SPEAKER IDENTIFICATION/VERIFICATION USING NEURAL NETWORK DISTANCE MEASURES

Claude A. Norton, III, Stephen A. Zahorian and John Kurtz
Department of Electrical and Computer Engineering
Old Dominion University, Norfolk, Virginia 23529

INTRODUCTION

A method for using a collection of binary neural network discriminative classifiers is described and evaluated for a speaker identification/verification system. In the first step of the procedure, feed-forward 2-layer networks are trained to discriminate every pair of “users” for the system. The outputs of this collection of networks are used as a distance metric for each user to each other user. These distance measures create a profile for each user, representing the position of that user relative to all other users. Statistics of these profiles form low-dimensionality “signatures” for each user. These signatures are used to train secondary neural networks (one for each user) to distinguish each user from all other users (i.e., impostors for that user). With this system, a speech sample from an unknown speaker is first classified in terms of the “closest” user, and then further determined to be either that user or an impostor. Experimental results using all 630 speakers from the TIMIT and NTIMIT data base are presented. Average results, in terms of an unknown speaker correctly identified or properly rejected are 99% and 88% respectively for the two data bases using 7 sentences for training and 3 sentences for testing from each unknown speaker.

METHOD

All speech segments were first analyzed using a 29 term Discrete Cosine Transform Coefficient (DCTC) expansion (Zahorian and Jagharghi, 1993). This analysis was performed on a 30 ms Hamming windowed speech frames computed every 10 ms, using a bilinear warping factor of .25. The zero order (“DC” term) of the DCTC analysis, which is similar to cepstral analysis, was not used. These spectral features were then scaled to have a mean of zero and standard deviation of 0.2. For each pair of users in the system, a two-layer feed forward fully interconnected “primary” network with 10 hidden nodes was trained with backpropagation to discriminate the two users in that pair. The number of network updates was determined by maximizing performance on training speech. Thus for M users, a total of M * (M-1) /2 networks were trained, with M-1 networks devoted to discriminating each particular user from the other M-1 users.

The next step of the training procedure was to use the primary networks to create sample profiles for each user, and for impostors of that user, as follows. Each training segment of speech for each user was evaluated by the M-1 networks pertinent to that user for each frame of speech. The average levels of these M-1 networks were then used to form an M-1 dimensional vector, giving a measure of the distance of that user to each of the other users. Another M-1 dimensional vector was formed from the standard deviations of each network output (computed over all the frames in the training segment) to give an estimate of the consistency of the distance measures in the first M-1 dimensional vector. For each user, these steps were completed both with samples of speech from that user, and with samples of...
speech from all the other users (impostors for that user). For each of the two (M-1) dimensional vectors characterizing valid and impostor training speech, the mean and 3 central moments were computed to form signatures for each user, and for impostors of that user. Finally a secondary network, with the same configuration as the primary networks, was trained to discriminate each user from impostors of that user. This entire process was repeated for each of the M users. Thus a total of M binary secondary networks were trained.

The evaluation phase consisted of first determining which of the users an unknown speaker was most similar to, using only the primary networks mentioned above. Thus this step was performed using the same method as previously reported for Binary-Paired Partitioning (BPP) speaker identification. (Rudasi and Zahorian, ICASSP 91). Using the processing method described above for training, statistics were computed for the best ‘candidate’ user, and the secondary network specific to that user was invoked to make the user/impostor decision.

**EXPERIMENTS**

Experiments were conducted with both the clean and telephone versions of TIMIT. Experiments were conducted for 49 users selected both sequentially and randomly. The random case noted MIT (because of it’s origin as a set of test speaker for phone classification experiments at MIT) represents an accurate gender distribution throughout the speaker population. The remainder of the database was used as impostors (i. e., 581 impostors). For each case, the user discrimination neural networks were trained using four sentences (7-10) from each user. Profiles were generated with 7 training sentences (4-10). Final evaluation was performed with the remaining 3 sentences (1-3), for 1, 2, and 3 sentences. (Note sentences are ordered as SA, SI, SX.)

The average of the user acceptance and impostor rejection rate was evaluated as a performance index (PI), as a function of input test speech length. The results for TIMIT plotted in Figure 1 range from 73% correct, using .16 seconds of evaluation speech to over 99.7% correct, using about 8 seconds (3 sentences) of evaluation speech. There was very little difference between using sequentially selected or randomly selected users. Results for a similar experiment, except based on the NTIMIT data, are given in Table 1 and range from 72% accuracy (with .16 seconds of test speech) to 86% accuracy with three sentences of test speech.

**CONCLUSION**

A new method has been presented for speaker identification/verification. A collection of primary 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 with clean speech show that the method can achieve nearly 100% accuracy with 49 users. The method also extends to noisy speech with reduced performance.
Figure 1: Performance Index as a function of input speech length

Table 1. Speaker identification/verification accuracy for NTIMIT for various test speech lengths.

<table>
<thead>
<tr>
<th>Length (s)</th>
<th>User % Correct</th>
<th>Impostor % correct</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.19</td>
<td>27.6</td>
<td>96.3</td>
<td>62.0</td>
</tr>
<tr>
<td>0.38</td>
<td>35.8</td>
<td>96.4</td>
<td>66.1</td>
</tr>
<tr>
<td>0.76</td>
<td>44.2</td>
<td>96.6</td>
<td>70.4</td>
</tr>
<tr>
<td>1.52</td>
<td>55.8</td>
<td>96.5</td>
<td>76.2</td>
</tr>
<tr>
<td>3.05</td>
<td>96.3</td>
<td>96.7</td>
<td>80.0</td>
</tr>
<tr>
<td>6.10</td>
<td>73.3</td>
<td>97.2</td>
<td>85.3</td>
</tr>
<tr>
<td>9.15</td>
<td>75.0</td>
<td>97.2</td>
<td>86.1</td>
</tr>
</tbody>
</table>

REFERENCES

