TEXT-INDEPENDENT TALKER IDENTIFICATION WITH NEURAL NETWORKS

Laszlo Rudasi and Stephen A. Zahorian

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

ABSTRACT

This paper introduces a binary partitioned approach to classification which is applied to talker identification using neural networks. Neural networks have been shown to work exceptionally well for small but difficult classification tasks. Their application to large (i.e. having more than 10 to 20 categories) however is limited by a dramatic increase in required training time. The time required to train a single network to perform N-way classification is nearly proportional to the exponential of N. In contrast, the binary partitioned approach requires training times on the order of N^2. Our experimental evidence also suggests that the binary partitioned neural network approach requires less training data then the use of a single large network. The binary partitioned approach was used to develop a talker identification system for the 47 male speakers belonging to the Northern dialect region of the TIMIT data base. The system performs with 100% accuracy in a text-independent mode when trained with about 9 to 14 seconds of speech and tested with 8 seconds of speech.

1. INTRODUCTION

Over the past several years many studies have been conducted in automatic speaker identification or verification. There are so many variations in the particulars of the problem definition and in the data bases used to test speaker identification algorithms that meaningful comparisons are difficult to make. For example many studies are conducted with "laboratory" quality speech while others are conducted with "telephone-quality" speech. The amount and type of speech materials used for training and testing varies considerably among the studies. However, in the majority of previous studies, the number of speakers has been restricted to a relatively small number such as ten or twenty. Many previously presented classification schemes for speaker identification do not scale up well to a large number of categories (speakers). In this paper, we present a method for partitioning a large classifier using a large number of small classifiers, and we give evidence to show that this method is particularly well-suited to speaker identification for large speaker groups.

There are several possible ways to partition a large classification task. Each involve applying subclassifiers to an unknown sample, each of which eliminates one or more incorrect categories. One approach is to use a tree-like network of classifiers, with each classifier sorting unknown samples into distinct groups. As an unknown sample works its way down the classification tree the number of categories per group is successively reduced until, at the bottom of the tree, only one category per group remains. For example, for the speaker identification problem, with speakers in the "natural" groups of men, women, and children, classification could be accomplished with four classifiers - one to sort men, women and children and one classifier each to determine the particular speaker within each group [2]. The classification problem could also be partitioned with a tree structure of binary classifiers. Each classifier would sort incoming data into one of two groups. At each step of classification half the remaining possible categories are eliminated. With this approach the number of required subclassifiers for an N category problem is N-1, and the number required for decision making is the order of LOG_2N. Assuming that there are no errors made at the beginning levels of the binary decision tree, no subclassifier would need to make a decision about an input sample whose category was not represented in its training set. Such an approach is not limited to two-way separations of data at each step. However this method is limited by the need to find "good" partitionings of the categories for each classifier.

The approach presented in this paper is another alternative to partitioning the classification problem. In particular we propose using a large number of binary classifiers, with each classifier trained to distinguish only between two categories. With N categories, there are a total of N*(N-1)/2 pairs of categories. Therefore, a total of N*(N-1)/2 binary classifiers are required, each trained to discriminate between one specific pair of categories. If each of these pairs can be successfully discriminated, the overall classification problem can be solved. An unknown sample could be classified with the pair-wise classifiers using a total of N-1 binary decisions. That is, since each binary decision eliminates one category from contention, only one category remains after the N-1 decisions.

The major advantage of this approach is that the categories need not be grouped. There is no need to find "natural" partitions. Rather each classifier can be highly tuned to discriminate between the two members of its particular pair. A potential disadvantage is the requirement for a large number of classifiers. For example with N=47, 1081 binary-pair classifiers are required versus only 46 binary-group classifiers. There is also no guarantee that a system based on binary-pair classifiers will perform well in a real classification problem. However, in this paper we will show that the binary-pair classifier not only performs well for speaker identification but also has advantages over a single large classifier.

2. NEURAL NETWORKS FOR TALKER IDENTIFICATION

2.1. Single large network

Neural networks have been applied to many classification tasks with great success. In several previous studies, and as a control for the present study, a two layer, feedforward, fully interconnected, memoryless neural network,

CH2977-7/91/0000-0389 $1.00 © 1991 IEEE

trained with backpropagation was used for speaker identification. The main variables with this method are the number of hidden nodes, the amount of training data, the amount of training iterations, the learning rate, and the test speech length. Variables which are of lesser importance, or which have more or less been optimized in many previous studies, are the number of layers, the type of non-linearity function used (sigmoid), the method of initializing the weights, and the momentum term [1]. The output nodes of the network each correspond to one of the categories (speakers). For each input pattern, the network is trained to have the output corresponding to the correct category high, while keeping the other outputs low. During classification of an unknown sample, the output nodes are accumulated over the number of frames in the sample, and the category with the greatest sum is chosen. Since the neural net is memoryless, the averaging is external to it. Thus the network makes many independent soft decisions each of which is based on a small segment of speech. This approach contrasts considerably with first computing speech statistics, such as covariance matrices, etc. and then classifying based on long term average properties [3,4].

The memoryless feed-forward architecture makes decisions based only on static features, since each frame is considered independent of its neighbors. By increasing the time window, or by adding short term memory to the network, such as with a recurrent or time-delay neural network, the ability to utilize dynamic information is added. This causes some performance improvement [2], but at the expense of significantly increased training time. In the present study, only memoryless feed-forward networks were used.

The main problem with the use of one large network for talker identification is that training time increases exponentially with the number of categories. Thus large problems become unsolvable. Furthermore, the addition of a new talker to an existing system generally requires retraining the entire network. These problems are solved by the use of binary-pair neural networks. This new approach also offers the advantage of modularity, which could be useful in implementation.

2.2. Binary-pair neural networks

The solution proposed in this paper is to replace one large network with a large number of much smaller networks. The binary-pair approach is not restricted to neural nets, but seems particularly appropriate for the case of neural nets. As discussed earlier, N*(N-1)/2 small classifiers are trained, each to distinguish between two of the N categories. Each of these small binary (two-way) neural nets are independent of the others as well as the training data of the non-relevant categories.

There are two fundamentally different approaches for using these binary classifiers to classify an unknown sample. They both classify with 100% accuracy as long as each of the binary classifiers performs correctly. The performance of the two approaches, however may degrade differently if not all the binary classifiers work perfectly. The first, and simplest approach to classifying an unknown sample is to run the sample through all the binary classifiers and tallying the "votes" for each of the categories. If all the binary classifiers work properly then only the correct category will receive a perfect score, while the best possible competing score is one less. A slight variation of this method is to recognize that the output of each of the binary neural net classifiers is a soft decision. The method of summing soft decisions will be referred to as the "global soft decision search" in the experimental section to follow.

The other basic approach may be thought of as a series of elimination rounds. First, with N categories all the categories are paired in N/2 pairs, and one of each pair is eliminated by the application of the binary classifier corresponding to the pair. The winners of each round are then paired again in the next round until only one category survives. (If N is odd, the one "left-out" category automatically advances to the next round.) One potential advantage of this method, if implemented on a serial computer, is that only N-1 binary classifications are required. The number of these that must make the correct decision is on the order of $\log_2 N$. This classification method will be referred to as the "binary tree search" in the experimental section.

3. EXPERIMENTS

3.1. Database

The speaker identification experiments were performed on a subset of the DARPA TIMIT Acoustic Phonetic Continuous Speech Database. The database consists of 10 sentences from each of 420 talkers as follows: two dialect calibration (SA) sentences, three random contextual variant (SI) sentences, and five phonetically compact (SX) sentences. The two SA sentences are the same for each of the talkers. The SI and SX sentences vary from talker to talker. This speech data was sampled at 16 kHz with a 16 bit A-D.

For these experiments the 47 male talkers belonging to the Northern dialect region were used. The SA and SI sentences were used for training and the SX sentences for testing. Thus it should be noted that of the 5 training sentences, only the SA sentences were the same for all the speakers. Presumably less training data would have been required if all the training sentences were the same for all speakers. All the testing sentences were different for each speaker. In experiments with less than 47 categories, the talkers were chosen based on the alphabetic ordering of the speaker initials.

3.2. Preprocessing of the acoustic data

The feature set used for encoding the speech signal was a form of cepstral coefficients computed as follows. First, the speech signal was high-frequency preemphasized with transfer function $1-.95z^{-1}$. The speech signal was then windowed using a 32 msec Hamming window, with a 10 ms frame spacing. The magnitude spectrum was computed using a 512 point FFT for each frame. The spectrum was then log amplitude scaled and frequency warped with a bilinear transform with a coefficient of .6. Fifteen cepstral coefficients were then computed over a frequency range of 150 Hz to 6000 Hz and used as features for additional processing. Using statistics computed from the entire training set, the cepstral coefficients were normalized for zero mean and a standard deviation of one for each coefficient.

Low energy frames were discarded as input to the speaker identification system. Low energy frames were defined as those with a normalized zero-order cepstral coefficient of less than -1. The threshold was determined from pilot experiments which indicated that low energy frames were poor predictors of talker identity. This threshold eliminated about 20 percent of the frames.

3.3. Experimental design and results

The experiments were designed with the following four goals in mind:

(1) to compare the performance of the binary-pair partitioned approach to that obtainable with one large network;

(2) to compare the relative training time requirements of the two methods;

(3) to compare the soft global search and the binary tree search methods of classifying using binary neural net classifiers;

(4) to examine the performance degradation of the binary partitioned approach as more categories are added.

In each experiment, the network or networks used were two layer, fully interconnected, memoryless, feed-forward, with sigmoid non-linearities. Network weights were initialized with random values uniformly distributed from -0.05 to 0.05. Each was trained with the backprop method with a fixed learning rate (0.1 - 0.3) and fixed momentum term (0.7). The output targets were 0.999 for the node corresponding to the correct category, and 0.001 for the other(s). The training and testing data sets were the same in each experiment as discussed in sections 3.1. & 3.2.

For experiments 1, 3, and 4, networks were trained until an empirically determined threshold of performance was reached on the training data. Typically this threshold was 65% to 75% of training data frames correctly classified, for the binary partitioned approach, and 30% to 50% for the single large network. Note that some binary networks were thus trained for more iterations than others. The lower threshold was used for the large network, because the errors were distributed over several nodes rather than just one as for the binary-pair networks. The exact threshold used depended on the number of hidden nodes and also, for the case of the single large network, on N.

3.3.1. Experiment 1.

This experiment was designed to compare the performance of the partitioned neural-network method to that attainable with a single large network. Three cases were considered: N=5, N=10, and N=15. Each of the six systems was optimized iteratively with respect to the following variables: number of hidden nodes, learning rate, and amount of training. For each of the six systems, the result of the best run is shown in Figure 1.

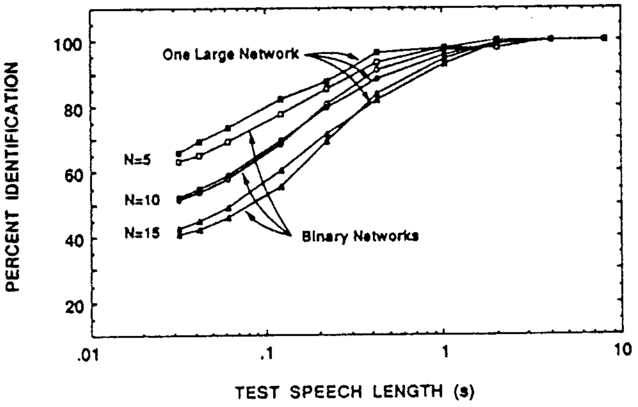

Figure 1. Speaker identification performance comparison of one large network versus a system of binary-pair networks.

The basic result is that performance is almost the same for the two classification schemes. For short duration segments, and for N=5, the single network classifier is slightly better. But more importantly, for N=10 and N=15, for longer speech segments (which are needed for reliability, and are therefore of more interest) the partitioned system performs slightly better. The fact that the single network degrades more with the increase in N tends to suggest that it needs more training data per talker than the partitioned system as N increases.

It is worth noting that the single network benefited more from the optimization process. The partitioned system is less sensitive to changes in the variables, which further suggests that the small binary nets are less sensitive to overtraining and need less training data. However, for both cases, the results were not overly sensitive to the network parameters. For example, as the number of hidden nodes changed by a factor of two from the "optimum" value, performance degraded only slightly. The numbers of hidden nodes used for the data plotted in Figure 1 were 10, 20, and 45 for the single networks with N = 5, 10, and 15 repsectively, and 6 for the case of all the binary pair networks.

3.3.2. Experiment 2.

This experiment was designed to compare the training time of the single network to that of the binary partitioned system as a function of N. Both systems were required to perform the same task, which was to correctly identify N speakers, each represented by a single 4 second segment of test speech. Each system was tested frequently during training to determine the time at which the 100% threshold was reached. The results, shown in Figure 2. reflect only the training time. The figure clearly shows that training time increases much more rapidly with N for the large network versus the binary-partitioned system of networks.

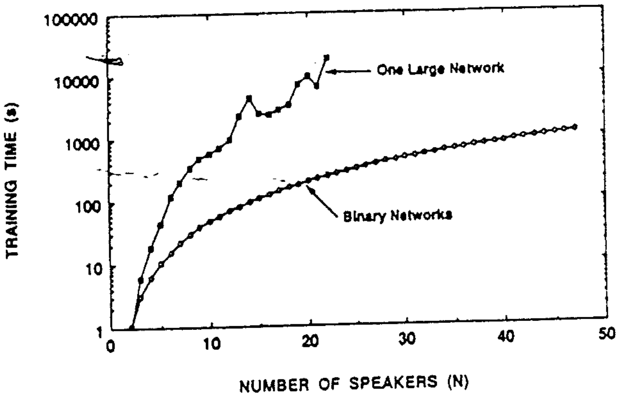

Figure 2. Training time comparison of one large network versus a system of binary-pair networks.

Each of the large single networks had 2*N hidden nodes, which was found optimal for performance in Experiment 1. (In a few side experiments with N=10, 20 hidden nodes were found close to optimal with respect to training time as well.) Two runs were made for each value of N, with learning rates of 0.1 and 0.2. The two results were then averaged.

The training time requirement of the binary partitioned approach was averaged from a much larger number of runs. 190 binary classifiers were trained, (corresponding to N=20). They each trained to the required performance threshold in 0.47 to 2.35 seconds, with an overall average of 1.05 seconds. The expected value of the training time for a system with any N was, therefore determined to be 1.05*N*(N-1)/2.

The computer used for these experiments is based on a 16 kHz Intel 80386 processor with an 80387 coprocessor. The programs were written in Microsoft Fortran V5.0 running under OS/2.

3.3.3. Experiment 3.

This experiment compares the two methods of classifying using binary pair networks. These methods are the global soft decision and the binary tree search as discussed in section 2.2. Both methods used the same set of binary classifiers. First, the binary classifiers were each trained to get a recognition rate of at least 75% on their respective training data sets on individual speech frames. Figure 3. shows that the binary tree search consistently performs somewhat better than the global soft search method.

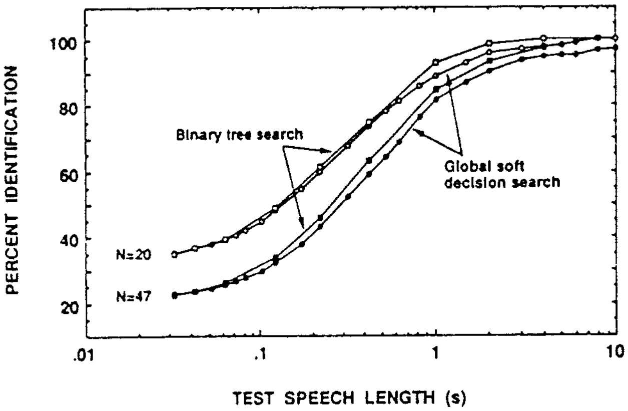

Figure 3. Speaker identification performance comparison of binary-pair neural network system for two evaluation methods.

3.3.4. Experiment 4.

This experiment shows recognition results for the binary partitioned system as a function of test speech length for N = 5, 10, 20, and 47. The results, shown in Figure 4. were obtained with the binary tree search method, and "optimum" values for hidden nodes, training thresholds, etc. from previous experiments. As expected the results degrade somewhat as N increases, in the sense that a larger length of test speech is required to reach 100% performance. However, even for 47 speakers, 100% performance is reached with 8 seconds of speech data.

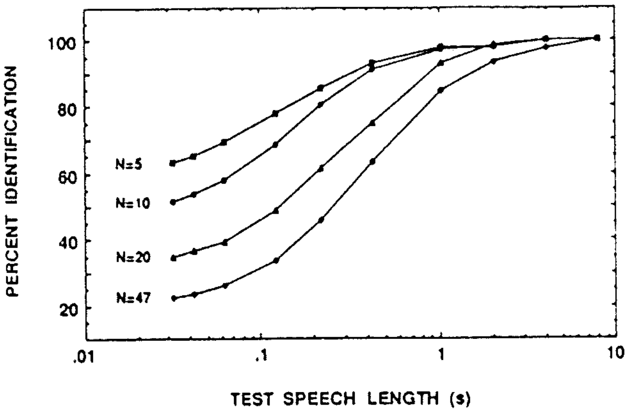

Figure 4. Speaker identification performance for binary-pair neural networks for varying numbers of speakers.

4. CONCLUSIONS

In this paper we have introduced a new method for partitioning a large classification problem using N*(N-1)/2 binary pair classifiers. The binary pair classifier has been applied to a talker identification problem using neural networks for the binary classifiers. This partitioned approach performs comparably, or even better, than a single large neural network. For large values of N (> 10), the partitioned approach requires only a fraction of the training time required for a single large network. For N = 47, the training time for the partitioned network would be about two orders of magnitude less than for the single large network. The talker identification results obtained in this study appear to be better than any previously published results for tasks of similar complexity.

REFERENCES

[1] J. Oglesby and J. S. Mason (1990), "Optimization of neural models for speaker identification," ICASSP-90, pp. 261-264.

[2] L. Rudasi and S. A. Zahorian (1990), "Text-independent talker identification using recurrent neural networks," J. of Acoust. Soc. Amer. 87, Suppl, S104.

[3] M. Savic and S. K. Gupta (1990), "Variable parameter speaker verification system based on hidden markov modeling," ICASSP-90, pp. 281-284.

[4] W. Ren-hua, H. Lin-shen and H. Fujisaki (1990), "A weighted distance measure based on the fine structure of feature space: application to speaker recognition," ICASSP-90, pp. 273-276.