THE APPLICATION OF BINARY-PAIR PARTITIONED NEURAL NETWORKS TO THE SPEAKER VERIFICATION TASK

Claude A. Norton III, Stephen A. Zahorian and Zaki B. Nossair

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

KEYWORD LIST

List 1: Pattern Recognition

List 2: Binary-pair partitioned neural networks; Speaker verification; Cepstrum analysis: Speech signals; Decision making; Decision analysis; Kaiser window; IDFT; Learning rate; Learning parameters; Network architecture; Neural network algorithms; Neural network applications; Neural network classifiers; Pattern classification; Pattern categorization; Performance evaluation.

ABSTRACT

A method is presented for the application of binary-pair partitioned neural networks in the task of speaker verification. The binary-pair partitioned neural network is a previously developed technique used for speaker identification [1]. The training and evaluation procedures are discussed, as well as the selection of the verification thresholds. For a verification task of 30 users and 41 impostors an accuracy of 96.3 percent was achieved using 13.5 seconds of input speech extracted from the DARPA/TIMIT database [2]. For input speech lengths as low as 2.7 seconds the system maintains a 86.9 percent accuracy.

INTRODUCTION

Speaker verification which is accurate, fast and reliable is an important task with many possible applications. A good speaker verification system should not sacrifice reliability and accuracy for speed; but, should work toward the goal of consistently good verification with a minimum of input speech. To achieve such performance a speaker verification system must have a solid foundation built upon past work with speaker identification. The approach in this paper is based on a robust technique for speaker identification which has been presented by Rudasi and Zahorian [1].

This paper will describe a technique for the extraction of acoustic information from a speech signal and the representation of that information in a manner suitable for classification with a binary-pair partitioned neural network (BPP). We will also discuss the development and determination of a thresholding technique which partitions the speaker space into two categories: users and impostors.

BACKGROUND

Speaker verification is the logical extension of speaker identification. In the identification problem we are faced with a fixed number of possible speakers, M. The identification task is to identify the present speaker as one of the M possible speakers. The verification task includes not only this simple classification, but the system must also be able to reject an infinite set of 'impostor' speakers. Thus, the speaker space consists of a set of acceptable speakers, M (users), and an infinite set of possible speakers, N, including N-M impostors. The verification task includes not only correctly identifying users, but also rejecting impostors. The speaker space is best represented by a simple Venn diagram (figure 1). With the added degree of complexity inherit in speaker verification comes practical applications, such as: keyless entry systems, voice verification in conjunction with personal identification number for secure transactions (i.e., credit cards, banks, etc.) and the use of voice verification as a security feature.

CLASSIFICATION TECHNIQUE

The classification method used in this research is called binary-pair partitioning (BPP). Binary-pair partitioning is a special case of group partitioning. It uses M*(M-1)/2 two-way classifiers to make an M-way decision. Each binary decision is made between a pair of categories or speakers. Thus, there are M-1 decisions relevant to each user in set M. For classification these decisions are combined to produce an overall decision The purpose for this type of partitioning is improve accuracy and reduce training time. Using a conventional single network for an M-way decision requires training time exponentially proportional to the number of users (M). Using the BPP network can reduce the training time to as little as $LOG_2(M)$. Implementation of the BPP network requires two steps in the classification process:

1. "Elemental" classifiers are trained to distinguish between every possible pair of speakers in set M (users).
2. Test data is then run through each elemental classifier trained in step 1. The decisions made in this process are combined to produce an overall decision based on all the binary-pairs.

The performance of the BPP network classifier has been established in several other studies [1,3,4].

For the task of speaker verification the BPP neural network was adapted for use with the addition of two threshold factors. The first is an overall performance threshold, called the primary threshold (PT). The second is a threshold based on a quality measurement from each of the binary networks, called the secondary threshold (ST). The PT is implemented as follows. For an unknown speaker, the BPP network computes a 'distance' measure to each possible user. This step essentially computes a pseudo-probability that the unknown speaker is each of the users. The results are scaled in a range of 0 - 1. The higher value (1) would indicate a high probability of the unknown speaker being the current user. A low value (0) would indicate the opposite. Thus, the user with the maximum value for these comparisons would be recognized as the correct speaker, subject to the PT. If the maximum value is above the PT then the unknown speaker is considered a user subject to the ST. If the maximum value is below the PT then the unknown speaker is determined to be an impostor.

The ST is implemented using a measure of the quality of each of the binary networks which is computed during the training phase. That measurement is represented by a vector in which each element is determined by the absolute difference between the desired output and the actual output of each network. This absolute difference is computed on a frame by frame basis and then averaged over the number of frames used in training the network for each user. This average value becomes an element in the vector which represents the measure of quality. The use of this threshold is based upon the assumption that the relative performance of each of the binary classifiers remains consistent between training and test data for the same user. By relative performance we mean the performance of each binary network with respect to the others. During the training phase the quality vector of each binary network is recorded. This forms a vector of length M-1 (recall that M is the number of users). During the testing phase an unknown segment of speech is classified as the closest possible user. Once the closest user is determined and the normalized sum of the network outputs exceeds the PT the measure of quality is computed for the unknown segment of speech. A distortion measure, using Euclidean distance, is made against the quality measure determined during training. If the distortion measure is lower than the ST value, then that segment of speech is identified as belonging to the proposed user. If the distortion measurement is larger than the threshold then that segment of speech is classified as belonging to an impostor. The use of each threshold as a solitary threshold was examined and it was determined that consistently higher performance is achieved using both thresholds.

It must be noted that our method of classification, that is predetermining the closest user first, is far more stringent than other approaches currently in use [5]. In alternate methods, each unknown speaker first claims an identity as each of the possible users and then that claim is either substantiated or rejected. Thus for a case with M users and 1 impostor, M*(M-1) different classifications are made. Using our approach the same case would yield only one classification: the impostor would first be classified as the closest user, then a determination would be made as to the validity of that claim. As a result of these differences, our method yields more modest results. The threshold, which is empirically derived, is the determining factor for effective performance. It is a probabilistic border between the user and the impostors. Thus, the threshold can be manipulated to either enhance recognition or rejection based on the particular application. This flexibility is extremely useful in applications which require stringent security, or applications which require only minimal security but where rejection of a valid speaker would be detrimental.

FEATURES

The features, or characteristics, used for recognition are based on acoustic information contained in the speech signal. The features used, Discrete Cosine Transform Coefficients (DCTC), compactly represent the magnitude frequency components of speech as a series of coefficients. A 30 term expansion was used to represent each frame of speech with one term representing the pitch. Forty millisecond frames were selected at 20 millisecond intervals. Thus, each frame of speech included a 20 millisecond overlap. A Kaiser window was used for smoothing (coefficient of 5.33) [3]. The DCTCs were computed over a frequency range of 0 Hz to 8000 Hz. The data was then scaled, linearly, using the following relationship:

$$X_{\_i} = \frac{X_i - \overline{X}}{5 \bullet \sigma_x} 0$$

Where X_i' is the scaled value and equal to the original coefficient minus the original mean, divided by five standard deviations. The purpose for this scaling of the feature set was to normalize the range of the data. This type of scaling results in faster network training and in some cases better overall performance. This combination of features and scaling has proved most effective in previous work in speaker identification [1,4].

EXPERIMENTS

The database used in all experiments was the DARPA/TIMIT Speech Corpus [2]. The speech corpus is composed of 630 speakers, both male and female, from different dialect regions in the United States. The dialect region variations are logically grouped into eight categories. For each speaker 10 sentences are recorded the first two sentences are identical for all speakers, the remaining eight are different. The experiments were performed using speakers of common dialect origins (dialect region 2) and the same gender. These guidelines were chosen to create the most challenging classification task.

The experiments were conducted with training data composed of 5 sentences from 30 male speakers. The training set was composed of the 5 SX sentences for each speaker. The testing data was composed of the remain 5 sentences (SA and SI). The content of the SX (training) sentences vary with each speaker. This training set constitutes set M, or the users. Several network parameters were experimentally varied and optimum values determined. The parameters which were varied include: the network learning rate, number of training iterations, number of features used in training and the number of hidden nodes in the network.

Three criteria were established for evaluation of performance: rejection rate, recognition rate and performance index. Included in the first two criteria are Type I and Type II statistical errors [6]. Briefly, a Type I error is incorrectly rejecting a valid user, while a Type II error is incorrectly recognizing an impostor (see table 1). Recognition rate is defined as the percentage of correctly identified users to the total number of users. Rejection rate is defined as the percentage of correctly rejected impostors to the total number of possible impostors.

A factor was established and named the performance index (PI), which represents the mean of the recognition and rejection rates as defined above. The PI is a more reasonable method of tracking both recognition and rejection rates as a function of other parameters such as: length of input speech, number of hidden nodes in the network, number of training iterations and threshold. The experiments were conducted in two phases: basic optimization and final experimentation. During the basic optimization stage a small set of speakers was used: 10 users and 10 impostors. This small data set was used to facilitate rapid turn-around time for modification of network training parameters. Several tests were initially performed with this small data set to determine general optimization parameters, such as: thresholds, network learning rate, training iterations, number of features and nodes in the hidden layer. The value of primary threshold was varied from .60 to .77 and the PI was computed as a function of the threshold, and it was determined that the best primary threshold was at .72. The secondary threshold was varied from 1.44 to 1.50 for a fixed primary threshold and the PI was monitored as a function of length of input speech (see figure 2). This figure displays the three best ST values as a function of input speech length. From the graph we can see a secondary threshold of 1.44 is superior for shorter segments of speech; while, a value of 1.50 is suitable for longer lengths of speech. The performance of the system improved by increasing the number of features used (up to 30). The training rate is a value which controls the initial magnitude of the weight adjustments during the neural network training phase. Values of .15, .25 and .35 were experimented with and the best performance for training was achieved at .25. The number of training iterations was varied from 150,000 to 600,000 by factors of 2. Good performance was achieved at 150,000 iterations with only a slight performance increase at 300,000 and 600,000 iterations. The slight gain in performance was not worth the additional training time required. The number of hidden nodes in the network was varied from 10 to 30. More hidden nodes allows for a complex decision to be established in the feature space; however, this is at the cost of network generalization. In the experiments a hidden layer of 30 nodes was found to be somewhat superior and thus was chosen.

RESULTS

The results achieved following the techniques presented were very promising. Based on 71 speakers, all male and of common dialect origins, the performance index was as high as 96.3 percent with 13.5 seconds of input speech (see figure 3). It can also be seen from this figure the system performs exceptionally well with as little as 2.7 seconds of input speech, degrading only to a performance index of 86.8 percent.

CONCLUSION

A technique for speaker verification based on the use of binary-pair partitioned neural networks has been described and evaluated. The subsequent results for this system are very encouraging. Experimentation is continuing with the goal to minimize the amount of input speech required to maintain good system performance. The speaker database will be expanded to include all 630 speakers available in the DARPA/TIMIT database.

ACKNOWLEDGMENTS

This research was supported by National Science Foundation grant # IRI9217431.

REFERENCES

[1] L. Rudasi, S.A. Zahorian, "Text-Independent Speaker Identification using Binary-pair Neural Networks," *IJCNN-92*, pp. IV: 679-684.

[2] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM," NTIS order number PB91-100354.

[3] S.A. Zahorian, Z.B. Nossair, C.A. Norton, III, “A Partitioned Neural Network Approach For Vowel Classification Using Smoothed Time/Frequency Features,” *EUROSPEECH-93*, pp. II: 1225-1228.

[4] L. Rudasi, S.A. Zahorian, “Text-Independent Talker Identification with Neural Networks,” *ICASSP-91*, pp. 389-392

[5] C. Montacié, J. Le Floch, “Discriminant AR-Vector Models for Free-Text Speaker Verification,” *EUROSPEECH-93*, pp. I:161-164.

[6] J. Freund, G. Simon, *Modern Elementary Statistics*, eighth edition, Prentice-Hall, 1992.

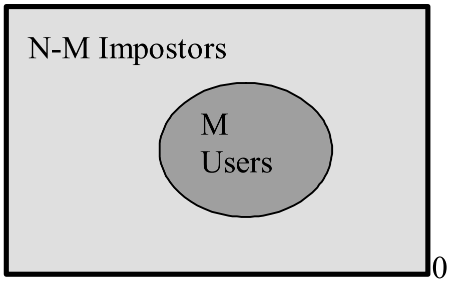


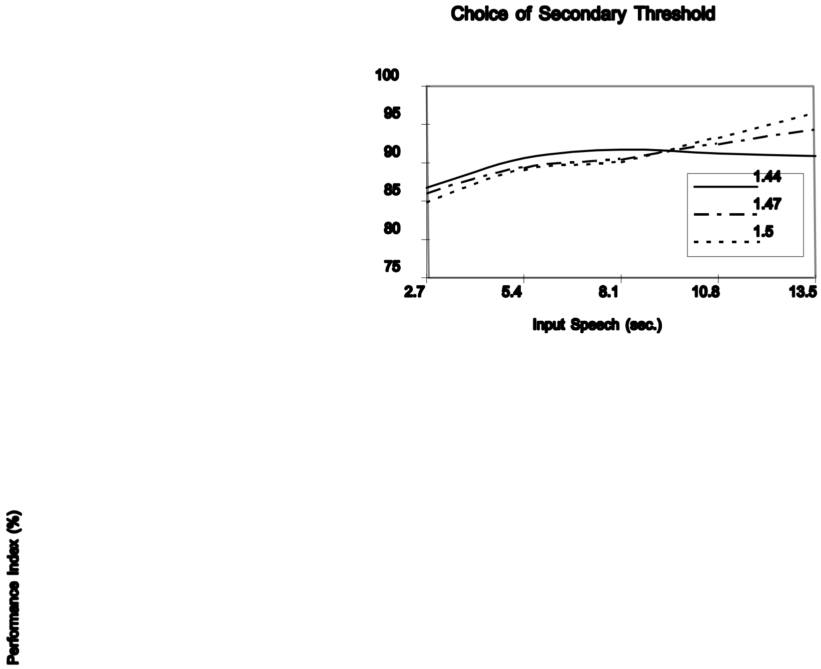
Figure 1: Venn diagram showing speaker universe N, acceptable speakers, M, and impostor speakers, N-M.

	Accept H_0	**Reject H_0**
H_0 is true	Recognition	Type I Error
H_0 is false	Type II Error	Rejection

Figure 3: Venn diagram showing speaker universe N, acceptable speakers, M, and impostor speakers, N-M.

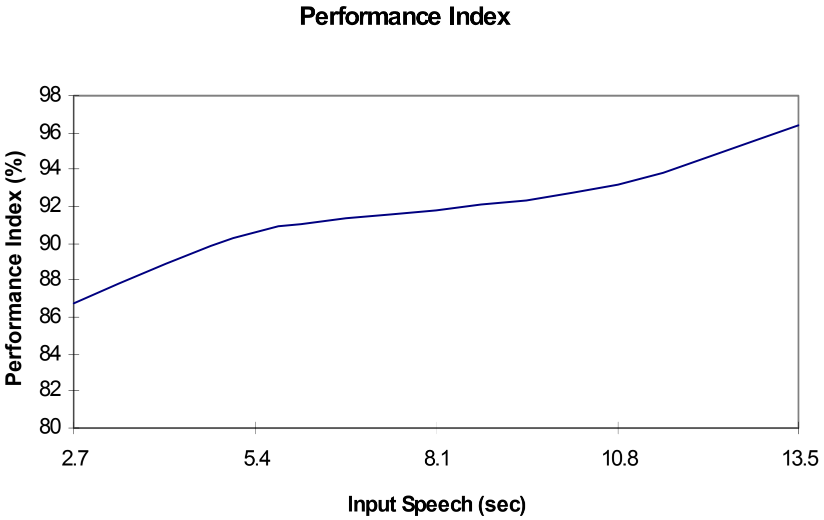
	Accept H_0	**Reject H_0**
H_0 is true	Recognition	Type I Error
H_0 is false	Type II Error	Rejection

Table 4: Type I errors, Type II errors, rejection and recognition criteria for any given hypothesis H_o.



0

Figure 5: Performance index as a function of input speech length for various secondary threshold values.



0

Figure 6: Results of verification experiments using 71 male speakers from a single dialect region, using 'optimal' parameter values.