Automatic Speaker Recognition System using the Discrete Hartley Transform and an Artificial Neural Network

Christopher N. Gedo and John C. Eremic

Naval Postgraduate School
Monterey, California 93943

Abstract

A system is presented which will identify speakers that it was trained to recognize with 100% accuracy. This system will also classify inputs that it has not been trained to recognize as unknown. The high degree of confidence offered by this system is attributed to careful processing of the input data sets. Only easily distinguishable segments of the speech samples are passed as inputs to the artificial neural network (ANN). Additionally, sufficient input information is provided with the training inputs to enable the ANN to rapidly train and subsequently classify speakers. This system may be readily implemented in hardware using existing technology. Hardware implementation will render near real-time performance. Applications that would be enhanced by the use of this system include electronic surveillance and computer security.

1 Introduction

The speaker identification system consists of a speech sampler, a signal pre-processor, an ANN, and a network post-processor. The system description begins with a discussion of the techniques used to sample and window data. Pre-processing techniques including definitions of the short-time energy, zero-crossing rate, Discrete Hartley Transform (DHT), and characteristic eigenvector of each speech segment are presented. An example using data from the DARPA TIMIT speech corpus illustrates the pre-processing procedure. Neural network employment including feature vector formulation and post-processing of the network output concludes the system description. Experimental results are provided and demonstrate system effectiveness.

2 Sampling and windowing technique

The speech is sampled at 8 kHz and segmented using a 128 point (16 ms) rectangular window. A short window ensures the stationarity of data within a segment and allows maximum selectivity when choosing segments to process from short-duration speech samples. Each segment overlaps the previous segment by 64 samples. The overlap effectively samples the rectangular window output at its Nyquist rate. The bandwidth of the rectangular window is

$$BW_{rect} = \frac{1}{NT} = \frac{8000}{128} = 62.5\text{ Hz}. \tag{1}$$

3 Pre-processing technique

The short-time energy and zero-crossing rate (ZCR) are used to differentiate between voiced and unvoiced speech. The discrete convolution

$$E_n = \sum_{m=-\infty}^{\infty} s^2(m)w(n-m) \tag{2}$$

where *s(m)* is the speech signal and *w(m)* is the windowing function defines the short-time energy for the n^{th} segment[1]. It emphasizes the differences in amplitude between voiced speech, unvoiced speech and non-speech portions of the input data set.

The ZCR provides a simple, but accurate, means of spectral measurement [2]. This spectral measure is used to estimate the fundamental vocal tract frequency and characterize the voiced or unvoiced nature of speech segments. The fundamental frequency of a signal as a function of the zero-crossing rate is

$$f_0 = \frac{ZCR}{2} f_s \tag{3}$$

where f_s is the sampling frequency. The ZCR of the n^{th} segment is

$$ZCR_n = \frac{1}{2} \sum_{m=-\infty}^{\infty} \left|\operatorname{sgn}[s(m)] - \operatorname{sgn}[s(m-1)]\right| w(n-m) \tag{4}$$

where the sign function is defined to be

$$\operatorname{sgn}(x) = \begin{Bmatrix} 1, x \geq 0 \\ -1, x < 0 \end{Bmatrix}. \tag{5}$$

As used in the speaker identification system, both the short-time energy and ZCR for each speech segment are compared to thresholds. These thresholds are adjusted until the desired percentage of speech segments pass. This thresholding procedure ensures that the data processed and subsequently presented to the ANN is the most distinctive in nature.

Segments with energy and ZCR above the thresholds are further processed by taking their Discrete Hartley Transform (DHT). The DHT developed by Bracewell [3] is defined to be

$$H[s(m)] = \sum_{m=0}^{N-1} s(m) \cdot \left[\cos\left(\frac{2\pi km}{N}\right) + \sin\left(\frac{2\pi km}{N}\right) \right]. \qquad (6)$$

The relationship between the DHT and the DFT is

$$H[s(t)] = M(\omega) \cdot [\cos\phi(\omega) + \sin\phi(\omega)] \qquad (7)$$

where $M(\omega)$ is the DFT magnitude and $\phi(\omega)$ is the DFT phase[4]. Equation (7) illustrates the convenient way that the DHT incorporates magnitude and phase information into a single real term.

The DHT coefficients of the selected speech segments are further processed by correlating them. The k^{th} lag of the autocorrelation of a speech segment is defined as

$$r_k = \frac{1}{N} \sum_{m=0}^{N-1} H(m) H(m+k) \qquad (8)$$

where $H(m)$ refers to the DHT coefficients for the segment, N is the window length, and $0 \le k \le M-1$. An $M \times M$ dimensional toeplitz matrix of the form

$$\mathbf{R}_n = \begin{bmatrix} r_0 & r_1 & \cdots & r_{M-1} \\ r_1 & r_0 & \cdots & r_{M-2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{M-1} & r_{M-2} & \cdots & r_0 \end{bmatrix} \qquad (9)$$

is formed for the n^{th} selected segment. Computation of the eigenvector associated with the largest eigenvalue of $\mathbf{R}_n$ completes the pre-processing of a speech segment. This eigenvector characterizes a processed speech segment and is presented along with the short-time energy and ZCR as the input feature vector to the ANN.

To illustrate the pre-processing procedure, the phoneme *"iy"* is isolated for two male speakers from region one of the DARPA TIMIT speech corpus. A single segment is processed for each speaker. The time domain speech segments and resulting eigenvectors with $M = 16$ are shown in Fig. 1 for the two speakers.

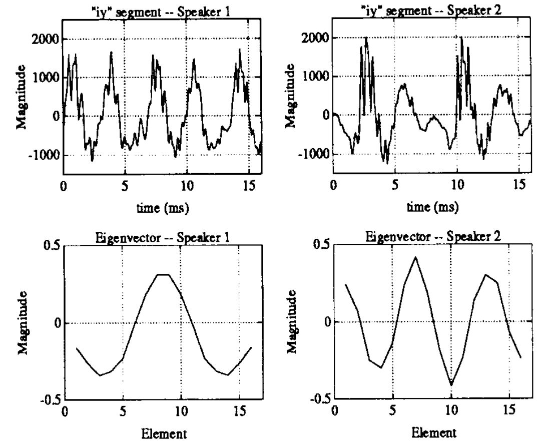

Fig. 1: Time domain speech segments and processed characteristic eigenvectors for two speakers.

To illustrate the influence of the phase on the eigenvectors, the same segments were processed using FFT magnitudes instead of the DHT coefficients. The resulting eigenvectors are shown in Fig. 2.

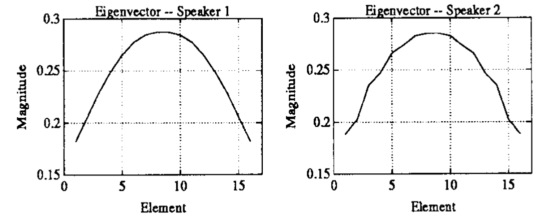

Fig. 2: Eigenvectors derived from FFT magnitudes for the speech segments of Fig. 1.

Comparing the eigenvectors of Fig. 1 to those of Fig. 2, it is clear that including phase information in the preprocessing procedure produces more distinct feature vectors.

4 The artificial neural network

Although the ANN is a major part of this system, it is completely generic in nature and its function is defined completely by the input vectors and the desired output. A standard Back-Propagation (BP) or Learning Vector Quantization (LVQ) network is used for the speaker identification system. Whereas the LVQ network uses a Kohonen layer to compute the Euclidean distance between the input vector and each processing element's (PE's) weight vector, the BP network uses the hyperbolic tangent for its PE transfer function[5].

The number of input PE's is governed by the number of elements in the input vector and the number of desired outputs independent of type of network employed. Each desired output has a corresponding input PE with a linear transfer function. There are normally three or fewer hidden layers. The number of PE's for the first hidden layer is usually the same as the input layer. If additional hidden layers are used, they typically have fewer PE's than the input layer. Figure 3 illustrates a typical BP ANN architecture.

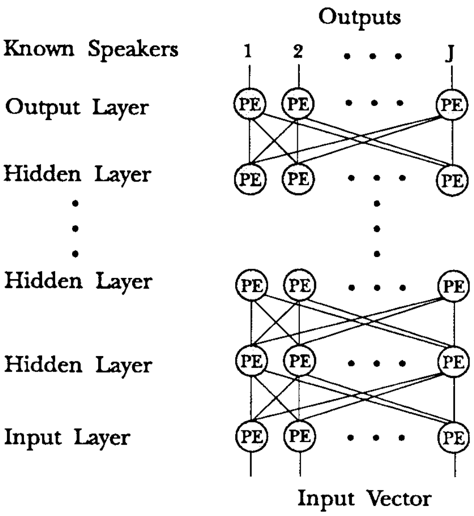

Fig. 3: Typical BP ANN architecture.

The ANN of the speaker identification system must be trained to recognize the known speakers before it can be used. A training file of input vectors consisting of the short-time energy, the ZCR, and the principal eigenvector processed from the correlated DHT coefficients is required. Additionally, each input training vector is appended with a desired output code having the same number of elements as there are outputs. These outputs represent the known speakers. The output code consists of a *one* for the desired output and a *zero* for all other outputs. Figure 4 shows the contents of a training file of input vectors for a two known speaker system ($J = 2$).

The short-time energy and ZCR are used to clarify whether a sample represents the voiced or unvoiced speech of a particular speaker. This accelerates the learning rate of the ANN by reducing the ambiguity caused by having two types of input vectors with very different characteristics representing the same speaker. It is desirable to provide the ANN with as many training inputs as possible to ensure accurate classification.

The ANN is trained until the desired accuracy is obtained. Once trained, the speaker identification system may be used.

```
Short-Time Energy, ZCR, E(1), E(2),  . . .  , E(M), 1, 0
Short-Time Energy, ZCR, E(1), E(2),  . . .  , E(M), 1, 0
                                  •
                                  •
                                  •
Short-Time Energy, ZCR, E(1), E(2),  . . .  , E(M), 1, 0
Short-Time Energy, ZCR, E(1), E(2),  . . .  , E(M), 0, 1
Short-Time Energy, ZCR, E(1), E(2),  . . .  , E(M), 0, 1
                                  •
                                  •
                                  •
Short-Time Energy, ZCR, E(1), E(2),  . . .  , E(M), 0, 1
```

Fig. 4: Typical training file of input vectors for a two known speaker system.

A speech sample is collected and processed. A file of input vectors similar to those of Fig. 4 without the training codes is presented to the ANN.

5 ANN post-processing

The outputs of the ANN must be interpreted so that a classification can be assigned to each speech sample. The ANN outputs for the feature vectors of the chosen segments from each speech sample are compared to a threshold, typically > 0.5, and then tallied. This threshold is required for BP networks because the output values normally range from 0 to 1. Conversely, an LVQ network quantizes its outputs at 0 or 1 and does not require a threshold. If a high percentage of the outputs are tallied for one category with a corresponding low percentage for the others, a classification is made. If a sample does not meet this criterion, a classification of unknown is assigned.

6 Experimental results

The system has been tested on a limited case with superior results. Speech samples were obtained from three male speakers who exhibited similar pitch levels in their speech. The ANN was trained to recognize two of the speakers using eight samples of speech per speaker. Each sample was two seconds in length. Samples from the third speaker were reserved to test the system's ability to reject an unknown speaker. With short-time energy and ZCR thresholds set to retain 12% of the speech segments and 16 element eigenvectors, 350 feature vectors per speaker were presented to the ANN for training. A BP ANN was trained for 60,000 cycles and then tested. The results of that test are shown in table 1.

The output value threshold to assign a choice was set at 0.8. The BP ANN correctly identified both speakers with significantly more accuracy than it did the unknown speaker. The same data was presented to an LVQ ANN. In this case, the unknown speaker was rejected with more confidence than with the BP ANN because the outputs were essentially split between the two known speakers. Table 2 shows the results of the test using an LVQ ANN.

Speaker	Choice 1	Choice 2
1	58/72	0/72
2	4/72	55/72
unknown	111/205	44/205

Table 1: Results of test with two known speakers, one unknown speaker, and a BP ANN.

Speaker	Choice 1	Choice 2
1	71/72	1/72
2	11/72	61/72
unknown	102/205	103/205

Table 2: Results of test with two known speakers, one unknown speaker, and an LVQ ANN.

When the output is considered in a statistical sense the ANN demonstrates the ability to classify speakers with no ambiguity. It is interesting to note that in this BP case, the RMS error never converged during the training phase before conducting the test. This indicates that a restrictive RMS error threshold is not necessary to guarantee robust performance of the system.

7 Conclusions

The speaker identification system was successfully tested. Although the test was limited, results indicate that it will perform well in more general cases. It is important to use phase information when pre-processing the data; the DHT conveniently combines phase and magnitude information into a single real term. Quantities such as the short-time energy and ZCR characterize the voiced or unvoiced nature of a speech segment and are necessary to expedite training and testing of the ANN. The thresholding scheme to choose which speech segments to process ensures that only distinctive portions of the speech sample are presented to the ANN for classification.

References

[1] O'Shaughnessy, Douglas, *Speech Communication, Human and Machine*, Reading, Massachusetts: Addison-Wesley, 1990.

[2] Rabiner, L.R. and Schafer, R.W., Digital Processing of Speech signals, Englewood Cliffs, New Jersey, 1978.

[3] "Discrete Hartley Transform", K.N. Bracewell, *J. Opt. Soc. Am.*, Vol 173, No12, pp 1832-1835, Dec 83.

[4] "The Hartley Transform Applied to Speech Coding," E. Chilton, *IEE Colloquium on Speech Coding*, IEE London, Nov 89.

[5] Rumelhart, D.E. and McClelland, J.L., "Parallel Distributed Processing: Explorations in the Micro-Structure of Cognition", Vol I, *Foundations*, MIT Press, 1986.