REUSABLE BINARY-PAIRED PARTITIONED NEURAL NETWORKS FOR TEXT-INDEPENDENT SPEAKER IDENTIFICATION

Stephen A. Zahorian

Department of Electrical and Computer Engineering
Old Dominion University
Norfolk, VA 23529

ABSTRACT

A neural network algorithm for speaker identification with large groups of speakers is described. This technique is derived from a 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 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. 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 the TIMIT database, the number of pair-wise classifiers can be reduced by approximately a factor of 5, with only minor degradation in performance when 3 seconds or more of speech is used for identification. Using all 630 speakers from the TIMIT database, this method can be used to obtain over 99.7% accuracy. With the telephone version of the same database, an accuracy of 40.2% can be obtained.

1. INTRODUCTION

There are several well-established techniques for speaker recognition/identification (for example, see Gish and Schmidt for a tutorial article). Techniques include both parametric methods such as Gaussian Mixture Models (Reynolds and Rose, 1995) and nonparametric methods such as ones using vector quantization (Soong et al., 1985; Matsui and Furui, 1991) or ones which use neural networks (Bennani and Gallinari, 1991; Rudasi and Zahorian, 1991 and 1992). The neural network approach, which is used in this paper, although potentially 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 prohibitively long. 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. 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 a single large neural network is that it significantly reduces the training time and requires less 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 technique for reducing the number of pair-wise networks required by the BPP approach. This technique will also be referred to as reusable binary pair partitioning (RBPP).

The remainder of this paper provides an explanation for this method and summarizes some experiments with the TIMIT and NTIMT databases used to evaluate this technique.

2. REUSABLE BINARY-PAIRED PARTITIONING METHOD

The basis for expecting that pair-wise networks can be used to separate many speaker pairs stems from the observation that speakers are likely to be clustered in similar groups. If a binary network is trained to separate two speakers, with one speaker from each of two widely separated groups, that network is also quite likely to be effective in separating the other speakers between those two groups. For example, if a certain network is trained to separate a specific female speaker from a specific male speaker, it is quite likely that that network will also separate many other female/male speaker pairs.

To take advantage of this speaker clustering in a systematic way, we begin by arbitrarily selecting the first two speakers in our speaker population and then training 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 all 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 accounted for. In practice, as shown in the experimental section, this method can be used to greatly reduce the number of pair-wise networks that would have been required by the BPP approach.

In our implementation of this method, 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 for that speaker pair. This process may also completely eliminate 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.

3. 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 RBPP method significantly reduces the number of pair-wise networks with very little degradation in identification accuracy. For all experiments each pair-wise network was a memoryless, feed-forward, multi-layer perceptron and was configured to have one hidden layer of 5 nodes and one output node. Backpropagation was used for training these networks with 200,000 network updates using an initial learning rate of 0.25. The learning rate was reduced by a factor of .96 every 5000 network updates. A momentum term of .6 was used.

The TIMIT and NTIMIT speech databases were used for testing. These data bases each contain 10 sentences for each of 630 speakers 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. The NTIMIT data base contains the same data as the TIMIT data, except that all the speech materials were transmitted over phone lines.

In all experiments 20 cepstral coefficients (CC0 to CC20) were computed for each speech frame as follows. First, a second order high frequency pre-emphasis filter with a broad peak around 3 kHz was applied to the speech signal. The second step was to compute a 1024 point FFT from each 32 ms Kaiser-windowed (coefficient of 5.33) frame of speech data with the window advanced by 16 ms. The following step was to compute the amplitude spectrum, logarithmically scale it, and then frequency warp it with a bilinear function using a coefficient of .45. The next step was to compute the 20 cepstral coefficients as the cosine transform of the scaled magnitude spectrum over the frequency range 0 to 8000 Hz for TIMIT, and 300 Hz to 4000 Hz for NTIMIT.

Experiment I

This experiment was conducted to investigate tradeoffs between accuracy and number of networks needed as a function of a "threshold" parameter. Note that, in this application, since 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, it was very straightforward to define a threshold as some number between .5 and 1.0. In particular, the average neural network output level for all speech frames was compared to the threshold value to determine if a given network was suitable for discriminating between two speakers. The experiment was conducted using all 102 speakers of dialect region 2 with the threshold value changed from 0.55 to 0.75 in steps of 0.05. For comparison the experiment was also conducted with the original BPP method, that is using all 5151 networks. The numbers of networks used for each of the threshold values are shown in table 1a. Figure 1 shows the identification accuracy for these 102 speakers as a function of the amount of speech used for identification from each speaker and for different values of the threshold, and for the complete set of networks. The figure shows that when the threshold is higher than 0.70, there is no significant improvement in performance, and the performance is nearly equal to that of the BPP method.

This experiment was repeated using the same 102 speakers, but with the NTIMIT data base. For this case the threshold values were varied from .55 to .70. The numbers of networks used for each of the threshold values are shown in table 1b.

Threshold	.55	.60	.65	.70	.75	BPP
Networks	35	127	353	781	1686	5151

Table 1a. Number of networks required for a given threshold value for TIMIT data base experiments.

Threshold	.55	.60	.625	.65	.675	BPP
Networks	156	691	1197	1849	2480	5151

Table 1b. Number of networks required for a given threshold value for NTIMIT data base experiments.

The most dramatic difference between this case and the data from TIMIT is that the overall accuracy is severely degraded, as expected. For this case, performance of the RBPP method is also approximately equal to that of the complete BPP method, for thresholds of .675 or more. However, many more networks are needed to reach this threshold than for the case of TIMIT.

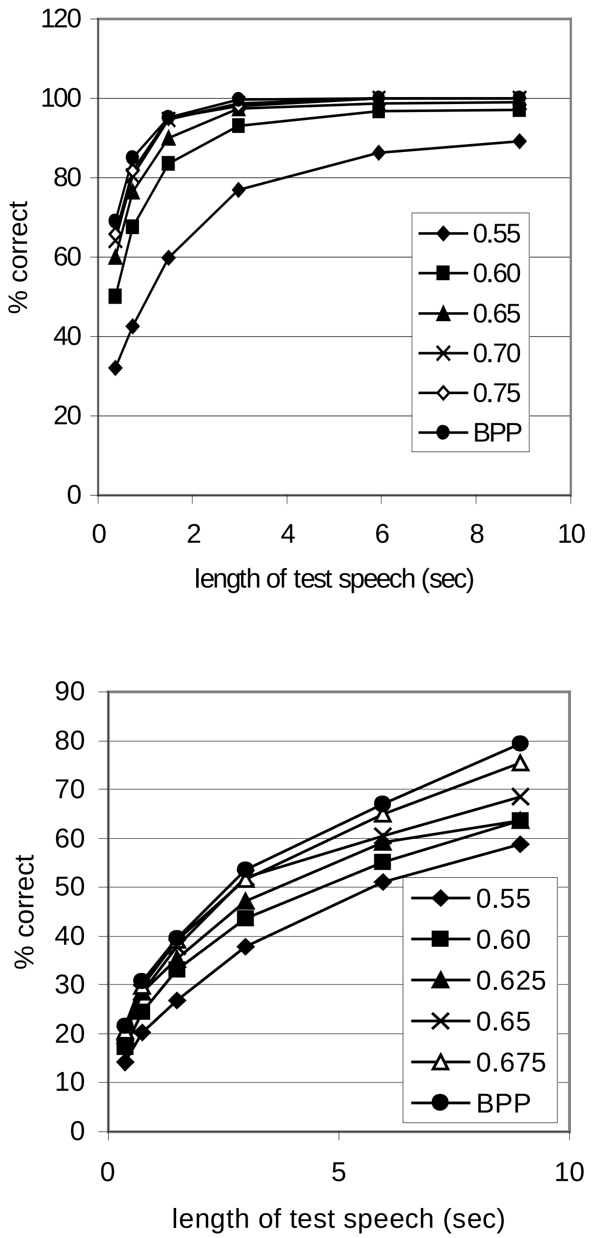

Figure 1. Speaker identification accuracy as function of test speech length for various thresholds for replacing networks. The upper panel is for TIMIT, and the lower panel is for NTIMIT.

Experiment II

The purpose of this experiment was to experimentally determine the reduction in the number of required pair-wise networks as the number of speakers increases. We varied the number of speakers from 25 to 400 speakers. The number of networks required by the BPP approach are, of course, determined solely by the number of speakers. Figure 2 shows the number of networks required by each of the two approaches as a function of the number of speakers. A threshold of .70 was used for the RBPP method for the TIMIT data, and a threshold of .625 for the case of NTIMIT. As the figure shows, for the thresholds used, the number of networks required for the RBPP is approximately 1/10 of the number needed for the BPP approach for each number of speakers. Thus, the computational load of the recognizer is also reduced by about a factor of 10.

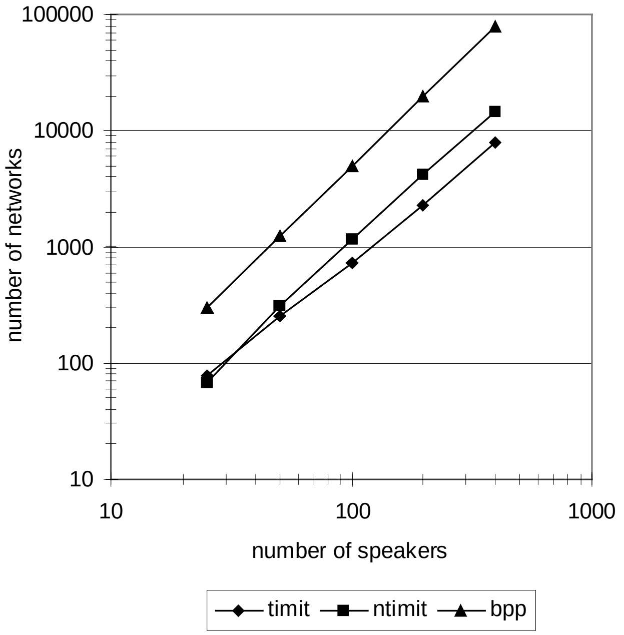

Figure 2. The number of networks required for the reusable binary paired partitioned classifier as a function of the number of speakers.

Experiment III

This experiment was conducted to evaluate the performance of the RBPP approach when applied to a large number of speakers. For this purpose, the RBPP system was trained for all 630 speakers for both the TIMIT and NTIMIT data bases. A threshold of .75 was used for the TIMIT data and .600 for the NTIMIT data base. A total of 43000 pair-wise networks were computed for TIMIT and 22,000 for NTIMIT as compared to 198,900 networks which would have been needed by the BPP approach. Figure 3 shows the performance achieved by the system for the 630 speakers as a function of the amount of speech used for evaluation. For each data base, two curves are drawn: one is the accuracy considering only if the correct choice, the other is the accuracy if the correct speaker is among the top 5 choices.

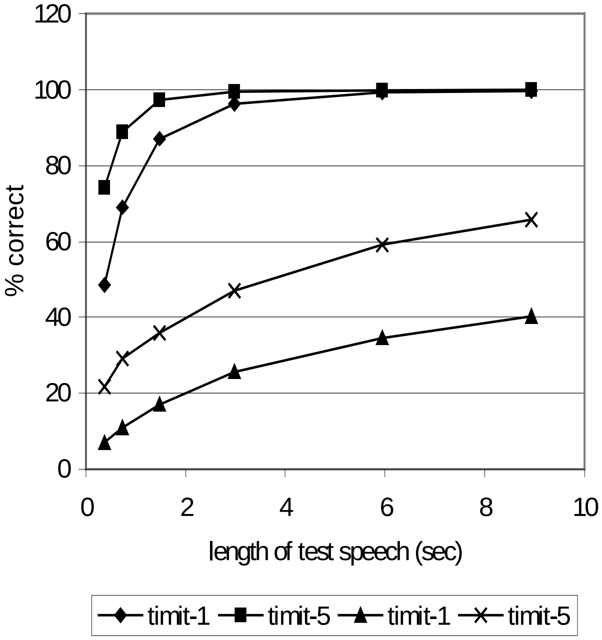

Figure 3. Performance achieved by the RBPP classifier for a 630 speaker identification task, as a function of the amount of speech used for evaluation.

For the 630 speakers of TIMIT, we obtained 96.4% accuracy when 3 seconds of speech (one sentence) was used. This accuracy improves to 100%, as the test length increases to 9 seconds and if a speaker is considered "correct" if among the top 5 choices of the classifier. For the case of NTIMIT, the corresponding accuracy rates are 25.6% and 65.7%. Compared to another study which reported speaker identification results for all 630 speakers of TIMIT (Reynolds et al., 1995), these results are very similar for the case of TIMIT, but lower for NTIMIT. It should be noted however, that there are several other differences between our work and this previous work, including the fact that the two sentences with the same speech materials for all speakers (the SA sentences) were used as training sentences in this previous study, thus not making the tests totally text independent. In our work, the SA sentences were used for testing only, thus not biasing the training process.

3. CONCLUSION

The RBPP technique described in this paper was shown 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 with this technique is nearly as good as the BPP approach, provided the number of networks is reduced by no more than a factor of approximately 10.

The method provides a very convenient mechanism for trading off accuracy versus computational demands and storage requirements. If lower accuracy can be tolerated, a lower threshold for replacing networks can be used to reduce the number of networks needed.

An apparent extension of the method would be to retain the reused networks, using all data from speakers in the two groups separated by the network. Presumably, this would further reduce the number of networks needed for a given level of performance. It could also allow better scaling properties to very large speaker populations.

4. ACKNOWLEDGEMENT

Portions of this work were supported by NSF grant IRI-9217436.

5. REFERENCES

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

[2] Gish, H., and Schmidt, M. (1994), "Text-Independent Speaker Identification," IEEE Signal Processing Magazine, October 1994, pp. 18-32.

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

[4] Reynolds, D. A., Zissman, M. A., Quatieri, T. F., O'Leary, G. C., Carlson, B. A. (1995), "The Effects of Telephone Transmission on Speaker Recognition Performance," ICASSP-95, pp 329-332.

[5] Reynolds, D. A., and Rose, D. A. (1995), "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," IEEE Trans. Speech, Audio Processing, 3, pp 72-83.

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

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