T-1}\}$, the log-likelihood function of a GMM is as following:

$$\mathcal{L}(X^{(l)}, \lambda_s) = \sum_{t=l}^{l+T-1} \log P(\mathbf{x}_t \,|\, \lambda_s) \quad s = 1, \ldots, S. \tag{2}$$

Thus, the segment, $X^{(l)}$, is assigned to a label of registered speakers, s^*, on the basis of the maximum likelihood principle; that is

$$s^* = \underset{1 \le s \le S}{\operatorname{argmax}}\, \mathcal{L}(X^{(l)}, \lambda_s). \tag{3}$$

III. Methodology

In this section, we present our methodology on how to use the interspeaker information to improve the performance of a model-based speaker identification system. We first give our motivation based on the GMM-based speaker identification scheme. Then, we present our method for capturing interspeaker information with a neural network. For improving generalization, we further propose a query-based learning algorithm for training a neural network. All the technical components mentioned above constitute our methodology.

A. Motivation

As pointed out in the introduction, misidentification for a GMM may result from limited training data and possible flaws in a learning process. These factors cause a GMM not to model speaker's characteristics perfectly such that misidentification can occur on a mismatch condition. To indicate this problem, we first demonstrate the performance of a GMM-based speaker identification system. A training set is used to train each GMM for a set of speakers, and an alternative data set recorded in another session, hereinafter called *validation set*, is used to observe the variation of those GMMs' likelihood values. To simplify presentation, we assume that the likelihood values of a GMM are subject to the Gaussian distribution

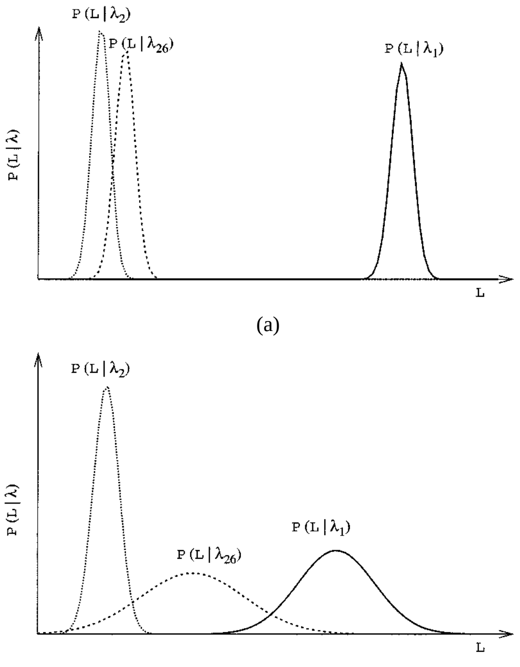
$$P(\mathcal{L} \mid \lambda_s) = \frac{1}{\sqrt{2\pi}\sigma_s} \exp\left(-\frac{(\mathcal{L}-\mu_s)^2}{2\sigma_s^2}\right). \tag{4}$$

Here $\mathcal{L}$ are the likelihood values defined in (2) for a specific speaker model s. μ_s and σ_s^2 are their mean and variance that can be estimated based on either a training or a validation set.

Now we take an example to present our points. For utterances belonging to speaker 1, their feature vectors are fed to different speaker models. Fig. 1 illustrates distribution of likelihood values produced by three speaker models, λ_1, λ_2, and λ_{26}, denoted by $P(\mathcal{L} \mid \lambda_1), P(\mathcal{L} \mid \lambda_2)$, and $P(\mathcal{L} \mid \lambda_{26})$, respectively. For utterances in the training set, it is observed in Fig. 1(a) that the distribution of likelihood values corresponding to the true speaker model, λ_1, is separated from that of others. It indicates that speaker 1 is always correctly identified in terms of utterances in his/her training set. For utterances in the validation set, however, such a result does not remain, as illustrated in Fig. 1(b). In this circumstance, the overlap between the $P(\mathcal{L} \mid \lambda_1)$ and $P(\mathcal{L} \mid \lambda_{26})$ implies that some utterances belonging speaker 1 may be wrongly identified as belonging to speaker 26, according to the decision strategy defined in (3). Such misidentification is unavoidable in reality but provides one kind of interspeaker information in terms of a validation set. In order to improve the performance of a model-based speaker identification system, how to utilize interspeaker information is an open problem. In the sequel, we present a neural-network method to capture such interspeaker information.

B. Capture interspeaker Information With a Neural Network

Prior to the presentation of our method, we first present a schematic diagram of our hybrid system, consisting of GMMs and a neural network, in Fig. 2 in order to understand our method better. In Fig. 2, GMMs are employed to characterize speakers in terms of their voice, while the neural network is used to lower misidentification made by GMMs through the use of interspeaker information. As a consequence, our system is trained in a two-stage way. Assume that there are S speakers registered in the system. In the first stage, S speaker models are created with GMMs based on a training set, where one GMM models a speaker. In the second stage, a neural network, working on the output space of GMMs, is employed to take advantage of the aforementioned interspeaker information conveyed in the output space of GMMs based on a validation set. During training, the neural network tends to learn how to lower misidentification resulting from GMMs in a supervised learning way. Once the two-stage learning is finished, the system, working in a cascade way, is applicable to any unknown voice token in a testing set for decision-making, as shown in Fig. 2. In the sequel, we present a systematic methodology to establish such a system for speaker identification.



(a)

(b)

Fig. 1. Distribution of likelihood values (scores) corresponding to three speaker models, λ_1, λ_2, and λ_{26}, denoted as $P(\mathcal{L} \mid \lambda_1)$ (solid line), $P(\mathcal{L} \mid \lambda_2)$ (dashed line), and $P(\mathcal{L} \mid \lambda_{26})$ (dotted line) for utterances belonging to speaker 1. (a) Distribution estimated on a training set. (b) Distribution estimated on a validation set (different from the training set).

For the use of interspeaker information, our idea is to build a mapping between misidentification and its correction by a neural network in terms of a validation set. In doing so, the remaining task is how to design supervised learning pairs based on the validation set for training the neural network.

Assume that there are total L speech segments belonging to S speakers in the validation set. For a speech segment, $X^{(l)} (l = 1, 2, \ldots, L)$, all the GMMs, $\lambda_s (s = 1, \ldots, S)$, produce a set of likelihood values $\mathcal{L}(X^{(l)}, \lambda_s) (s = 1, \ldots, S)$. Thus we assemble these likelihood values into an input vector to the neural network, $\tilde{\mathcal{L}}^{(l)} = [\tilde{\mathcal{L}}_1^{(l)}, \tilde{\mathcal{L}}_2^{(l)}, \ldots, \tilde{\mathcal{L}}_S^{(l)}]^T$, via the following normalization:

$$\tilde{\mathcal{L}}_s^{(l)} = \frac{\mathcal{L}_{\max}}{\mathcal{L}_{\max} + 0.5 - \mathcal{L}(X^{(l)}, \lambda_s)} \tag{5}$$

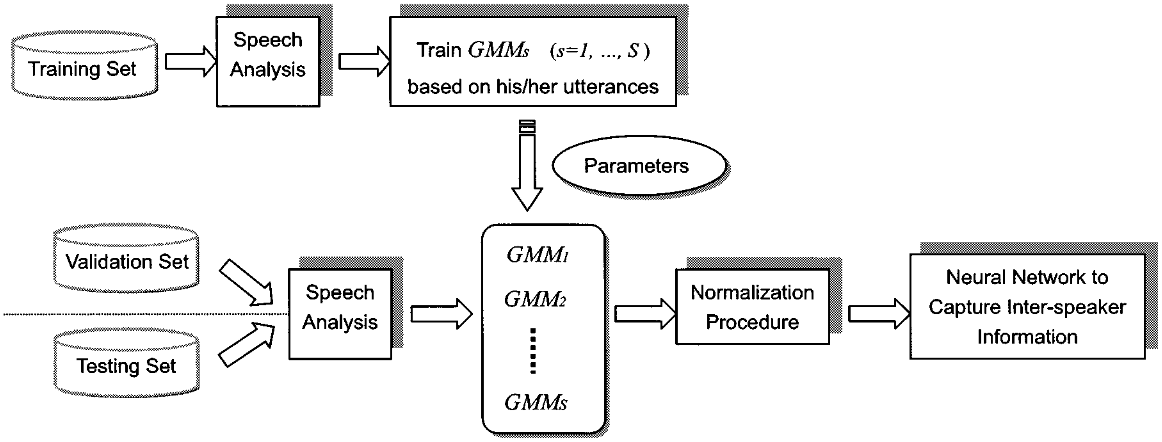


Fig. 2. The schematic diagram of our hybrid speaker identification system.

where $\mathcal{L}_{\max} = \max_{1\le s\le S, 1\le l\le L}\{\mathcal{L}(X^{(l)}, \lambda_s)\}$. The above normalization is designed to facilitate training of the neural network.

In order to discriminate between different speakers, it is desirable that misidentification yielded by GMMs [cf. Fig. 1(b)] is corrected while those right decisions made by GMMs keep unchanged. In other words, the target output of the neural network should be the right label for an input speech segment no matter what GMMs outputs are. Traditionally, the target output is encoded according to the one-of-S scheme, where the sth element of the target vector equals one if the input belongs to speaker s and other elements equal zero. Recent studies showed that an efficient encoding scheme for target outputs could improve the generalization capability of a neural network [12]. Motivated by this work, we utilize interspeaker information on misidentification to design an encoding scheme for target outputs in supervised learning pairs.

To acquire interspeaker information on misidentification, we define a *confusion matrix* that properly describes both classification and misidentification results by GMMs on a validation set. The confusion matrix, CM, is in the following form:

$$\mathrm{CM} = \begin{pmatrix} n_{11} & n_{12} & \cdots & n_{1S} \\ n_{21} & n_{22} & \cdots & n_{2S} \\ \vdots & \vdots & \vdots & \vdots \\ n_{S1} & n_{S2} & \cdots & n_{SS} \end{pmatrix}.$$

Here n_{ij} denotes the number of the speech segments belonging to speaker i but assigned to speaker j by GMMs according to the decision strategy in (3). On the basis of the confusion matrix, we derive a conditional probability of misidentification that all the utterances, X_i, belonging to speaker i is assigned to speaker j

$$P(j \mid X_i) = \frac{n_{ij}}{\sum_{s=1}^{S} n_{is}}. \tag{6}$$

Here, $P(j \mid X_i)(j \ne i)$ is viewed as the misidentification rate of a model-based speaker identification system in terms of utterances belonging to a specific speaker. Thus, the confusion matrix (CM) provides the interspeaker information via the conditional probability, $P(j \mid X_i)$.

On the basis of the conditional probability, the target output $T(j \mid i)$ in supervised pairs is designed accordingly as follows:

$$T(j \mid i) = \begin{cases} 1, & \text{if } P(j \mid X_i) > 0, j = i \\ -1, & \text{if } P(j \mid X_i) > 0, j \ne i \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Equation (7) suggests a new encoding scheme, hereinafter called *rival penalized encoding rule* (RPER), based on interspeaker information. That is, the jth element in the target vector is assigned to $T(j \mid i)$ for an input vector resulting from any utterance belonging to speaker i. Thus, our encoding scheme tends to punish those GMMs resulting in misidentification while the correct decisions made by the GMMs remain.

In such an encoding scheme, target vectors are distinct to ensure that each speaker class has an unique target codeword $\mathbf{T}_s = [T(1 \mid s), \ldots, T(S \mid s)]^T$. In contrast to the one-of-S encoding scheme, moreover, the Hamming distance of codewords between the true speaker's and the impostor's is enlarged such that the misidentification by GMMs could be corrected more effectively. Based on the supervised learning pairs given above, the activation function of the neural network should be the hyperbolic tangent function. Here, we emphasize that our encoding scheme punishes only those GMMs leading to misidentification in terms of the decision strategy in (3) and distinguishes from the traditional one-of-S encoding scheme where all the components except the right one are punished.

In terms of our RPER encoding scheme, the decision strategy of the neural network during testing is as follows. Given an unknown speech segment, the neural network yields an output vector, $\mathbf{o} = [o_1, \ldots, o_S]^T$, via the outputs of GMMs. Then the distance between the output vector and each target codeword $\mathbf{T}_s (s = 1, 2, \ldots, S)$, needs to be calculated by the distance measure as

$$d(\mathbf{o}, \mathbf{T}_s) = \sum_{j=1}^{S} |o_j - T(j \mid s)|. \tag{8}$$

On the basis of the minimal distance, the unknown speech segment is assigned to a label of registered speakers, s^*, i.e.,

$$s^* = \operatorname*{argmin}_{1\le s\le S} d(\mathbf{o}, \mathbf{T}_s). \tag{9}$$

C. A Query-Based Learning Algorithm for Active Learning

In empirical studies on our method above, we find that a validation set plays an important role for sufficient use of interspeaker information with a neural network. However, neural network is traditionally treated as a passive learner with randomly chosen inputs, which might limit the generalization capability of our method. Recent studies [11], [15] showed that query-based learning could be used for active learning. The basic idea underlying the query-based learning is that the next input depends on those used previously so as to improve the generalization ability of the learner. For our purpose, we present a query-based learning algorithm to actively select the input data of interest during training.

In the query-based learning, new queries are chosen according to some heuristics. One useful hint is to choose those queries that maximize the expected information content, e.g., work in [15], [27]. Motivated by the idea, we select those input data around the decision boundaries as the candidates of next queries since these samples have uncertainty in decision-making and, therefore, they could provide maximal information in determining the correct decision boundaries. As argued in the statistical learning theory [28], such samples likely play a more important role in generalization.

As a consequence, our query-based learning algorithm is described as follows.

1) Train the neural network first on the output space of GMMs by using 10% samples randomly selected from a given validation set.
2) For each remaining sample belonging to speaker s in the validation set, $\tilde{\mathcal{L}}^{(l)}$ $(l = 1, \ldots, L_s)$, the neural network trained in step 1) yields an output vector $\mathbf{o}^{(l)}$. Calculate the distance between the output and the target codeword corresponding to speaker s, $d^{(l)}(\mathbf{o}^{(l)}, \mathbf{T}_s)$, in terms of (8).
3) Select the sample of maximal distance as next query, which could characterize the shape of decision boundaries

$$l^* = \underset{1 \le l \le L_s}{\operatorname{argmax}}\, d^{(l)}(\mathbf{o}^{(l)}, \mathbf{T}_s).$$

4) Append the selected data to the validation set and retrain the neural network on the resulting validation set.
5) Repeat steps from 2) to 4) until the prespecified number is reached.

Through the data selection by our query-based learning algorithm, a new validation set is formed for training the neural network, which could lead to better generalization as demonstrated in the next section.

IV. Simulations

In this section, we present comparative results on the KING speech corpus [5] to demonstrate the effectiveness of our method proposed in Section III.

This corpus consisting of wide-band and narrow-band sets is a benchmark acoustic database especially for text-independent speaker identification. The wide-band set was collected with a high quality microphone in a quiet room, while the narrow-band set was collected by telephone handset through a long distance telephone channel. In each set, all speakers are male and ten sessions for each speaker were recorded from a week to a month apart. It is reported [5] that some data in the wide band set were unfortunately missing, which results in different population in two sets.

As handled in [22], the preprocessing for the text-independent speaker identification system is performed as follows: 1) preemphasizing with filter response $H(z) = 1 - 0.95z^{-1}$; 2) 32 ms Hamming windowing without overlapping; 3) removing the silence and unvoiced part of speech in terms of short-term average energy; and 4) extracting 16-order Mel-scaled cepstral feature vector from each short-term frame. All the simulations are performed on a PC (Pentium III 500) of the Microsoft Windows'98 platform.

Results in [8], [22], and [24] indicate that GMM performs quite well for text-independent speaker identification. In simulations, we adopt GMM as a representative of model-based approaches to build a model-based speaker identification system. Then a three-layered perceptron is employed to capture interspeaker information from output space of GMM. For comparison, a neural-network based speaker identification system is also built, where a multilayer perceptron (MLP) is created based on feature vectors extracted from utterances of all the speakers in the original validation set without active learning.

The performance of speaker identification systems in this paper is defined as misidentification and identification rates, i.e.,

$$\begin{aligned}&\text{Misidentification Rate}\\ &\quad = \frac{\text{\# incorrectly identified segments}}{\text{\# total testing segments}} \times 100\%\\ &\text{Identification Rate} = 100\% - \text{Misidentification Rate.}\end{aligned}$$

In simulations, testing speech segments of a specific length are adopted to evaluate the overall performance of the MLP-based, GMM-based, and our hybrid speaker identification systems. In order to demonstrate the effectiveness of the RPER method, we also use the 1-of-S encoding scheme in our hybrid systems to train the three-layered perceptron. For comparison, moreover, a new validation set is dynamically produced by our query-based learning algorithm, while random selection is used during training.

A. Results on Wide-Band Set

In simulations on the wide-band set, we use utterances of 49 speakers collected in all ten sessions, $S01$–$S10$, for experiments. For evaluating our method thoroughly, we adopt the two-session training, where speech of around 70 s recorded in two sessions is used, and the single-session training, where speech of around 40 s recorded in a single session is used, to train 49 GMMs of 32 mixture components (the same structure is used in [22]). Obviously, the data recorded in multiple sessions cover more variabilities than that contained in a single session. To capture the interspeaker information from the output space of GMMs, a three-layered perceptron consisting of 49 input nodes, 60 hidden nodes, and 49 output nodes is employed based on a cross-validation procedure. A portion of training data and an alternative session constitute an original validation set, as a basis, to train the neural network. For each speaker, speech segments of this validation set are selected actively by our query-based learning algorithm to form a new validation set for better generalization. To simplify the presentation, the resulting validation set consisting of 60 speech segments (about 15 s) by active selection is denoted as AV60, and another resulting validation set consisting of 80 speech segments (about 18 s) by random selection is denoted as RV80. Thus, all the remaining utterances in this database are used for test. In addition, such a simulation is repeated by the use of data belonging to different sessions for reliability. Totally, ten trials are performed; five for the two-session training and five for the single-session training. In each trail, different sessions are chosen to form a training set and an original validation set. As a result, constitutions of ten trials are listed in Table I. For comparison, moreover, we also train another three layered perception consisting of 16 input, 40 hidden, and 49 output nodes (a cross-validation method has been adopted for model selection) directly on the original validation set without active learning.

Fig. 3 depicts an example to demonstrate how our method works during test in trial 1. All the speech segments belonging to speaker 1 on testing sets are fed to the GMM-based and the hybrid (RPER) systems. Misidentification caused by GMMs and that of the neural network are illustrated in Fig. 3. This example well demonstrates how misidentification caused by six GMMs has been lowered by the three-layered perceptron to different extents.

Now we report the detailed testing results in trail 1. Table II summarizes identification rates of the MLP-based, GMM-based, and our hybrid systems on different testing sessions. From Table II, we observe that the MLP-based speaker identification system performs quite poor since the MLP, working directly on the speech feature space, could encounter a more complicated decision-making problem along with the catastrophic or cross-talk effects during training. It is evident from simulation results that our hybrid systems outperform both the MLP-based and GMM-based speaker identification systems no matter which target encoding scheme is used. In particular, our RPER encoding scheme leads to better generalization for the three-layered perceptron in contrast to the 1-of-S encoding scheme. Note that the neural network trained with the 1-of-S encoding scheme sometimes does not improve the performance of the GMM-based speaker identification system, such as for utterances in sessions $S07$ and $S10$, while our encoding scheme consistently leads to improvements in all the sessions. This result empirically shows that our encoding method alleviates the catastrophic effects during training. From the comparative results of the hybrid systems based on AV60 and RV80, it is shown that the active selection by our query-based learning algorithm achieves better overall performance than random selection. By our active selection, fewer training data are used but yield higher identification rates.

To demonstrate the role of our active learning in more detail, Fig. 4 depicts the evolutionary identification process as the number of queries increases. For comparison, the same process by a random selection is also shown in Fig. 4. It is observed from Fig. 4 that the active learner performs similar to a random selection initially. After 25 queries, however, the active learner achieves the considerably lower misidentification rates in contrast to the random selection. In general, more data or queries likely carry richer information unless appended data are redundant. Apparently, our simulation results are consistent with this statement; i.e., as the number of samples (queries) increases, the performance is improved no matter which kind of selection is used.

TABLE I
THE LIST OF CONSTITUTIONS IN TEN TRAILS (WIDE-BAND SET)

	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5
Training set	S01, S02	S03, S04	S05, S06	S07, S08	S09, S10
Validation set	S02, S03	S01, S03	S01, S06	S01, S07	S01, S09

	Trial 6	Trial 7	Trial 8	Trial 9	Trial 10
Training set	S01	S03	S06	S09	S10
Validation set	S01, S06	S01, S03	S01, S06	S01, S09	S01, S10

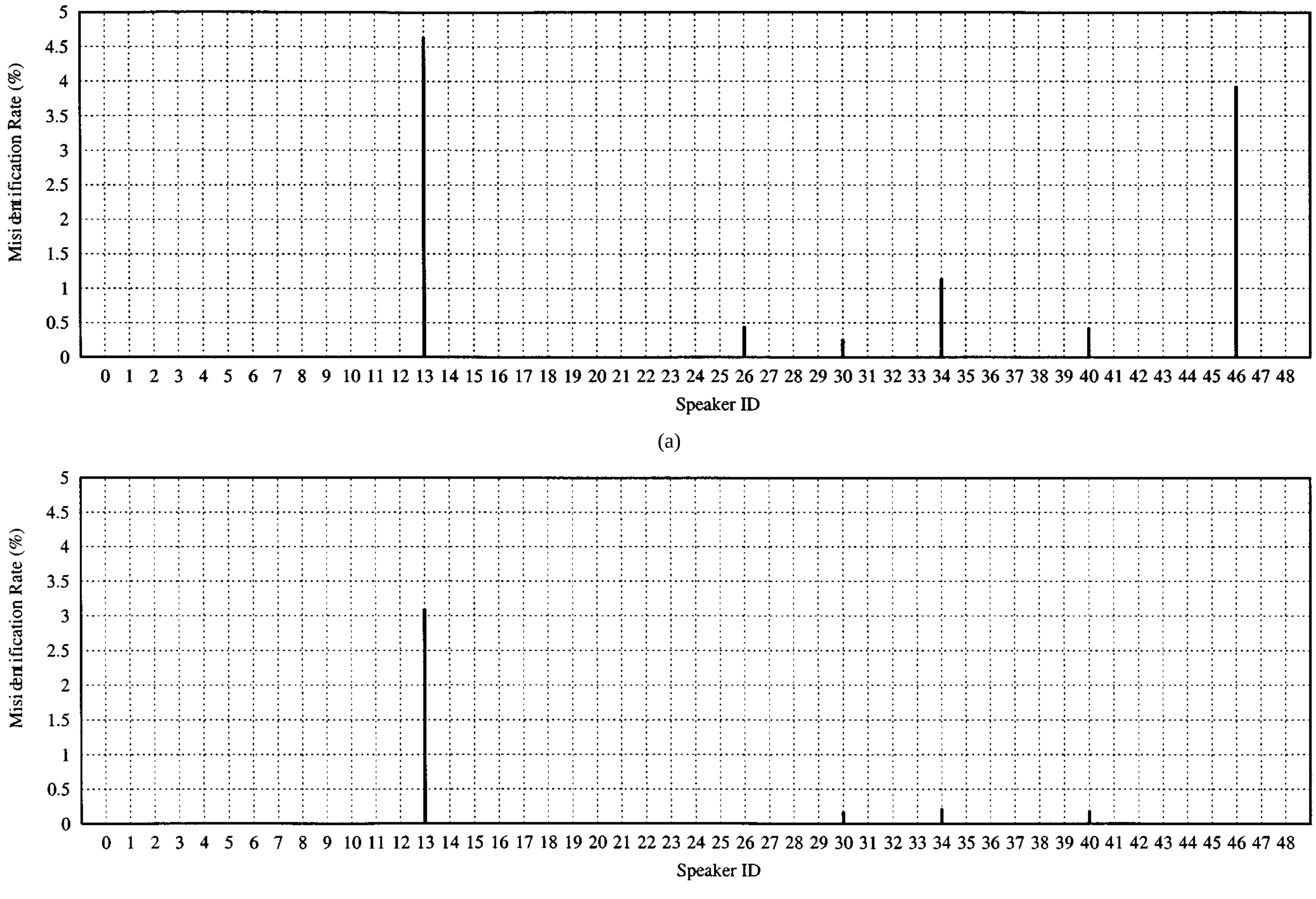
Fig. 5 shows the overall performance of all trials by averaging identifications rates of the GMM-based and our hybrid systems. For explicit comparison, we group the results into two groups in terms of training duration; one for those trails of the single-session training and the other for those trails for the two-session training. It is evident from Fig. 5 that our hybrid systems, trained with short training duration (40 s) of a single session, raise the averaging identification rate up to 4.1% in contrast to the GMM-based method. Similarly, our hybrid systems, trained with longer training duration (70 s) of two sessions, also result in improvements.

For further comparison, we also use testing speech segments of different lengths to evaluate the performance of the GMM-based and hybrid systems. Fig. 6 illustrates misidentification rates of the GMM-based and hybrid (RPER) speaker identification systems as testing lengths vary from 1.6 to 8.0 s for comparison. From Fig. 6, our method results in continuous improvements regardless of lengths of testing speech segments.

Finally, we compare the CPU time for training two hybrid systems with that for training the MLP-based and GMM-based systems. The two hybrid (RPER) systems are based on RV80 and AV60, respectively. Fig. 7 illustrates training time taken by different methods. From Fig. 7, it is observed that the hybrid systems take much shorter time than an MLP-based system. It implies that our method provides a more efficient way to use neural networks for speaker identification though other efforts can be made as well to reduce the computational load, e.g., the work in [13] and [21]. It is observed that the active selection by our query-based learning algorithm spends slightly longer time than the random selection. Nevertheless, it appears logical because each query in our active selection is produced by calculating the outputs of all unused samples in input space but such calculation is not needed in the random selection.

B. *Results on Narrow-Band Set*

In the narrow-band set, the limited bandwidth and distorted transmission channel cause speech quality to be degraded severely. In particular, there are differences in spectral characteristics between sessions $S01$–$S05$ and sessions $S06$–$S10$, because



(a)

(b)

Fig. 3. Misidentification rates of the GMM-based and the Hybrid (RPER) systems for all the testing utterances belonging to speaker 1 (wide-band set, trial 1). (a) Misidentification rates produced by the GMM-based speaker identification system. (b) Misidentification rates produced by the Hybrid (RPER) system.

TABLE II
IDENTIFICATION RATES (%) OF THE MLP-BASED, GMM-BASED, AND OUR HYBRID SPEAKER IDENTIFICATION SYSTEMS FOR TESTING SPEECH SEGMENTS OF 8 s (WIDE-BAND SET, TRIAL 1)

Method	Validation	Testing Set							
	Set	S04	S05	S06	S07	S08	S09	S10	Average
MLP		75.23	70.05	60.10	61.47	63.26	55.64	60.92	63.81
GMM		89.61	90.76	87.58	91.21	91.46	84.97	93.00	89.79
Hybrid (1-of-S)	RV80	89.79	94.29	87.89	90.78	92.14	86.34	91.51	90.39
Hybrid (RPER)	RV80	91.46	96.43	88.67	91.35	93.65	89.39	94.14	92.15
Hybrid (RPER)	AV60	92.20	96.43	89.97	91.72	91.68	90.97	93.63	92.37

speech is passed through different local telephone channels [5]. Signal-to-noise ratio (SNR) for sessions $S06$–$S10$ is about 10 dB worse than that for sessions $S01$–$S05$.

For simulations on the narrow-band set, we adopt a preprocessing procedure similar to that for the wide-band set. Moreover, the mean subtraction technique [22] is applied in preprocessing and the weighted Mel-scaled cepstrum is further used for feature extraction [14], which results in the robustness to noise and degraded speech.

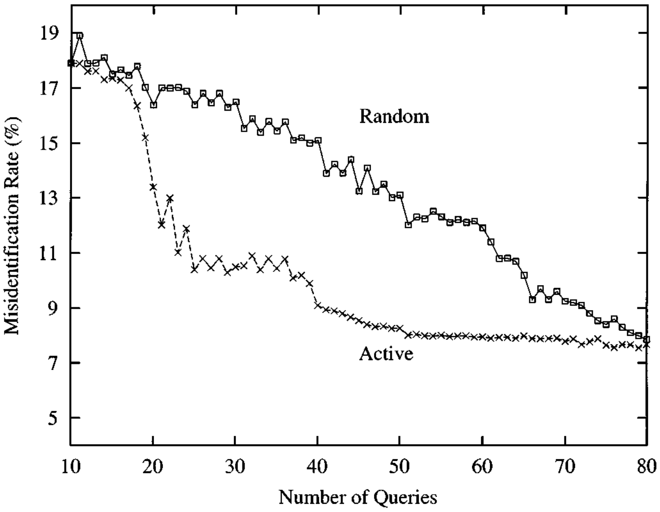


Fig. 4. Overall misidentification rates of hybrid (RPER) speaker identification system on testing sets (wide-band set, trial 1) by our query-based learning algorithm and a random selection as the number of queries increases. Here, speech segments of 8 s are used for test.

There are 51 speakers with all ten sessions used in our simulations. In our simulations, two trials are performed. Accordingly, 51 GMMs of 32 mixture components are employed to model these speakers. In trial 1, utterances of 100 s recorded in sessions $S01, S02$, and $S03$ are used to train GMMs. In trial 2, utterances recorded in sessions $S05, S06$, and $S07$ constitute a training set of the same duration. In two trails, a three-layered perceptron consisting of 51 input nodes, 60 hidden nodes, and 51 output nodes is employed to capture the interspeaker information based on the original validation set consisting of speech

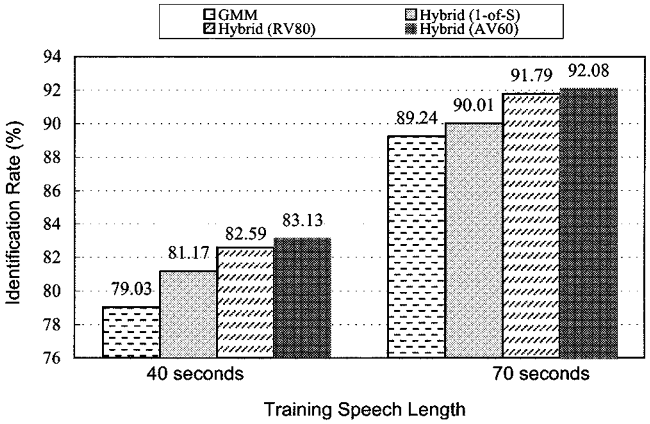


Fig. 5. Overall identification rates of GMM-based and Hybrid speaker identification systems based on GMMs trained with speech of different lengths in terms of the two-session and the single-session training (wide-band set). Here, speech segments of 8 s are used for test.

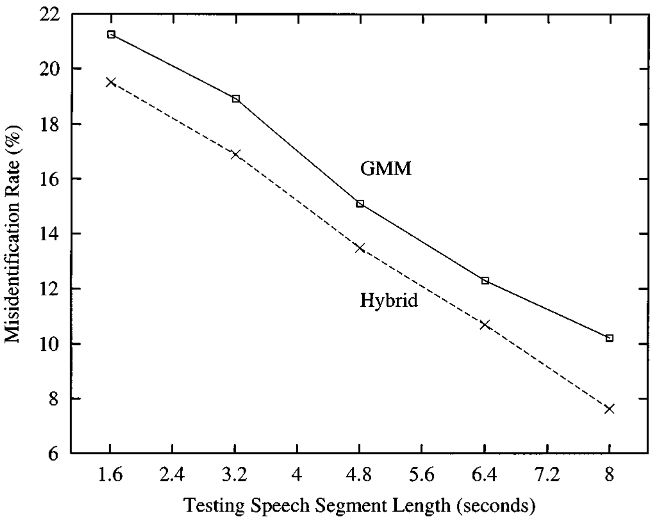


Fig. 6. Overall misidentification rates of the GMM-based and the hybrid (RPER) speaker identification systems in terms of different testing speech segments spanning from 1.6 to 8.0 s (wide-band set, ten trials).

belonging to sessions $S01$ and $S06$. Similarly, our query-based learning algorithm is used to form a new validation set during training. Other six sessions are used for test. For comparison, the same training set in each trail is also used to train an MLP-based speaker identification system of the structure consisting of 16 input, 40 hidden, and 51 output nodes.

Tables III and IV summarize identification rates of our hybrid systems on testing sets in two trails along with the corresponding results of the MLP-based and GMM-based systems for comparison. It is evident for simulation results that our hybrid (RPER) systems lead to significant improvements in comparison with the MLP-based, GMM-based and hybrid (1-of-S) systems, even though the SNR in some sessions, e.g., session $S07$, is lower than 20 dB. Moreover, simulation results indicate that our query-based learning algorithm performs better than random selection, even with fewer training data (60 segments by our active selection versus 80 segments by the random selection).

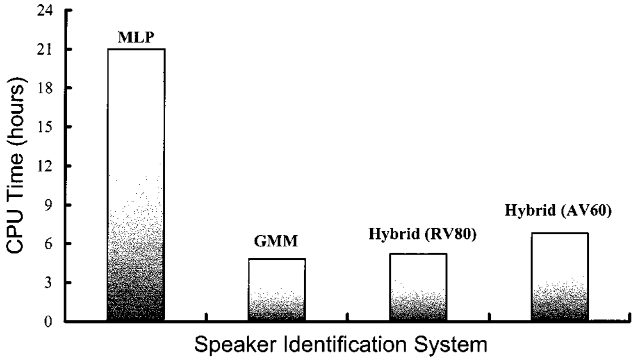


Fig. 7. The CPU time for training of the MLP-based, GMM-based, and hybrid (RPER) speaker identification systems in terms of the two-session training (wide-band set).

TABLE III
IDENTIFICATION RATES (%) OF THE MLP-BASED, GMM-BASED, AND OUR HYBRID SPEAKER IDENTIFICATION SYSTEMS FOR TESTING SPEECH SEGMENTS OF 8 s (NARROW-BAND SET, TRIAL 1)

Method	Validation	Testing Set						
	Set	S04	S05	S07	S08	S09	S10	Average
MLP		51.49	50.48	46.62	45.89	44.71	41.84	46.83
GMM		88.12	72.50	47.40	60.41	61.90	62.67	65.50
Hybrid (1-of-S)	RV80	85.47	73.29	51.12	62.95	69.78	69.51	68.68
Hybrid (RPER)	RV80	88.21	75.68	57.87	67.04	69.33	67.18	70.88
Hybrid (RPER)	AV60	88.21	79.28	63.01	64.94	70.13	70.65	72.69

TABLE IV
IDENTIFICATION RATES (%) OF THE MLP-BASED, GMM-BASED, AND OUR HYBRID SPEAKER IDENTIFICATION SYSTEMS FOR TESTING SPEECH SEGMENTS OF 8 s (NARROW-BAND SET, TRIAL 2)

Method	Validation	Testing Set						
	Set	S02	S03	S04	S08	S09	S10	Average
MLP		51.23	52.32	60.19	55.34	47.98	48.26	52.55
GMM		60.45	70.91	77.08	77.22	72.62	68.81	71.18
Hybrid (1-of-S)	RV80	67.03	73.19	82.49	77.47	73.54	73.36	74.51
Hybrid (RPER)	RV80	70.37	76.26	82.61	79.61	74.40	73.62	76.15
Hybrid (RPER)	AV60	72.71	78.58	86.61	83.30	75.46	79.87	79.42

For overall performance in two trials, Fig. 8 depicts the averaging misidentification rates of the GMM-based and our hybrid (RPER) systems as testing lengths vary from 1.6 to 8.0 s. Apparently, our method consistently results in significant improvements. For further comparison between our active learning and random selection in trail 1, we also show their evolutionary identification process in Fig. 9. Although both active and random selection can reduce the misidentification rates as the number of queries increases, our active selection method leads to better generalization on this noisy database.

V. CONCLUSION

We have presented a novel connectionist method to improve a model-based speaker identification system by introduction of interspeaker information. Simulation results on the KING database show that our method leads to a considerable improvement for a GMM-based speaker identification system. The proposed encoding scheme based on interspeaker information results in

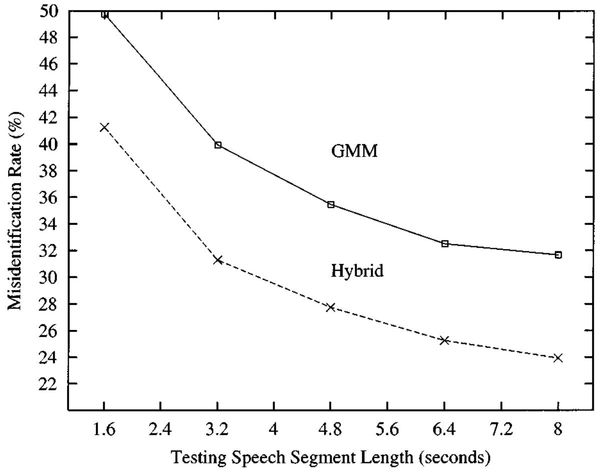


Fig. 8. Overall misidentification rates of the GMM-based and the hybrid (RPER) speaker identification systems in terms of different testing speech segments spanning from 1.6 to 8.0 s (narrow-band set, two trials).

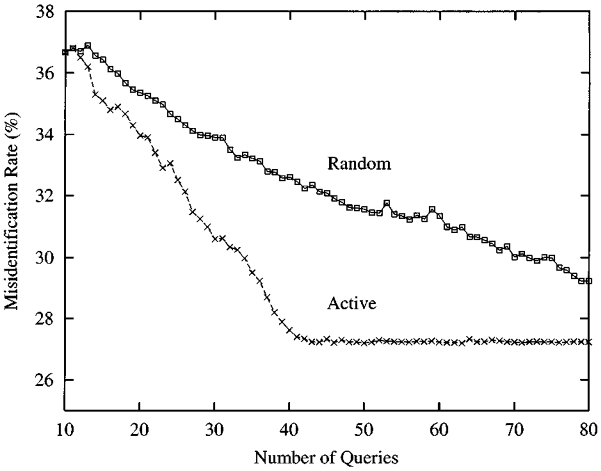


Fig. 9. Overall misidentification rates of hybrid (RPER) speaker identification system on different testing sets (narrow-band set, trial 1) by our active learning and random selection as the number of queries increases. Here, speech segments of 8 s are used for test.

better generalization in contrast to the traditional one-of-S encoding scheme, and our query-based learning algorithm can dynamically generate an effective validation set by active learning, which leads to better generalization. In addition, our method yields faster training in contrast to those connectionist speaker identification methods with classification directly on the speech feature space. Thus, our method would provide a fast way to update an speaker identification system once new speech data are available.

Appending a new user is a substantial task for any realistic operational system. It is unavoidable either in our hybrid system. Unlike other connectionist speaker identification systems, it is done in our method by appending a statistical speaker model and retraining the neural network separately. Thus, an open problem, how to update the neural network in our system fast, still remains to be studied in the future, though this neural network works on the output space of statistical speaker models where fewer data are required for training.

Essentially, our method can be viewed as an application of the stack generalization principle in machine learning [29]. In this sense, the methodology presented in this paper provides a framework to reduce misidentification in speaker identification by the use of interspeaker information regardless of the computational apparatus used in this paper. Our earlier work indicated that the use of alternative computational apparatus under this framework yields the satisfactory result [1]. On the other hand, the idea underlying our methodology could be also extended to handle some acoustic modeling problems; e.g., our method is expected to yield better discrimination between two similar phonemes.

Acknowledgment

Authors are grateful to anonymous reviewers for their comments which improve the presentation of this paper.

References

[1] W. Q. Bao, K. Chen, and H. S. Chi, "An HMM/MFNN hybrid architecture based on stacked generalization for speaker identification," in *Proc. Int. Joint Conf. Neural Networks*, Anchorage, AK, 1998, pp. 367–371.

[2] Y. Bennani, "A modular and hybrid connectionist system for speaker identification," *Neural Comput.*, vol. 7, no. 4, 1995.

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

[4] J. P. Campbell, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[5] A. D. Carlo, M. Falcone, and A. Paoloni, "Corpus design for speaker recognition," in *Proc. ESCA Workshop Automat. Speaker Recognition, Identification, Verification*, Martigny, Switzerland, 1994, pp. 47–50.

[6] C. Che and Q. Lin, "Speaker recognition using HMM with experiments on the YOHO database," in *Proc. Europ. Conf. Speech Commun. Technol.*, Madrid, Spain, 1995, pp. 625–628.

[7] K. Chen, "A connectionist method for pattern classification with diverse features," *Pattern Recognition Lett.*, vol. 19, no. 7, pp. 545–558, 1997.

[8] K. Chen, L. Wang, and H. S. Chi, "Methods of combining multiple classifiers with different features and their applications to text-independent speaker identification," *Int. J. Pattern Recognition Artificial Intell.*, vol. 11, no. 3, pp. 417–445, 1997.

[9] K. Chen, D. H. Xie, and H. S. Chi, "A modified HME architecture for text-dependent speaker identification," *IEEE Trans. Neural Networks*, vol. 7, pp. 1309–1313, Sept. 1996. for errata see *IEEE Trans. Neural Networks*, vol. 8, pp. 455, Mar. 1997.

[10] D. A. Cohn, "Neural network exploration using optimal experiment design," in *Advances in Neural Information Processing Systems 6*, J. Cowan, G. Tesauro, and J. Alspector, Eds. San Mateo, CA: Morgan Kaufmann, 1994.

[11] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *J. Artificial Intell. Res.*, vol. 4, no. 1, pp. 129–145, 1996.

[12] T. G. Dietterich, "Solving multiclass learning problems via error-correcting output codes," *J. Artificial Intell. Res.*, vol. 2, no. 2, pp. 263–286, 1995.

[13] K. R. Farrell, R. J. Mammone, and K. T. Assaleh, "Speaker recognition using neural networks and conventional classifiers," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 194–205, 1994.

[14] S. Furui, "Recent advances in speaker identification," *Pattern Recognition Lett.*, vol. 18, no. 9, pp. 859–872, 1997.

[15] M. Hasenjager and H. Ritter, "Active learning with local models," *Neural Processing Lett.*, vol. 7, pp. 107–117, 1998.

[16] A. Higgins, L. Bahler, and J. Porter, "Speaker verification using randomized phrase prompting," *Digital Signal Processing*, vol. 1, no. 1, pp. 89–106, 1991.

[17] J. N. Hwang, J. J. Choi, S. Oh, and R. J. Marks II, "Query-based learning applied to partially trained multilayer perceptrons," *IEEE Trans. Neural Networks*, vol. 2, pp. 131–136, 1991.

[18] T. Isobe and J. Takahashi, “A new cohort normalization using local acoustic information for speaker verification,” in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Phoenix, AZ, 1999, pp. 841–844.

[19] J. C. Junqua and J. P. Haton, *Robustness in Automatic Speech Recognition*. Boston, MA: Kluwer, 1997.

[20] D. J. MacKay, “Information-based objective functions for active data selection,” , vol. 4, no. 4, pp. 590–604, 1992.

[21] J. Oglesby and J. S. Mason, “Optimization of neural models for speaker identification,” in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, San Francisco, CA, 1990, pp. 261–264.

[22] D. A. Reynolds, “A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification,” Ph.D. dissertation, Dept. Elect. Eng., Georgia Inst. Technol., 1992.

[23] ——, “Speaker identification and verification using Gaussian mixture models,” in *Proc. ESCA Workshop Automat. Speaker Recognition, Identification, Verification*, Martigny, Switzerland, 1994, pp. 27–30.

[24] D. A. Reynolds, R. B. Dunn, and J. J. McLaughlin, “The Lincoln speaker recognition system: NIST EVAL2000,” in *Proc. Int. Conf. Spoken Language Processing*, Beijing, China, 2000, pp. 470–474.

[25] A. E. Rosenberg, J. Delong, C. H. Lee, B. H. Juang, and F. K. Soong, “The use of cohort normalized scores for speaker recognition,” in *Proc. Int. Conf. Spoken Language Processing*, Albuquerque, NM, 1992, pp. 599–602.

[26] A. E. Rosenberg and F. Soong, “Recent research in automatic speaker recognition,” in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. New York: Marcel Dekker, 1992, pp. 701–738.

[27] S. Thrun and K. Moller, “Active exploration in dynamic environments,” in *Advances in Neural Information Processing Systems 4*, J. Moody, S. Hanson, and R. Lippmann, Eds. San Mateo, CA: Morgan Kaufmann, 1992.

[28] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[29] D. H. Wolpert, “Stacked generalization,” *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

Lan Wang received the B.Eng. degree in electrical and electronic engineering from Beijing Institute of Technology, Beijing, China, in 1993, the M.Eng. degree in signal processing from the Peking University, Beijing, China, in 1996, and is pursuing the Ph.D. degree at Cambridge University, Cambridge, U.K.

Currently, she is a Research Associate at the Center for Information Science of Peking University. Her research interests include speech and speaker recognition by neural networks and speech quality evaluation in mobile communication.

Ke Chen (M’97–SM’00) received the B.S. and M.S. degrees from Nanjing University in 1984 and 1987, respectively, and the Ph.D. degree in 1990, all in computer science.

From 1990 to 1992, he was a Postdoctoral Researcher at Tsinghua University. During 1992–1993, he was a Postdoctoral Fellow of Japan Society for Promotion of Sciences and worked at Kyushu Institute of Technology. Since 1994, he has been on the Faculty of Peking University, where he is now an adjunct Professor. From 1996 to 1998, he was a Visiting Professor at The Ohio State University. During 2000–2001, he held a Visiting Researcher position in Microsoft Research Asia and a visiting professorship at Hong Kong Polytechnic University. He has been on the Faculty of The University of Birmingham since 2001. He has published more than 80 technical papers in refereed journals and international conferences. His current research interest includes statistical pattern recognition, machine learning with an emphasis on neural computation, and their applications to machine perception.

Dr. Chen was a recipient of the TOP Award for Progress of Science and Technology from the National Education Ministry in China, a recipient of NSFC Distinguished Young Principal Investigator Award and a recipient of “Trans-Century Talented Professional” Award from the China National Education Ministry. He is a member of the IEEE Computer Society, the IEEE Systems, Man, and Cybernetics, Society, and a member of INNS.

Huisheng Chi (M’81–SM’88) received the B.Sc. degree from the Department of Radioelectronics, Peking University, China, in 1964.

He joined the Department of Radioelectronics, Peking University, as an Assistant in 1964. In 1972, he led the University’s research efforts in developing a satellite communications system that utilized Spread Spectrum technology. He began researching the use of an Optical Fiber system with regards to telephone traffic, the first-ever application of this type of technology in China. Then, from 1983 to 1990, he worked on a large-scale research project that involved both Intelligent VSAT and TDM/CDMAA technology. In 1986, he joined the Chinese National Laboratory on Machine Perception as a Deputy Director and then as a Director. He is not only currently the Executive Vice President of Peking University and Vice Chairman of the University’s Academic Committee, but also acts as the University’s Ph.D. candidate advisor. Professor Chi has also published a number of important papers and books in the field, such as Digital Speech Signal Processing. His main research has focused on speech signal processing, telecommunications, and neural networks.

Prof. Chi acts also as the Vice Chairman of the Neural Network Council of China (CNNC), Vice Chairman of China Society of Geographic Information System (GIS), a Governor of INNS, Consultant of Microsoft Research (China), Academician of International Euro-Asia Academy of Science. In 1994 and 1995, he received the INNS “Neural Network Leadership Award.” In 1999, as a member of team in their research on the ANN, he won the first-class prize for promoting scientific and technological development from the Ministry of Education.