HYBRID NETWORK BASED ON RBFN AND GMM FOR SPEAKER RECOGNITION

Wei–Ying Li   *Douglas O'Shaughnessy*
INRS—Télécommunications, Université du Québec
16, Place du Commerce, Verdun, Québec, Canada H3E 1H6
wyli@inrs-telecom.uquebec.ca

EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

ABSTRACT

In this paper, a hybrid network based on the combination of Radial Basis Function Networks (RBFNs) and Gaussian Mixture Models (GMMs) is proposed and used for speaker recognition. The hybrid network is a hierarchical one, where a GMM is built for each speaker and an RBFN is built for each group of speakers. The GMMs and RBFNs are trained independently. The RBFNs are used as a first stage coarse classifier and the GMMs are used as the final classifier. For each RBFN, only the first several candidates are chosen to take part in the final classification. The hybrid system is used for the SPIDRE database speaker recognition. Some experiments were carried out to choose the proper structure and parameters of RBFNs and GMMs. After using RBFNs, about 40% speakers were excluded without decreasing the performance. If the most confusable speaker sets in GMMs are grouped into RBFNs, the performance of GMMs can be increased more by using RBFNs.

1. INTRODUCTION

In recent years, GMMs have been successfully used in speaker recognition [1]. The training procedure of GMMs is based on Maximum Likelihood Estimation (MLE), whose discriminative power is limited. With an increase of the number of the recognition classes, both the speed and performance of GMMs decrease. In order to increase the recognition performance, many new approaches have been proposed, among which neural networks based on discriminant training are the most promising methods [2, 3, 4]. Many experiments show that both the discriminant training methods and MLE training can provide good performance, if there are enough model parameters, there is sufficient training data, the priori probabilities are known, and the modeling assumptions fit the data distributions. However, if these conditions are not satisfied, discriminant training can provide better performance [2, 3, 4].

There are many kinds of neural networks which have been successfully used for speech and speaker recognition, among which RBFNs are being widely used because of their faster training speed [2, 3, 4]. Like some other kinds of neural networks, RBFNs can approximate any function, if there are sufficient hidden nodes and the model parameters are trained properly. However, if a single RBFN is used to recognize a large group of classes, its structure will become more complex and training become more difficult, which result in lower performance. In order to use the discriminant training abilities of RBFNs and to overcome their shortcomings, we proposed a hybrid system, which is based on the combination of GMMs and RBFNs.

2. THEORY OF RBFN AND GMM

2.1. Radial basis function networks

The RBFN is a kernel function classifier, which uses the local areas formed by some simple kernel functions to make the decision boundaries among the recognition classes. The RBFN is a three-layer-node feedforward network (two-layers of connection weights). The input nodes correspond to the input pattern. The number of the input nodes is equal to the dimension of the input vectors. Each hidden node is an RBF node, which evaluates the input pattern by using the kernel function, and takes the evaluated score as its output. Each hidden node has two parameters, center and width. There are many kinds of functions that can be used as the kernel function in the hidden layer of RBFNs. Experiments show that the form of the hidden nodes is not critical to the performance of RBFNs [2, 3, 4]. In our RBFNs, each hidden node is a Gaussian basis function,

$$\Phi_j(X) = Kexp(-\frac{1}{2}(X-\mu_j)^T C_j^{-1}(X-\mu_j)), \quad 1 \leq j \leq H,$$

where K is a constant, H is the number of hidden nodes, Φ_j is the output of the j-th hidden node for input X, C_j and μ_j are the covariance matrix and mean vector of the j-th hidden node, which corresponds to the center and width of a hidden node, respectively. If we suppose that in the P dimensional feature space, the different elements are independent, then the covariance matrix is diagonal.

$$\Phi_j(X) = Kexp(-\frac{(X_p-\mu_{jp})^2}{2\sigma_{jp}^2}), \quad 1 \leq j \leq H,$$

where the index p means the p-th element of the corresponding vector. In such a case, the distribution of $\Phi_j(X)$ is a hyper-ellipsoid in the P dimensional space. The variance of each dimension, which is the width of the Gaussian kernel, is determined by the length of the axis of each corresponding dimension. If we further suppose that the variances of different dimensions for the j-th base function are equal, then $\Phi_j(X)$ is a hyper-sphere in the P dimensional space. As for the Mel Frequency Cepstral Coefficients (MFCC) of speech signals, the variations of different dimensions are quite different, so it is more accurate to use a hyper-ellipsoid instead of a hyper-sphere to describe the distribution of each base function.

Each output node represents a recognition class, and the value of each output node is the weighted summation of the outputs of all the hidden nodes. The class whose output node gives the highest value is chosen as the recognition result. The output nodes fulfill the following functions:

$$y_i(X) = \sum_{j=0}^{H} W_{ij} \Phi_j(X), \quad 1 \leq i \leq O,$$

where O is the number of output nodes, W_{ij} is the connection weight between the j-th hidden node and the i-th output node, $\Phi_0(X)$ is the bias hidden node and $\Phi_0(X) = 1$. The parameters of the RBFNs can be trained layer by layer, with the centers and widths of the hidden nodes trained first and the connection weights between the hidden nodes and the output nodes trained later. This is just the reason why RBFNs can be trained faster than other neural networks. The first layer can be trained by any clustering method, for example, LBG or K-means clustering methods. The second layer can be trained by the Least Mean Square (LMS) algorithm or be treated as a singer layer perceptron and trained by Error Back Propagation (EBP). The training procedure of RBFNs is a discriminant one; patterns from each class should be used at the same time.

2.2. Gaussian mixture models

A GMM is a special Hidden Markov Model (HMM), which has only one state and no state transition parameters. The only state of the GMM is also called the output node. The value of the output node is the weighted summation of the outputs of each mixture, and represents an evaluation score of the GMM to the input pattern. Each GMM has several mixtures, which are Gaussian distribution functions, the same as the hidden nodes in an RBFN. For a GMM, its output and mixtures are defined by:

$$y(X) = \sum_{j=1}^{M} W_j \Phi_j(X),$$
$$\Phi_j(X) = \mathcal{N}(X, \mu_j, C_j), \quad 1 \leq j \leq M,$$

where $\Phi_j(X)$ is the output of the j-th mixture, $y(X)$ is the total evaluation score of input X, W_j is the mixture weight, M is the number of mixtures, $\mathcal{N}$ is the normal distribution, C_j and μ_j are the covariance matrix and mean vector of the j-th mixture, respectively. It can be seen that the form of a mixture in GMMs is similar to that of a hidden node in RBFNs.

For each GMM, the means and covariances of the mixtures and the mixture weights can be initialized by using the same clustering method as that used for training the hidden nodes of RBFNs. These values can then be optimized by using the Expectation-Maximization (EM) algorithm [1]. While the values of the connection weights in an RBFN can be positive, negative or zero, the values of the mixture weights in a GMM can only be positive. For each recognition class, a GMM can be built independently using only the training patterns of the class.

3. GMM/RBFN HYBRID SYSTEM

Though RBFNs have been used successfully when the number of recognition classes is small, it is difficult to build just one RBFN to recognize many classes. With an increase in the number of recognition classes, not only the training of an RBFN becomes quite difficult, but also its recognition performance decreases greatly. As for a GMM, its recognition rate also decreases when there are more recognition classes. If the error parts of GMMs and RBFNs do not overlap, then it is possible to improve the performance of GMMs by using RBFNs.

The hybrid system has two stages. The first stage consists of a set of RBFNs, each corresponding to a group of speakers. The RBFNs are used as coarse classifiers. After the test pattern passes the first stage, only the most possible candidates are selected to take part in the second stage match using GMMs. Thus, the number of recognition classes in the second stage reduces. Of course, sufficient candidates must be chosen in the first stage to guarantee that the correct class will be included in the second stage.

4. DATABASE AND FEATURES

The SPIDRE (Speaker Identification Research) database is used in our experiments. It consists of text-independent telephone speech and contains 45 target speakers (male and female). Each speaker has 4 conversations, two of them coming from the same handset and the others from different handsets. The maximum length of each conversation is 5 minutes. In the SPIDRE speech, the background is noisy and channel noise is high. When doing feature extraction, the length for frame analysis is 30 ms and the frame shift is 10 ms. The Hamming window is used. By using the frame energy, the silence parts of each conversation can be judged and skipped. For each frame, the 15 dimensional MFCCs and corresponding first order delta MFCCs are calculated. Since the frame energy may change for different speakers, different conversations and at different times, it may not be a stable factor for a speaker's information. So the energy element is omitted and only 14 dimensional MFCCs are used. For each speaker, one of the two conversations coming from the same handset of each speaker is randomly selected for training; the other 3 conversations were used for testing. The first 23 speakers are used and there is no overlap between the test patterns in our experiments.

5. EXPERIMENTS

Since the hybrid network is a combination of GMMs and RBFNs, it is important to increase the performance of both GMMs and RBFNs before building the hybrid network. Several experiments have been done to decide the proper structure and parameters of RBFNs and GMMs. For each RBFN in our experiments, all speakers have equal amounts of training speech, and the actual length used of each training conversation is the same, so the RBFN will not be biased on any particular speaker.

5.1. Number of hidden nodes in RBFNs and mixtures in GMMs

There is no theoretical guide for choosing the proper number of hidden nodes in RBFNs and the number of mixtures in GMMs. Using the same training sets and test sets, we tested the performance of RBFNs with different hidden nodes (from 64 to 512) and GMMs with different mixtures (from 16 to 160). The experiments show that for an RBFN whose number of input nodes and number of output nodes are given, a proper number of hidden nodes should be chosen. As for GMMs, when the number of mixtures increases, the performance does not vary much. Too few hidden nodes/mixtures may not be sufficient for modeling the distributions in the vector space, while too many may not be precisely estimated from a specific training database. The more the hidden nodes or mixtures, the longer the training and recognition time. When choosing the number of hidden nodes, the number of training patterns and the number of RBFN output nodes should also be taken into consideration. In our later experiments, 256 hidden nodes and 64 mixtures are used and for RBFNs and GMMs, respectively.

5.2. The lower limit of covariance

The minimum value of the covariances of the hidden nodes in RBFNs and mixtures in GMMs cannot be too small, otherwise the singularity can cause divergence of RBFNs or decrease the performance of RBFNs and GMMS [1]. We also tested the performance of RBFNs and GMMs with different values of the lower limit of covariance and found that to get better performance for our database, $0.01 \leq \sigma_{min}^2 \leq 0.03$ and $0.03 \leq \sigma_{min}^2 \leq 0.1$ should be set for GMMs and RBFNs, respectively.

5.3. Delta CEP features

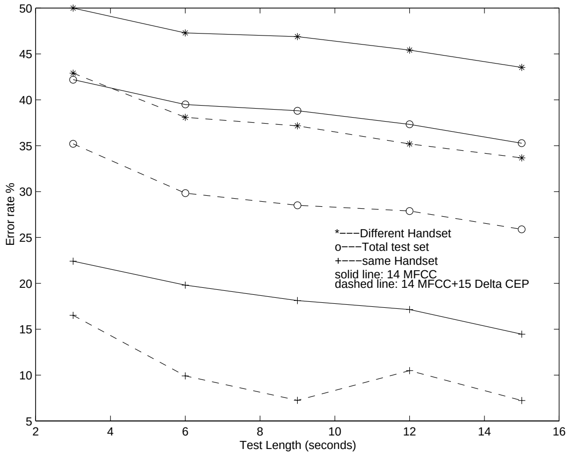

Figure 1 Error rates of GMMs with/without using delta CEP(8 speakers, handset adaptation)

Delta CEP is the differential version of the MFCC, which characterizes the variation of cepstral parameters along the time axis and is also called transitional features. When using only delta CEP in our RBFN, the performance is quite lower. Many experiments show that using dynamic features (delta CEP) together with static features (CEP) can improve the performance of a speech recognition system. Some comparative experiments were carried out to see if the additional delta CEP can improve the performance of GMMs and RBFNs. From Fig. 1, it can been seen that after using the delta CEP, the performance increased for both the different handset and same handset case. The same results were obtained for RBFNs. The performance of RBFNs does not increase as much compared with that of GMMs.

5.4. Cross-training

When only one conversation is used for training, the performance is not high. In order to see if the performance can be increased by adding more training patterns, especially different handset patterns, a conversation in the test set coming from a different handset than the original training conversation was randomly selected and added to the training set. Fig. 2 shows the results of GMM with/without cross-training.

After cross-training, the performance of the same handset decreased, while the performance for the whole test set increased. The improvement of performance of the test set results from the great increase of the different handsets. The great difference between the performance of same handset and different handsets was greatly decreased by using cross-training. Similar results were obtained for RBFNs.

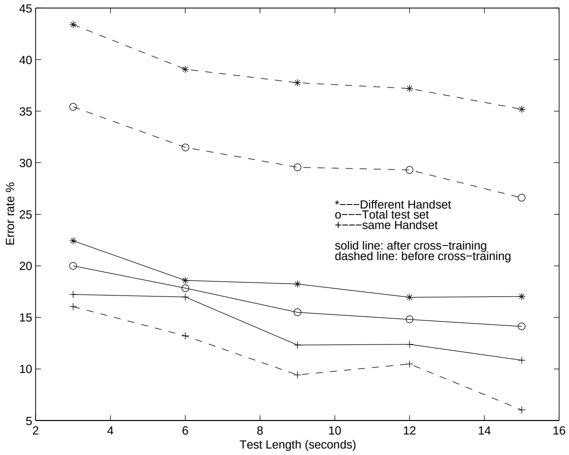

Figure 2 Error rates of GMMs with/without cross training(8 speakers, handset adaptation, 29 dimensional data)

5.5. Handset adaptation

For both GMMs and RBFNs, the performance of different handsets was lower. In order to increase the recognition rate of different handsets, we used handset adaptation for all conversations. For each conversation, after silence parts are omitted, an average vector (mean) of the conversation is calculated and subtracted from each frame. The results of GMMs before/after the handset adaptation are shown in Fig. 3.

After adaptation, the performance for GMMs increased. The difference between the performance of the same handset and the different handsets does not decrease much. While doing mean vector subtraction, the speaker's information may also be eliminated to some extent. The ability of this kind of handset adaptation is limited.

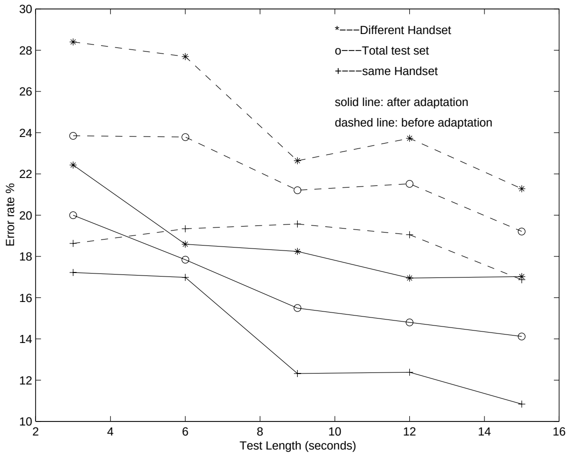

Figure 3 Error rates of GMMs with/without handset adaptation (8 speakers, cross-training. 29 dimensional data)

5.6. GMMs/RBFNs

The more the output nodes in each RBFN, the better its discriminative power can be used, and at the same time, the more difficult to train the RBFN and the lower its performance. Since GMMs can not recover the inclusion error of RBFNs, it is quite important for the correct class to be chosen by RBFNs in the first stage. We compared the inclusion rate and the complexity of RBFNs with different number of output nodes, and find that RBFNs with smaller number of output nodes are easier to train. In the hybrid network, 5 RBFNs were used with each has 5 output nodes except the last one has 3. The 23 speakers in SPIDRE were grouped into 5 groups (i.e.,RBFNs) according the sequence the speaker appeared.

A GMM was built for each speaker and an RBFN was built for each group of speakers. In the hybrid networks, the training sets and testing sets used in RBFNs and GMMs are same. The 14 CEPs and 15 delta CEPs were used and the handset adaptation was only used on the 14 CEPs part. For the RBFNs with 5/3 output nodes, when the first 4/2 candidates were selected to take part in the second stage classification, the performance of the hybrid network is the same as that of using only GMMs.

Fig. 4 shows the error rates of GMMs with/without using RBFNs when using crossing-training and choosing the first 3/2 candidates in RBFNs. The performance of the hybrid net dropped a little for short test patterns, and increased or did not change for longer test patterns compared with that of using only GMMs. Here about 40% recognition classes were excluded in the first stage classification. In our experiments, the speakers were grouped into the RBFNs by sequence. If the speakers, which are difficult to be distinguished by GMMs, are grouped into the RBFNs, it is possible to increase the overall performance.

In our experiments, the performance of GMMs or RBFNs on the training sets is always high. For a few test conversations, most of their testing patterns cannot be recognized correctly by GMM or RBFN. Some different-handset conversations give low error rates, but other same-handset conversations give high error rates. Although some speakers claimed that they used the same handset, actually they might have used different ones if there were several telephone handsets in their house. This may be the reason for the high error rates of some same-handset conversations.

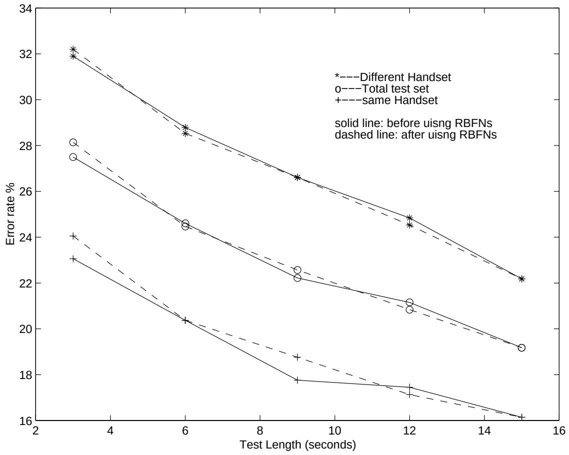

Figure 4 Error rates of GMMs with/without using RBFNs(23 speakers, handset adaptation, cross-training, 29 dimensional data)

6. CONCLUSIONS

In this paper, a GMM/RBFN hybrid network was proposed and used for speaker recognition. Some experiments were carried out to choose the proper structure and parameters of RBFNs and GMMs. Without decreasing the performance, about 40% recognition classes were excluded by using RBFNs. If the most confusable speaker sets in GMMs are grouped into RBFNs, the performance of GMMs can be increased more greatly.

References

[1] D. A. Reynolds, et al, *Robust text-independent speaker identification using Gaussian mixture speaker models.* IEEE Trans. on Speech & Audio Processing, Vol.3, No. 1, pp. 72-83, Jan. 1995.

[2] E. Singer, et al, *A speech recognizer using radial basis function neural networks in an HMM framework.* ICASSP'92, pp. 629-632.

[3] R. Lippmann, et al, *Hybrid neural network HMM approaches to wordspotting.* ICASSP'93, pp. 565-568.

[4] S. E. Fredrickson, et al, *Text-independent speaker recognition using neural network techniques.* Fourth Intl. Conf. on ANN, pp. 13-18, June 1995.

[5] W. Reichl, et al, *A hybrid RBF-HMM system for continuous speech recognition.* ICASSP'95. pp. 3335-8 vol. 5.

[6] M. Ceccarelli, et al, *Sequence recognition with radial basis function networks: experiments with spoken digits.* Neurocomputing, Vol: 11 No. 1 pp. 75-88 May 1996.

[7] R. Renals, et al, *Phoneme classification experiments using radial basis functions.* Proc. Intl. Joint Conf. on Neural Networks, Vol. 1, pp. 461-467, June 1989.