1-1-1998

Voice Recognition using Neural Networks

Ganesh K. Venayagamoorthy
Missouri University of Science and Technology

Viresh Moonasar

K. Sandrasegaran

Recommended Citation

G. K. Venayagamoorthy et al., "Voice Recognition using Neural Networks," *Proceedings of the 1998 South African Symposium on Communications and Signal Processing, 1998. COMSIG '98*, Institute of Electrical and Electronics Engineers (IEEE), Jan 1998.
The definitive version is available at https://doi.org/10.1109/COMSIG.1998.736916

VOICE RECOGNITION USING NEURAL NETWORKS

Ganesh K Venayagamoorthy, Viresh Moonasar and Kumbes Sandrasegaran*
Electronics Engineering Department, M L Sultan Technikon, Durban, South Africa
gkumar@saiee.org.co.za
**Institute for Information Sciences and Technology (IIST), Massey University, New Zealand*
K.Sandrasegaran@massey.ac.nz

Abstract— **One solution to the crime and illegal immigration problem facing South Africa is the use of biometrics techniques and technology. Biometrics are methods for recognizing a user based on unique physiological and/or behavioural characteristics of the user. This paper presents the results of an ongoing work in using neural networks for voice recognition.**

Keywords— **Voice recognition, Neural Networks**

I. INTRODUCTION

Crime and illegal immigration in South Africa has reached unprecedented levels. Although the gap between the rich and the poor is the main cause of this crime, it is the person in the middle who ends up paying for it. Long term and short term solutions are needed urgently. Conventional keys, access codes, access cards are proving ineffective as they are easily lost, stolen, copied, observed or left at home. The goal of this work is to come up with innovative, but inexpensive, solutions to the crime problem using emerging technologies such as biometrics, artificial intelligence, computer networks, signal processing, etc. The long-term goal of the work is to design a black box that will identify a known user based on characteristic features such as speech, image, etc.

Biometrics are methods for recognizing a user based on unique physiological and/or behavioural characteristics of the user. These characteristics include finger prints, speech, face, retina, iris, hand-written signature, hand geometry, wrist veins, etc. Biometrics system are being commercially developed for a number of financial and security applications. The task performed by this system can be classified into identification and verification. Identification involves identifying a user from a database of user characteristics whereas verification involves authenticating a user's identity using a pattern in its database.

Of all the above mentioned human traits used in Biometrics, the one that humans learn to recognize first is the voice characteristic. Infants can identify the voice of their mothers and telephone users can identify a caller on a noisy telephone line. Furthermore, the bandwidth associated with the speech is also much smaller than the other image based human traits. This implies quicker processing and smaller storage space.

Speaker recognition systems can be divided into two namely: text-dependent and text-independent systems. In text-dependent systems, the user is expected to use the same text (keyword or sentence) during training and recognition sessions. A text independent system does not use the training text during recognition session. Both systems perform the following tasks: feature extraction, similarity analysis and selection. Feature extraction uses the spectral envelope to adjust a set of coefficients in a predictive system. One voice sample can then be compared for similarity with another sample by computing the regression between the coefficients. This is similarity analysis. A number of normalization techniques have been developed to account for variation of the speech signals.

This paper will present the results of an ongoing work in using neural networks for voice recognition. This work should be contrasted with speech recognition where the goal is to identify the words spoken by a user. The goal of this work is to identify a user using his voice patterns. This justifies the use of the term "voice recognition". All users utter the same test word or a phrase for identification purposes. This paper will the present the results from a test phrase.

The speech signals corresponding to a test phrase of a group of people are recorded in voice files on a computer using sound recording software. The information in these files are converted from the time domain to the frequency domain using digital signal processing techniques.

The frequency spectra of the speech signal is used to train a neural network. The frequency range of the human voice (0.2 to 3.2 kHz) is converted to a vector. This vector forms the input to the neural network. The outputs of the neural network is the identity of the user.

After training the neural network using the recorded voice patterns, it is tested in a real-time environment to identify any of the groups of people trained. It is also tested with untrained users.

This paper will present the voice patterns of some users, explain the digital signal processing involved and describe in detail the neural network architecture employed and how it identifies the user and with what sort of accuracy.

II. RECOGNITION SYSTEM

A block diagram of a typical speech/speaker recognition system is shown in Fig. 1.

The system is trained to recognize a person's voice by each person speaking out a specific utterance into the microphone.

0-7803-5054-5.0029 $10.00

The speech signal is digitized and some signal processing is carried out to create a template for the voice pattern and this is stored in memory.

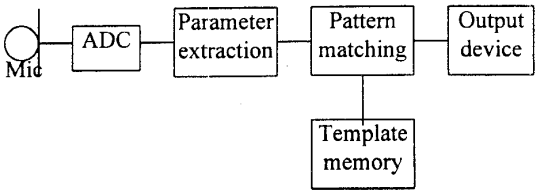

Fig. 1. Block diagram of a typical speech recognition system.

The system recognizes a speaker by comparing the utterance with the respective template stored in the memory. When a match occurs the speaker is identified. The two important operations in an identifier are parameter extraction, where distinct patterns are obtained from the utterances of each person and used to create a template, and pattern matching, were the templates are compared with those stored in memory. Usually correlation techniques are employed for pattern matching.

In this paper, the voice recognition system investigated is shown in Fig. 2.

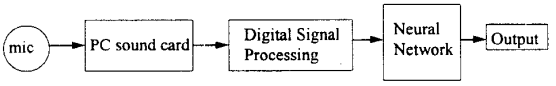

Fig.2. Block diagram of the voice recognition system.

In this paper, the voice recognition system is based on a known utterance. The system is trained on a group people to be recognized by each person saying a particular phrase. The voice is recorded on a standard 16-bit computer sound card using a directional microphone. A sampling rate of 16 kHz satisfying the Nyquist criterion is used. The voices are stored as sound files on the computer. Digital signal processing techniques are employed to convert these sound files to a presentable form as inputs to a neural network. The output of the neural network identities the speaker.

III. CHARACTERISTIC FEATURE EXTRACTION

The feature extraction plays a very important role in the speaker identification. Human speech can be sensibly interpreted using frequency-time interpretations such as a spectrogram. Frequency-energy interpretations and power spectral densities can be used to differentiate between speakers. Other methods that can be used for this purpose are the linear predictive coding and cepstral analysis, which are not reported in this paper.

The Fourier transform of discrete sample is given by relationship:

$$X(k+1) = \sum_{n=0}^{N-1} x(n+1) W_N^{kn} \qquad (1)$$

The discrete Fourier transform of the voice files are computed and the Power Spectral Density (PSD) is computed by taking the magnitude-squared result of the Fourier transform. These periodograms are then averaged and scaled. The PSD of a voice sample contains unique features attributed to an individual and these are used in our studies. These PSD values are presented in a vector form to the pattern matching network. The PSD of two different speakers is shown in Fig. 3 & 4. It can be seen from these figures that the PSD of speaker A differs from that of speaker B though there are similarities between the two. These PSDs are computed using the MATLAB signal processing toolbox [1].

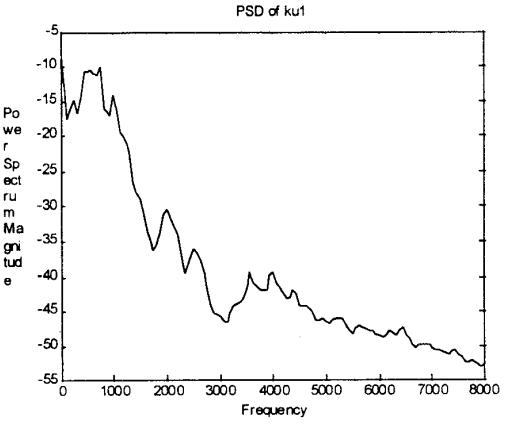

Fig.3. PSD of speaker A

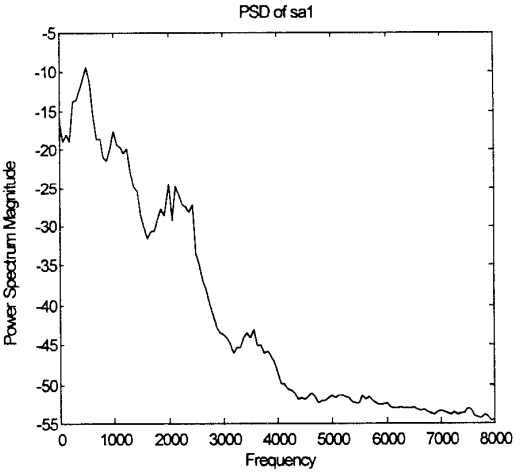

Fig.4. PSD of speaker B

Another signal processing technique to extract unique features for voice recognition is to take the time-dependent Fourier transform of voice signals. The time-dependent Fourier transform is the windowed, discrete-time Fourier transform for a sequence, computed using a sliding window. The spectrogram of a sequence is the magnitude of the time-dependent Fourier transform versus time.

The spectrogram of the same two speakers, A and B of Fig. 3 & 4 is shown in Fig. 5 & 6 respectively. Spectrograms of speaker A consists of more peaks than of speaker B. The spectrograms give a clear distinction between speaker A and speaker B.

In this paper, results using the PSD method of features extraction is only reported. Work with the spectrogram method is currently being investigated.

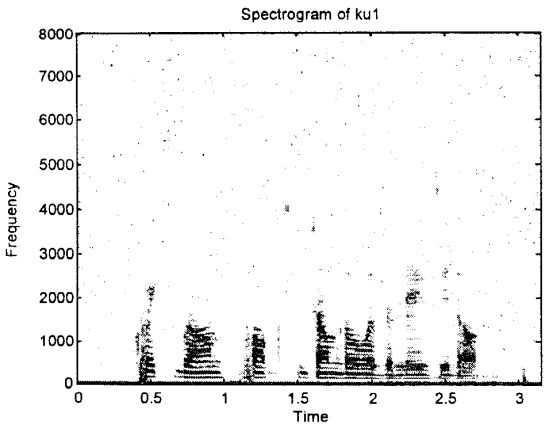

Fig. 5. Spectrogram of speaker A

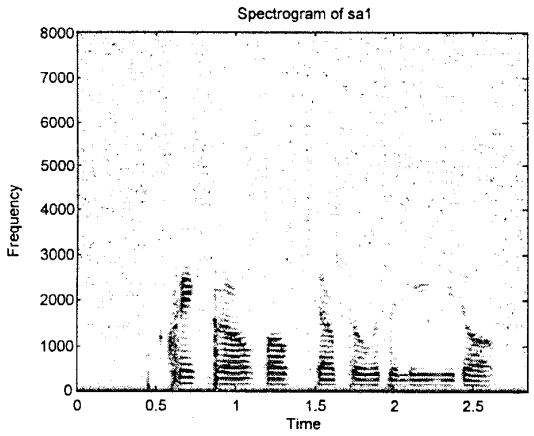

Fig. 6. Spectrogram of speaker B

IV. PATTERN MATCHING USING NEURAL NETWORKS

Artificial neural networks (ANNs) are intelligent systems that are related in some way to a simplified biological model of the human brain. They are composed of many simple elements, called neurons, operating in parallel and connected to each other in the forward path by some multipliers called the connection weights. Neural networks are trained by adjusting values of these connection weights between the network elements. Neural networks have self learning capability, are fault tolerant and noise immune, and have applications in system identification, pattern recognition, classification, speech recognition, image processing, etc. Backpropagation neural networks have used to identify bird species using recordings of birdsong [2].

In our application, ANN is used for pattern matching. The performance of the different neural network architectures were compared for this application.

A. Backpropagation Neural Network

A three layer feedforward neural network with a sigmoidal hidden layer followed by a linear layer is employed in this application for pattern matching. The neural network is trained using the backpropagation algorithm. A momentum term is used in the backpropagation algorithm to achieve a faster global convergence. In this application, an adaptive backpropagation learning method is employed that is the learning gain is adjusted during the training to enhance faster and global convergence. The three layer feedforward neural network architecture for this application is shown in Fig. 7. A bias value of 1 is used to enable each neuron to fire 100 %.

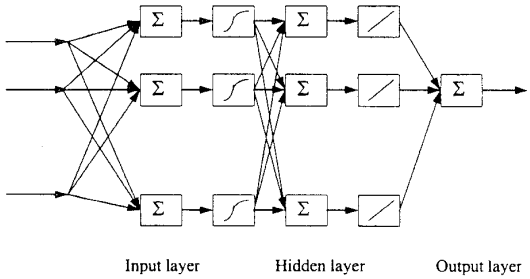

Fig. 7. Feedforward neural network

The ANNs were trained with seven voice samples recorded at different instants of time of four different speakers uttering the same phrase at all times. An initial learning rate, an allowable error and the maximum number of training cycles are the parameters that specified during the training phase. The neural network is constructed in the MATLAB environment [3]. In Fig. 8 the first diagram shows a plot of the sum squared error versus the number of epochs during the training phase. The sum squared error goal was reached in just 174 epochs. The second diagram of Fig. 8 shows the different learning rates used during the training.

A success rate of 100% was achieved when the ANN was tested with trained samples. However when untrained samples were used, only a 66% success rate was possible. This is due to the PSDs of the input samples not being consistent with that of the training samples. The ANN was tested with voice samples of people not known to it and it successful classified these samples as unidentified voices.

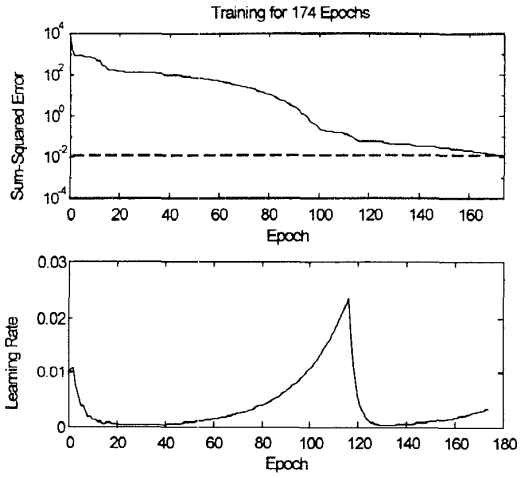

Fig. 8. Backpropagation training error curve and learning gain curve.

B. Self Organizing Maps

Self Organizing Maps (SOMs) are known to be better than other neural networks when it comes to autonomous data classification and these are widely used in condition monitoring and fault detection. The SOMs have different structures to sigmoidal feedforward neural networks.

The hidden layer neurons are arranged in a grid pattern and there is no output layer in this type of ANNs. Input vectors activate hidden neurons according to how similar their weights are to the input values. SOM training adjusts and rearranges these weights so that similar inputs activate the same or nearby neurons. Thus large amounts of data can be organized into unlabelled categories corresponding to various activation clusters on the hidden layer grid map. The organizing mechanism of Kohonen learning is used in training SOMs.

The SOMs were studied for our application of voice recognition. The training of SOMs was done with seven different voice samples of three different speakers uttering the same phrase at all times. A success rate of 89% was attained when tested with untrained samples. As number of speakers were increased to four, the success rate dropped to 78%.

The number speakers were increased further and a comparative study proved that the self organizing neural network has a better recognition ability compared to the backpropagation neural network.

V. HARDWARE

The work that has being reported in this paper was carried out on a Pentium 133 MHz computer with a 16 bit sound card. All the digital signal processing and neural network implementations were carried using the MATLAB signal processing and neural network toolboxes respectively.

This work will be extended to the implementation of a real time voice recognition system using neural networks on a digital signal processor once an optimized neural network and a better feature extraction method is arrived at.

VI. CONCLUSIONS

The use of artificial neural networks in voice recognition in our work has so far proved a fair amount of success especially with the self organizing neural networks. Studies in voice recognition using neural networks based on utterance containing certain phonemes will be investigated further [4]. With voice recognition systems employed as substitutes for conventional keys, access codes, access cards, etc., crime and illegal immigration problem can be effectively reduced.

REFERENCES

[1] P Krauss, L Shure, J N Little, "MATLAB Signal Processing Toolbox User's Guide", The Mathworks Inc., 1996.

[2] A L Mcilraith, H C Card, "Birdsong Recognition Using Backpropagation and Multivariate Statistics", *IEEE Trans on Signal Processing*, vol. 45, no. 11, November 1997.

[3] H Demuth, M Beale, "MATLAB Neural Network Toolbox User's Guide", The Maths Works Inc., 1996.

[4] R L Kashyap, "Speaker Recognition from a Unknown Utterance and Speaker-Speech Interaction", *IEEE Trans on Acoustics, Speech and Signal Processing*, vol. assp-24, no. 6, pp. 481-488, December 1976.