AUTOMATIC GENDER IDENTIFICATION BY VOICE

X.Y. Zhu and L.W. Cahill
Department of Electronic Engineering, La Trobe University

ABSTRACT - Gender identification is a sub-area of speaker identification. This paper reports a traditional pattern recognition method and a connectionist method of automatic gender identification. Six different acoustic features extracted from 16 millisecond segments of the vowel [i] were used and compared in the performance of gender identification. Fifty male speakers and fifty female speakers from the DARPA TIMIT speech corpus were tested in the experiments. The results show that speaker's sex can be identified accurately from very short segments of his or her voice by the proposed methods. For a large number of subjects, the connectionist model is superior to the traditional pattern recognition method in gender identification.

INTRODUCTION

Speaker identification has many applications and is of great interest in the speech research area. Gender identification is a subset of speaker identification. Automatic gender identification would allow search space to be cut in half in both speaker-independent speech recognition and speaker recognition.

This paper describes research into automatic gender identification. The speaker's sex is identified from short segments of his or her voice. A pattern recognition method and a connectionist model are proposed in the paper. The results are comparable with other algorithms, since the standard acoustic-phonetic database TIMIT was used in the experiments. The effectiveness of various acoustic features, which reflect vocal chord and vocal tract shape, are also compared.

SPEECH DATABASE

The data used in the experiments are from the DARPA TIMIT speech corpus (Zue et al., 1990). The corpus contains 6300 continuous sentences read by 630 speakers from 8 major dialect regions of the United States. The speech was sampled by a 16 bit A/D converter at 16 KHz. Acoustic-phonetic labels are included in the database.

Fussell compared six classes of phonemes in gender identification (Fussell, 1991). In all cases, vowels gave the best result. Therefore, in the present study, a vowel was used to distinguish sex. The test data are one hundred 16-millisecond-long (256 points) frames segmented from the central portion of the vowel [i]. The vowel [i] was segmented from *she* in the first calibration sentence *sr1*. Fifty male speakers and fifty female speakers from the 8 dialect regions were selected as test subjects. For each subject, only one frame was used.

ACOUSTIC FEATURES

The following six sets of acoustic features were extracted through linear prediction analysis (Markel & Gray, 1976):

- linear prediction coefficients (LPC);
- cepstral coefficients (CC);
- autocorrelation coefficients (ARC);

- reflection coefficients (RC);
- area functions of vocal tract (AF);
- log area ratios (LAR).

The analysis conditions were:

- filter order: 12;
- analysis frame: 256 points in the central portion of vowel [i];
- speech pre-emphasis factor: 0.95;
- analysis window: Hamming; and
- analysis: autocorrelation.

THE PATTERN RECOGNITION METHOD

In the pattern recognition method, a test speaker's acoustic features are compared with a male reference pattern and a female reference pattern. The reference patterns are the average features of all training subjects. The Euclidean distance matrix is used in the pattern recognition, since it has been shown to be the most efficient in gender identification (Childers, et al., 1988). The Euclidean D distance is defined as

$$D(X,Y) = |(X-Y)^T(X-Y)|^{1/2} \tag{1}$$

where X and Y are the test and reference vectors respectively, and T denotes transpose.

THE CONNECTIONIST MODEL

The connectionist model used in the experiment is a fully interconnected, feedforward, multi-layer perceptron (MLP). The MLP has thirteen inputs, eighteen units in a single hidden layer, and one output. The inputs of the MLP were the acoustic coefficients of the speaker. The output corresponded to the speaker's gender.

The model was trained by the back-propagation training algorithm (Hecht-Nielsen,1989). The desired output was 0.75 for female speakers, and 0.25 for male speakers. Male and female subjects trained the model alternately in the training process. Each training session involved training all the subjects in turn.

EXPERIMENTS AND DISCUSSION

Experiment 1

Experiment 1 aimed to compare the effectiveness of the various acoustic features in gender identification. Twenty five male subjects and twenty five female subjects were tested by the pattern recognition method The comparisons included all six acoustic features described above. The leave-one-out or exclusive training procedure was employed, i.e., the training procedure used all the data except the data of the test subject. The result of Experiment 1 is shown in Figure 1.

Experiment 2

Experiment 2 aimed to compare the connectionist model with the pattern recognition method. In this experiment, three sets of acoustic features: cepstral coefficients, reflection coefficients and area functions of vocal tract, were used. The exclusive training procedure was employed. To explore the performances of the methods dependent on the number of the test subjects N, identification rates were tested on various numbers of N. Figure 2 shows the results of the experiment for $N = 20$ (10 male and 10 female), $N = 50$ (25 male and 25 female), and $N = 100$ (50 male and 50 female). In the connectionist model, 1000 training sessions were completed, except for $N = 20$. Two hundred training sessions were carried out in $N = 20$ case, since it gave a better result than for the 1000 training sessions.

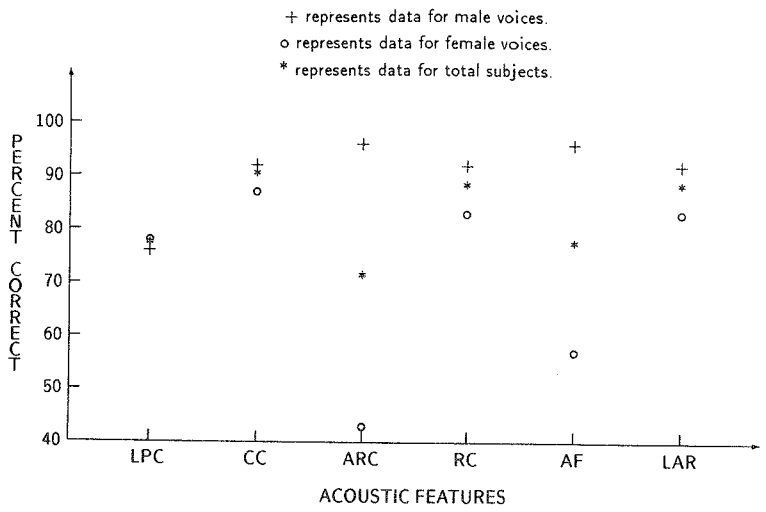

Figure 1. Comparisons of effectiveness of various acoustic features in gender identification using the pattern recognition method.

Discussion

The results presented in Figures 1 and 2 were obtained by using the exclusive training procedure. The inclusive training procedure was also used in the experiment, i.e., all the data, include the data of the test subject, were used to develop the reference pattern or model. In the test of the inclusive training, the connectionist model gave 100% identification rate in all the cases. This means that this one hidden layer MLP model is good enough to classify such large data into two classes. The pattern recognition with the inclusive training gave better results than those for the exclusive training in the case of $N = 20$. However, in the cases of large numbers of test subjects, the inclusive training was not shown significantly better than the exclusive training, since the test subject only had little contribution to the reference pattern.

If we study the performances of the methods for each individual subject, we make the following conclusions from the results of the experiments. First, the identification of male speakers is better than that of female speakers on average. In Figure 1, we can see that male speakers had a higher correct rate for all the acoustic features except for LPC. This is also true in experiment 2, (not shown in the results above).

Secondly, there were three *goat* subjects (one male and two female) in the test set. Using most of the acoustic features, the three subjects were wrongly identified in both methods. This performance is possibly dependent upon the tested portion of sound. Since only the vowel [i] was tested in the experiments, the above two points are not necessarily true for other vowels or other phonetics, such as consonants and nasals.

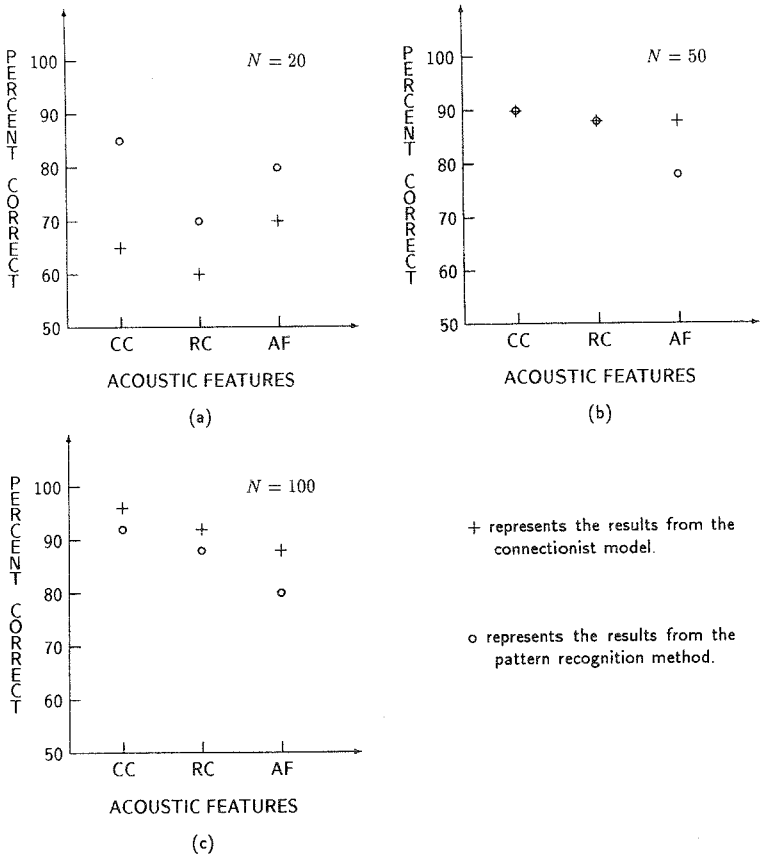

Figure 2. Comparisons of the connectionist model and the pattern recognition method. (a) Number of test subjects $N = 20$. (b) Number of test subjects $N = 50$. (c) Number of test subjects $N = 100$.

The third point is that the results of the two methods, and the results for the various acoustic features, are partly complementary. None of them is exactly same as any other. So it is possible to combine the two methods with different acoustic features to get a higher identification rate.

If the effectiveness of all the acoustic features is compared, from Figures 1 and 2, we can see that cepstral coefficients gave the best and most robust results in all the cases. This concurs with results of previous gender identification work (Fussell, 1991).

Comparing the performances of the connectionist model and the pattern recognition method, from Figure 2, we can see that the pattern recognition method gave stable identification rates irrespective of the value of N, while the connectionist model gave better results with larger values of N. This is because neural networks need a large set of training data. So their performances are better than those of conventional models in large populations, for instance, $N = 100$ in our experiment. For a small number of subjects (for example, $N = 20$ in our experiment), traditional methods are preferred. In experiment 2, we found that for large values of N, the connectionist model produced consistently better results than did the pattern recognition method for all

acoustic features. The better performance of the connectionist approach was most significant in the case of the AF feature. This may suggest that neural networks, in classifying gender, can make better use of a large information base.

CONCLUSIONS

From the discussion above, the following two conclusions may be drawn. The first one is that a high gender identification rate can be obtained from very short segments of speakers' voices. In the present study, up to 96% gender identification success is obtained from the test set of one hundred speakers by using the connectionist model with cepstral coefficients extracted from 16 millisecond segments of the vowel [i]. If various methods and a number of acoustic features are combined, even better performance may be reached. This proves that machine performance can exceed human performance in gender identification. The second conclusion is that connectionist models are superior to conventional models in classifying gender from large populations.

The present work has been based on the DARPA TIMIT speech corpus, where the quality of speech is high. Gender identification on lower quality speech need to be researched in future work.

REFERENCES

Childers, D.G., Wu, K., Bae, K.S. & Hicks, D.M. (1988) *Automatic recognition of gender by voice*, ICASSP'88, New York, NY, USA, 603-606.

Fussell, J.W. (1990) *Automatic sex identification from short segments of speech*, ICASSP'90, New Mexico, NM, USA, 409-412.

Hecht-Nielsen, R. (1989) *Neurocomputing*, (Addison-Wesley).

Markel, J.D. & Gray, A.H., Jr (1976) *Linear Prediction of Speech*,(Springer-Verlag: New York).

Zue, V., Seneff, S. & Glass, J. (1990) *Speech database development at MIT: TIMIT and beyond*, Speech Communication, Vol.9, 351-356.

A STUDY ON THE COMBINATION OF HIDDEN MARKOV MODELS AND MULTI-LAYER PERCEPTRON FOR SPEECH RECOGNITION

J. M. Song

Speech Technology Research Group
Department of Electrical Engineering
The University of Sydney

ABSTRACT - This paper presents an enhanced speech recognition algorithm by combining continuous hidden Markov modelling (HMM) with a multi-layer perceptron (MLP). The first stage of speech recognition, carried out by the HMM, selects a small group of candidates and projects incoming speech vectors into state normalized vectors. The second stage of MLP classifies each normalized vector generated by the HMM and determines the best candidate. In this architecture, the HMM plays a role of pre-classification, while the MLP is used for decision refinement. A simple speaker-independent isolated digits telephone speech database was used to test this approach. The result shows that the recognition performance increases from 92.9% to 93.8%.

INTRODUCTION

It is well known that hidden Markov modelling can be used successfully for automatic speech recognition. The power of this statistical pattern matching approach lies in the capability to modelling the spectral characteristics and capture the underlying temporal or dynamic structure of speech signal. However, the standard HMM based on the maximum likelihood criterion (ML) still suffers from some weaknesses caused by several assumptions. For instance, the HMM requires each acoustic feature vector to be independent, and the probability distribution of acoustic vectors in the corresponding space to be approximated by mixtures of Gaussian distributions. Even the model training paradigm remains questionable, i.e. each model is trained independently, rather than competitively. All these problems reduce the discriminative power in classification. Although some of these problems can be tackled within the HMM framework by using different optimum criteria, such as maximum mutual information (MMI) or corrective training, the advantages of these approaches are not totally clear.

There has been a lot of research devoted to the field of artificial neural networks (ANN) and it has been demonstrated that ANNs produce powerful discriminative functions to classifying static patterns. In ANN knowledge or constraints are not encoded in individual units, rules, or procedures, but distributed across many simple computing units. Uncertainty is modeled not as likelihoods or probability density functions in individual model, but by patterns of activity of many units. The training paradigm is competitive, i.e. not only does it increases the chance for correct responses, but it also discourages the mis-classified outputs. Once the training phase is finished, the classification task can be performed very fast due to the massive parallel structure of neural networks. Unfortunately ANNs also have certain weaknesses for use in speech recognition. For instance ANNs require a fixed-size vector applied to input layer and speech pre-segmented into speech units (word or sub-word) due to their inability to deal with the time sequential nature of speech. The research in the incorporation of HMM and ANN has become very attractive (Bourlard et al, 1992). The work presented in this paper combines of hidden Markov models and multi-layer perceptron together within a speech recognition system and demonstrates that a better recognition performance can be achieved by taking advantage of the both components.

THE HMM APPROACH

A probabilistic function of a hidden Markov chain is a stochastic process of two interrelated mechanisms, an underlying Markov chain having the finite number of states, and a set of random functions, each of which is associated with the individual state. At a discrete instant of time, the process is assumed to be in