SPEAKER VERIFICATION BASED ON SPEAKER POSITION IN A MULTIDIMENSIONAL SPEAKER IDENTIFICATION SPACE

CLAUDE A. NORTON, III AND STEPHEN A. ZAHORIAN
Department of Electrical and Computer Engineering
Old Dominion University, Norfolk, Virginia 23529

ABSTRACT:
A neural network technique is presented for the task of speaker verification. Using this technique each speaker is represented by a statistical profile derived from a vector of relative ‘distances’ to a set of other speakers. This profile uniquely represents each speaker in a multidimensional speaker space. The technique first uses binary-pair partitioned neural networks to discriminate each speaker from all other speakers in the space. This collection of networks computes pairwise distance measures among all users. The pairwise distance measures for each user, relative to all the other users, are combined to produce a profile by which that user is represented. A secondary neural network is then used to distinguish each user from all other speakers (impostors), based on these profiles. With this method, unknown speech can be accurately classified as user or impostor speech. Experimental results using the DARPA/TIMIT data base are presented.

INTRODUCTION

This paper describes a technique for speaker verification which characterizes each user by a distance measure relative to all other users in the speaker space. The approach used is based on the following steps: extraction of spectral features; training of initial neural networks for each pair of users; extraction of ‘profiles’ for each user; extraction of features from profiles, training of a secondary neural-network to identify each user; and classification of unknown speech as either ‘user’ or ‘impostor.’

BACKGROUND

Speaker verification is the logical extension of speaker identification. This leads to an infinite speaker space in which impostors far out-number the users. In contrast, for speaker identification, users form a small closed set. This difference leads to greater complexity in the verification process. In most applications, speaker verification is more useful than identification, thus leading to greater research interest in the verification task.

In a previous paper, we presented a robust speaker identification system based on a system of binary-pair partitioned neural networks (BPP) (Rudasi and Zahorian, 1992). Binary-pair partitioning is an alternative to group partitioning, using M*(M-1)/2 two-way, elemental, classifiers to make an M-way decision. Binary-pair partitioning has several advantages for classification tasks over a single large network, including better performance and reduced training time.

Since each of the BPP networks is trained discriminatively, the BPP method is easily applied to the identification task. In previous work, (Norton et. al., 1994), we presented a method to adapt the BPP method for verification. In that work we used a simple thresholding procedure at the outputs of the binary networks relevant to each user to distinguish that user from impostors. The method was based on the premise that the average levels of the outputs of networks trained to discriminate a user from other users would be higher for that user than for any impostor. However, as the number of users increased this method proved unreliable. The present paper gives a more robust method for discriminating users and impostors, but still using the basic discriminative BPP 'front end.'

The DARPA/TIMIT speech corpus was used in all experiments (Garofolo, et. al., 1991). The corpus is composed of laboratory-quality speech from 630 speakers, both male and female, from eight dialect regions in the United States. For each speaker 10 sentences were recorded, only two of which contain identical text across the speaker set.

METHODOLOGY

The following section highlights the basic approach used to characterize each user in terms of a unique "position" in the user-space. An initial BPP network was first trained using spectral features to distinguish between every possible pair of users. These networks were then used to generate profiles for each user, as described below. The user profiles, and similarly obtained impostor profiles, were then used to train a secondary neural network (or networks) to discriminate impostors and users. Impostor training profiles were obtained from each user posing as each of the other users. For the case of 20 users, there would thus be 19 training impostors for each user.

Primary Feature Selection

The initial BPP neural networks were trained using spectral features composed of 29-term DCT expansions of the log magnitude spectrum. These expansions were calculated from 32ms frames of speech spaced at 20ms intervals, each smoothed using a Kaiser window with â = 5.33 (Zahorian, et. al., 1993). The DCTCs were computed over a frequency range of 0-8000 Hz, then linearly scaled using a technique explained in (Rudasi and Zahorian, 1991). This type of scaling results in faster network training as shown in previous speaker identification work (Rudasi and Zahorian, 1991; Rudasi and Zahorian, 1992).

Primary Neural Network

Throughout the experimental phase, the primary BPP network configuration and topology remained constant. This network configuration has been effectively used in other similar research (Norton et. al., 1994; Rudasi and Zahorian, 1992; Rudasi and Zahorian, 1991). These networks, with hidden layer with 10 nodes and uni-polar sigmoid activation functions, were trained with back-propagation for 150,000 iterations (network updates, each based on the presentation of a single input vector). Training with more iterations did not significantly improve the results.

As mentioned above, the primary networks were used for pairwise comparisons of all users. For M users, there are M-1 networks which contribute to the recognition of each user. For example, for the case of twenty users, the BPP generates nineteen networks relevant to each user. Each of these networks is trained to perform one two-way discrimination (e.g., does the token belong to user one or two, one or three, one or four, and so on). The activation levels of each of these pertinent networks, as obtained from a representative speech sample, were used to create a profile for each user. In general, these levels would be higher for a sample of speech from that user versus a sample from an impostor. Additionally, however, some components of a profile would be expected to be higher than other components, depending on the relative similar of users. Thus these profiles create a 'signature' for each user. In effect, the 'signature' contains the 'distance' of each user to all other users in the user space. As described below, various methods were used to extract features from these profiles and to separate user and impostor profiles.

SECONDARY PROCESSING

Three distinct methods were used to separate users and impostors with the profiles mentioned above. In the first method, user profiles were represented by the mean, first central moment, second central moment and the third central moment. A similar set of 'features' were computed for a 'training' set of impostors. The basic premise was that the user moments would be distinctive from similar moments for impostors. These features were then used with a single secondary neural network to discriminate all users from all impostors.

For the second approach, the user profiles mentioned above were directly used. Thus each user was represented in more detail than for the first method. In addition, secondary neural networks were trained to discriminate the user profiles and impostor profiles for each user. For this method, using 20 users, 20 secondary neural networks were developed, in contrast to one secondary network for the first method.

In the final approach, statistics were again computed based upon the user profiles, as for the first method. Similar statistics (first and higher order central moments) were computed for profiles of temporal variance of pertinent network outputs across all frames of the speech sample. The two sets of statistics were combined to form a representation of each speaker. Thus, a typical feature vector has a constant size regardless of the number of users. In effect, statistics of statistics were used to represent each speaker. This method also used a separate network for each user to make the final user/impostor discrimination.

Note that for all methods, the evaluation phase first consisted of determining which of the users the unknown speaker was most similar to. For the first method mentioned, the statistical features were then computed using the profile computed for the networks of this 'candidate' user. The single secondary network was then used to determine whether the features belonged to the user or the impostor group. For the second method, the profile of the best 'candidate' user was computed, and the secondary network specific to that user was used to make the user/impostor decision. In the final method, statistics were computed based up on the best 'candidate' user, and the secondary network specific to that user was used to make the user/impostor decision. For both the first and third methods, the secondary feature vector remains constant in size as the user population varies.

EXPERIMENTS

In all experiments initial network training was conducted on 'training' speech and evaluation was conducted on separate 'test' speech. For the case of users, test speech consisted of different sentences from the same speaker. The number of users was either 10, 20, or 40, as noted below. Impostors were different speakers than those used in any of the training. Several experiments were conducted with each of the methods mentioned above. For each case, the primary neural networks were trained using four sentences (7-10) from each user.

EXPERIMENTAL RESULTS

Experiment Set 1

These experiments were used to evaluate the first method of secondary feature extraction. The statistical information for training the secondary neural network was computed using 3 sentences for each speaker. These experiments showed that good performance could be achieved using the first and second moments of the user profiles. The best evaluation rate was 93.7% based on approximately 2.7 seconds of speech (20 users and 80 impostors ???).

However, it was clear from the results of these experiments that the primary cue used to separate impostors and users was simply the mean level of the relevant networks. That is 'real' users would have higher levels of relevant networks than would impostors. It was also clear from the results that the features selected did not carry sufficient information to accurately discriminate users and impostors. Also, it did not seem likely that for a large number of users, the entire user population would form a distinct group from all impostors of all users.

Experiment 2

In this set of experiments we used the second method described for processing speaker profiles. These experiments were conducted with a varying number of users and a fixed number of impostors. In each case 80 impostors were used and 3 sentences were used to create user profiles. Evaluation results were computed on a single sentence basis. On average a single sentence in the TIMIT database is approximately 2.7 seconds of speech. The user set was varied from 10 - 40 by factors of 2.

Despite generally good results with this method, a major difficulty was related to the lack of sufficient training data to adequately train each of the secondary neural networks. Note that for training purposes, each user was utilized as an impostor for all the other users. Thus, for an M user case, for each training sentence, we would obtain 1 user profile token and M-1 impostor token profiles. For large numbers of users, the amount of impostor training data was much larger than the amount of user training data. Thus using a typical minimum mean square error network training criteria (over all data), the secondary networks would be biased to conclude that all speakers are impostors, since, on the average, this would be good decision. To overcome this limitation we modified the training algorithm of the neural network to alternate between user and impostor presentations. Thus the total number of weight updates for the two categories would be the same, even though this meant each specific user feature vector was presented to the network much more than each specific impostor feature vector.

Table 1 presents results of these experiments. The value called ‘select’ was used to indicate whether or not the new training method was employed. A ‘select’ value of 0 indicated the new weight update method was used, 1 indicated the traditional method was used.

The advantages of the new training method (select = 0) can be clearly seen in the 40 user results shown in Table 1. In this case each user category contained 3 training tokens, when 3 sentences were used for training. However, each impostor category contained 117 training tokens. Thus, with traditional training (select = 1) the network learns to classify almost all speakers as impostors.

Experiment 3

The final experiment tested the third method mentioned above. That is statistics of user profiles, and statistics representing variability in networks across time were used to represent each user. A separate network was used to separate each user, and impostors posing as that user.

In these experiments the training method used was identical to the method outlined in experiment 2 (select = 0). Profiles were generated based on 7 training sentences (4 - 10). Sentences 1 - 3 were used for evaluation purposes. Several cases of this experiment were conducted, computing performance as a function of input test speech length. The performance measure, performance index (PI), is the mean between user and impostor recognition rates. Experiments were based on 20 and 40 users cases. In all experiments the entire DARPA/TIMIT database was used, leading to 620 and 590 impostors, respectively. The results (see Figure 1) using the final approach indicate this method is very robust, even with a large number of impostors.

Table 1 goes here (to end of page)

Figure 1 Goes Here

Figure 1: Experiment 4, performance index as a function of input speech length

CONCLUSION

A new neural network method has been presented for speaker verification. A primary collection of networks is used for pairwise speaker discrimination, and then for speaker identification. For the purposes of verification, each user is represented by his position relative to other users, as determined by the primary networks. A secondary set of networks is used to determine whether an unknown speaker is a user or impostor. Experimental results show that the method performs with over 97% accuracy with 40 users, 590 impostors and one sentence of speech (about 2.7 seconds) for verification.

REFERENCES

C.A. Norton, S.A. Zahorian and Z.B. Nossair, “The Application of Binary-pair Partitioned Neural Networks to the Speaker Verification Task,” *ANNIE ‘94,* pp. 441-446.

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

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.

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.

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