# A NEURAL NETWORK CLUSTERING TECHNIQUE FOR TEXT-INDEPENDENT SPEAKER IDENTIFICATION

**ZAKI B. NOSSAIR AND STEPHEN A. ZAHORIAN**
*Department of Electrical and Computer Engineering*
*Old Dominion University, Norfolk, VA 23529*

***ABSTRACT:***

A clustering algorithm for speaker identification based on neural networks is described. This technique is modeled after a previously developed technique in which an N-way speaker identification task is partitioned into N*(N-1)/2 two-way classification tasks. Each two-way classification task is performed using a small size neural network which is a two-way, or pair-wise, network. The decisions of these two-way networks are then combined to make the N-way speaker identification decision (Rudasi and Zahorian, 1991 and 1992). Although very accurate, this method has the drawback of requiring a very large number of pair-wise networks. In the new approach two-way neural network classifiers, each of which is trained only to separate two speakers, are also used to separate other pairs of speakers. Thus, in effect, speakers are clustered according to each pair-wise classifier. This method is able to greatly reduce the number of pair-wise classifiers required for making an N-way classification decision, especially when the number of speakers is very large. For 100 speakers extracted from TIMIT database, we were able to reduce the required number of pair-wise classifiers by a factor of 5, with no degradation in performance when 2 seconds or more of speech are used for identification. We obtained 100% text-independent speaker identification accuracy for 200 speakers with approximately 6 seconds of speech from each speaker and 97% when 2 seconds of speech were used.

## INTRODUCTION

There are two well-established techniques for speaker recognition/identification. The first technique is based on VQ (Soong et al., 1985; Matsui and Furui, 1991) and the second is based on neural networks (Bennani and Gallinari, 1991; Rudasi and Zahorian, 1991 and 1992). The neural network based technique, although very accurate, has the drawback that when a large number of speakers (i.e., classes for a pattern recognizer) is considered, the training time required by the network becomes prohibitive. Additionally, the required amount of training data becomes very large. For this reason, some investigators partition the speaker identification task into a number of small tasks. Each of these small tasks requires a small size network which can be trained in a shorter amount of time and with less training data (Bennani and Gallinari, 1991; Rudasi and Zahorian, 1991 and 1992). One of these partitioning techniques is called binary pair partitioning (BPP) (Rudasi and Zahorian, 1991 and 1992). This BPP approach partitions an N-way speaker identification task with N*(N-1)/2 pair-wise classification tasks. Each of these pair-wise classification tasks is performed using a "small" neural

network. Each of these pair-wise networks is trained to separate only two speakers. That is, each pair-wise network is trained using speech data from the two speakers for whom the network is intended to separate. The decisions of these pair-wise networks are then combined to make the N-way decision. The pair-wise decisions are made on a frame by frame basis of speech and the decisions of each pair-wise network are averaged over all frames to determine one decision from each network. For the N-speaker identification task there are N*(N-1)/2 pair-wise decisions. From these pair-wise decisions there are N-1 decisions which are relevant to a certain speaker. The relevant decisions for each speaker are then averaged and used as an estimate for the a posteriori probability of that speaker. The advantage of using this BPP technique relative to the use of a single large neural network is that it significantly reduces the training time and requires less amount of speech per speaker for training. The disadvantage of this technique is that it requires a large number of pair-wise classifiers. The purpose of this paper is to introduce a clustering technique for reducing the number of pair-wise networks required by the BPP approach. This clustering technique will also be referred to as clustered binary pair partitioning (CBPP).

The following sections provide an explanation for the clustering technique and the experimental results obtained in evaluating that technique.

## CLUSTERING TECHNIQUE

In this clustering approach we make use of the similarity among subsets of speakers to reduce the number of pair-wise networks needed. To do this, we start by arbitrarily selecting the first two speakers in our speaker population and then train a network to separate these two speakers, using only the available training data for these two speakers. This trained network is then evaluated as to how well it can separate the other possible pairs of speakers in our population, using the training data of these speakers. A trained network is considered sufficient to separate other pairs of speakers if its performance, on the training data of these pairs of speakers, exceeds a certain threshold. This trained network is then used to replace those pair-wise networks which would have been required by the BPP approach to separate those pairs of speakers. Thus the networks which would have been needed for separating those pairs of speakers are eliminated. We then train another pair-wise network which was not eliminated by any of the previously trained pair-wise networks. Then we use that newly trained network to eliminate other pair-wise networks as described above. This process of training a network and eliminating or replacing other networks is iterated until all pair-wise networks are covered. Thus, in effect, speakers are clustered according to each pair-wise classifier. By this clustering method some of the trained pair-wise networks are able to replace hundreds of the pair-wise networks that would have been required by the BPP approach.

An important refinement on the basic algorithm as described above is that each newly-trained network is tested relative to all possible speaker pairs, including pairs for which there is an already trained network. If the newly-trained network is able to better separate two speakers than the previously selected network, it replaces the previous network relative to that speaker pair. This process may also completely eliminate the need for some of the initially trained networks. It also insures that the trained networks are used for best effectiveness and helps eliminate potential bias due to the ordering of the speakers.

## EXPERIMENTS

In order to evaluate this clustering method and compare it with the BPP approach, several experiments were conducted. The main goal of these experiments was to show that, for a large number of speakers, the clustering method reduces significantly the number of pair-wise networks with almost no degradation in identification accuracy and that the percent reduction in the number of pair-wise networks increases as the number of speakers increases. For all experiments each pair-wise network was a memoryless, feed-forward, multi-layer perceptron and was configured to have one hidden layer of 10 nodes and one output node. Ten hidden nodes were selected from pilot experiments. Backpropagation was used for training these networks with 75000 network updates using an initial learning rate of 0.2. The learning rates was reduced by a factor of .96 every 5000 network updates. A momentum term of .6 was used.

The TIMIT speech database (Lamel et al., 1986) was used for our experiments. This data base contains 10 sentences for each of 630 speakers and was sampled at a 16 kHz sampling rate. Five of these 10 sentences are phonetically-balanced sentences and are called SX sentences. Three of these 10 sentences are phonetically-diverse sentences and are called SI sentences. The other two sentences are dialect sentences and are called SA sentences. In all of our experiments seven sentences (5 SX sentences and 2 SI sentences) of each of the speakers were used for training and the other three sentences were used for evaluation.

In all experiments 29 cepstral coefficients ($CC_1$ to $CC_{29}$) were used for each speech frame. These cepstral coefficients were computed as follows. First a second order high frequency pre-emphasis filter given by:

$$y[n] = 0.3426\, x[n] + 0.4945\, x[n-1] - 0.64\, x[n-2] \qquad (1)$$

was applied to the speech signal. Pilot experiments demonstrated that this pre-emphasis, which has a peak at approximately 3 kHz, and which is a reasonably good match to the inverse of the equal-loudness contour, results in slightly better performance than does a first order pre-emphasis $y[n] = x[n] - .95\, y[n-1]$. The second step was to compute a 1024 point FFT from each 32 ms Kaiser-windowed (coefficient of 5.33) frame of speech data where the window was applied every 20 ms. The window length and shaping coefficient were chosen from pilot experiments. The next step in processing was to compute the amplitude spectrum, logarithmically scale it, and then frequency warp it with bilinear warping function using a warping coefficient of .25. The next step was to compute a cosine transform of the scaled magnitude spectrum. The cosine transform coefficients were computed over the frequency range 0 to 8000 Hz. These cosine transform coefficients are the cepstral coefficients. Thirty cosine transform coefficients ($CC_0$ to $CC_{29}$) were computed for each frame. The first coefficient $CC_0$ was eliminated and the other 29 coefficients were used to represent each frame. This number of coefficients per frame were determined from pilot experiments.

### Experiment I

This experiment was conducted to determine a threshold value to be used with the clustering approach, to show how the identification accuracy changes as the threshold value changes, and to show the number of pair-wise networks needed with each

threshold value when the number of speakers is fixed. The experiment was conducted using 100 speakers with the threshold value changed from 0.55 to 0.75 in steps of 0.05. Note that in this application, a neural network output of .5 implies no discrimination between the two speakers of a pair, whereas an output of 1.0 (or 0.0) implies perfect discrimination. Figure 1 shows the identification accuracy for these 100 speakers as a function of the amount of speech used for identification from each speaker and for different values of the threshold. The figure shows that when the threshold is higher than 0.65, there is no significant improvement in performance. Since the number of networks does increase as the threshold increases, this indicates that 0.65 is a suitable value to use for the threshold.

**Experiment II**

The goal of this experiment was to compare the performance of the clustering approach with that of the BPP approach. To do this comparison, we used each of the two approaches to classify 100 speakers (67 males and 33 females). For both


**Figure 0. Identification accuracy for 100 speakers as a function of amount of speech used for evaluation and for different threshold values.**

approaches 7 sentences (5 SX and 2 SI) of each of the speakers were used for training and 3 sentences (2 SA and 1 SI) were used for identification. For both methods the configuration of the pair-wise networks was identical. For the clustering approach a threshold of 0.65 was used. The two approaches were compared for several lengths of test speech. Figure 2 shows the performance of these two approaches as a function of the amount of speech used for identification. As we can see from this figure, the two approaches have identical performance when 2 seconds of speech or more are used from each speaker. However, for the clustering approach 984 pair-wise networks were required versus 4950 networks for the BPP approach. Thus, for the case of 100 speakers, the clustering approach resulted in about a 5 to 1 reduction in the number of networks.

**Experiment III**

This experiment was to show that the percent reduction in the number of required pair-wise networks increases as the number of speakers increases. For this purpose, we conducted several sub-experiments using the CBPP for various number of speakers to determine the number of networks required in each case. We changed the number of speakers from 25 to 200 speakers. The number of networks required by the BPP approach are, of course, determined by the number of speakers. Figure 3 shows the number of networks required by each of the two approaches as a function of the number of speakers. A threshold of .65 was used for the CBPP method. As the figure shows, the number of networks required by the clustering approach increases approximately linearly with the number of speakers, but with the BPP approach the number of networks increases with the square of the number of speakers. If 1000 speakers were used, the BPP approach would require 499,500


**Figure 0. Performance of the two partitioned neural network approaches versus the amount of speech used for evaluation.**


Figure 0. The number of networks required for each of the two partitioned neural network approaches as a function of the number of speakers.

networks whereas the clustering approach would require approximately 15,000 networks, or about 3% of the BPP networks.

**Experiment IV**

This experiment was conducted to evaluate the performance of the clustering approach when a large number of speakers is considered. For this purpose the CBPP system was trained for 200 speakers. A threshold of .65 was used. A total of 2907 pair-wise networks were computed compared to 19900 networks would have been computed by the BPP approach. Figure 4 shows the performance achieved by the system for the 200 speakers as a function of the amount of speech used for evaluation. As we can see from the figure, the system achieved 100% identification accuracy when 6 seconds of evaluation speech were used from each speaker and 97% when 2 seconds of speech were used. These results are very comparable or even higher than that obtained by other investigations (Bennani and Gallinari, 1991; Matsui and Furui, 1991).

## CONCLUSION

The clustering technique described in this paper has proven to be very effective in reducing the number of pair-wise networks required for an N-way speaker identification task compared to the BPP approach. The performance obtained using this clustering technique, using a threshold of .65, is almost identical to that of the BPP approach when the amount of speech used for evaluating each speaker is 2 seconds or more. The experiments have also shown that the percent reduction in the


**Figure 0. Performance achieved by the CBPP approach, for a 200 speaker identification task, as a function of the amount of speech used for evaluation.**

number of networks required increases as the number of speakers in the identification task increases.

For a 200-speaker identification task we obtained 100% accuracy when 6 seconds of speech were used. These results are very comparable or even higher than that obtained by other investigations.

## ACKNOWLEDGEMENT

This work was supported by NSF grant IRI-9217436.

## REFERENCES

Bennani, Y. and Gallinari, P. (1991), "On The Use Of TDNN-Extracted Features Information In Talker Identification," Proc. ICASSP-91, pp. 385-388.

Lamel, L. F., Kassel, R. H., and Seneff, S. (1986), "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," Proc. DARPA Speech Recognition Workshop, Report No. SAIC-86/1546, pp. 100-109.

Matsui, T. and Furui, S. (1991), "A Text-Independent Speaker Recognition Method Robust Against Utterance Variations," Proc. ICASSP-91, pp. 377-380.

Rudasi, L. and Zahorian, S. A. (1991), "Text-independent Talker Identification with Neural Networks," Proc. ICASSP-91, pp. 389-392.

Rudasi, L. and Zahorian, S. A. (1992), "Text-Independent Speaker Identification using Binary-pair Partitioned Neural Networks," Proc. IJCNN-92, pp. IV: 679-684.

Soong, F. K., et al. (1985), "A vector quantization approach to speaker recognition," Proc. ICASSP- 85, pp. 387-390.