

# COMPARISON OF TEXT-DEPENDENT SPEAKER IDENTIFICATION METHODS FOR SHORT DISTANCE TELEPHONE LINES USING ARTIFICIAL NEURAL NETWORKS

Ganesh K Venayagamoorthy and Narend Sundepersadh  
*gkumar@ieee.org & sundern@telkom.co.za*

Department of Electronic Engineering,  
M L Sultan Technikon,  
P O Box 1334, Durban 4000, South Africa.  
Tel. No. +27 31 3085381 & Fax No. +27 31 3085379

## ABSTRACT

The transition to democracy in South Africa has brought with it certain challenges. The main challenge is to get rid of crime and corruption. This paper presents a technique to combat white-collar crime in telephone transactions by identifying and verifying speakers using Artificial Neural Networks (ANNs). Results are presented to show that speaker identification is feasible and this is illustrated with two different types of ANN architectures and with two different types of characteristic features as inputs to ANNs.

## 1. INTRODUCTION

There is a vital need for speaker identification in all spheres of life. The most important being that this system will enable people to have secure access to information and property. It has significant advantages in electronic banking and internet access. Countless money is lost each year due to white collar crime, fraud, and embezzlement. In today's complex economic times, businesses and individuals are both falling victims to these devastating crimes. Employees embezzle funds or steal goods from their employers, then disappear or hide behind legal issues. Individuals can easily become helpless victims of identity theft, stock schemes and other scams that rob them of their money

White collar crime occurs in the gray area where the criminal law ends and civil law begins. Victims of white collar crimes are faced with navigating a daunting legal maze in order to effect some sort of resolution or recovery. Law enforcement is often too focused on combating "street crime" or does not have the expertise to investigate and prosecute sophisticated fraudulent acts. Even if criminal prosecution is pursued, a criminal conviction does not mean that the victims of fraud are able to recover their losses. They have to rely on the criminal courts awarding restitution after the conviction and by then the perpetrator has disposed of or hidden most of the assets available for recovery. From the civil law perspective, resolution and recovery can just be as difficult as pursuing criminal prosecution. Perpetrators of white collar crime are often difficult to locate and served with civil process. Once the perpetrators have been located and served, proof must be provided that the fraudulent act occurred and recovery/damages are needed. This usually takes a lengthy legal fight, which often can cost the victim more money than the fraud itself. If a judgement is awarded, then the task of collecting is made difficult by the span of time passed and the perpetrator's efforts to hide the assets. Often after a long legal battle, the victims are left with a worthless judgement and no recovery.

One solution to avoid white collar crimes and shorten the lengthy time in locating and serving perpetrators with a judgement is by the use of biometrics techniques for identifying and verifying individuals. Biometrics are methods for recognizing a user based on his/her unique physiological and/or behavioural characteristics. These characteristics include fingerprints, speech, face, retina, iris, hand-written signature, hand geometry, wrist veins, etc. Biometric systems are being commercially developed for a number of financial and security applications.

Many people today have access to their company's information systems by logging in from home. Also, internet services and telephone banking are widely used by the corporate and private sectors. Therefore to protect one's

resources or information with a simple password is not reliable and secure in the world of today. The conventional methods of using keys, access passwords and access cards are being easily overcome by people with criminal intention.

Voice signals as a unique behavioral characteristics is proposed in this paper for speaker identification and verification over short distance telephone lines using artificial neural networks. This will address the white collar crimes over the telephone lines. Speaker identification [1] and verification [2] over telephone lines have been reported but not using artificial neural networks.

Artificial neural networks are intelligent systems that are related in some way to a simplified biological model of the human brain. Attenuation and distortion of voice signals exist over the telephone lines and artificial neural networks, despite a nonlinear, noisy and unstationary environment, are still good at recognizing and verifying unique characteristics of signals. Multi-layer perceptron (MLP) feedforward neural networks trained with backpropagation algorithm have been applied to identify bird species using recordings of birdsongs [3]. Speaker identification based on direct voice signals using different types of neural networks have been reported [4,5]. The work reported in this paper extends the work reported in [5] to short distance telephone networks using ANN architectures described in section 4 of this paper.

The feature extraction, the neural network architectures and the software and hardware involved in the development of the speaker identification and verification system are described in this paper. Results with success rates up to 90% in speaker identification and verification over short distance telephone lines using artificial neural networks is reported in this paper.

## 2. SPEAKER IDENTIFICATION AND VERIFICATION SYSTEM

A block diagram of a conventional speaker identification/verification system is shown in figure 1. The system is trained to identify a person's voice by each person speaking out a specific utterance into the microphone. The speech signal is digitized and some digital signal processing is carried out to create a template for the voice pattern and this is stored in memory.

The system identifies 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 the parameter extraction and pattern matching. In parameter extraction distinct patterns are obtained from the utterances of each person and used to create a template. In pattern matching, the templates created in the parameter extraction process are compared with those stored in memory. Usually correlation techniques are employed for traditional pattern matching.

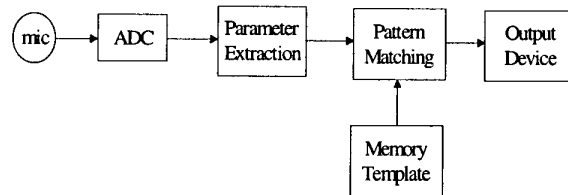


Figure 1: Block Diagram of a Conventional Speaker Identification/Verification System.

The speaker identification/verification system over telephone lines investigated in this paper using artificial neural networks is shown in figure 2.

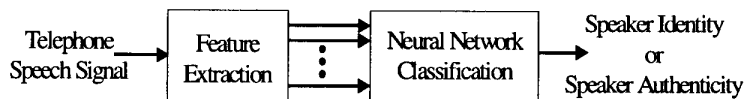


Figure 2: Block Diagram of the Speaker Identification/Verification System using an ANN.

In this paper, the speaker identification/verification system reported is a text-dependent type. The system is trained on a group of people to be identified by each person speaking out the same phrase. The voice is recorded on a standard 16-bit computer sound card from the telephone handset receiver. Although the frequency of the human voice ranges from 0 kHz to 20 kHz, most of the signal content lies in the 0.3 kHz to 4 kHz range. The frequency over the telephone lines is limited to 0.3 kHz to 3.4 kHz and this is the frequency band of interest in this work. Therefore, 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 used to convert these sound files to a presentable form as input vectors to a neural network. The output of the neural network identifies and verifies the speaker in the group.

### 3. FEATURE EXTRACTION METHODS

The process of feature extraction consists of obtaining characteristic parameters of a signal to be used to classify the signal. The extraction of salient features is a key step in solving any pattern recognition problem. For speaker recognition, the features extracted from a speech signal should be consistent with regard to the desired speaker while exhibiting large deviations from the features of an imposter. The selection of speaker-unique features from a speech signal is an ongoing issue. Findings report that certain features yield better performance for some applications than do other features. Ref. [5] have shown on how the performance can be improved by combining different types of features as inputs to an ANN classifier.

Speaker identification and verification over telephone network presents the following challenges:

- a) Variations in handset microphones which result in severe mismatches between speech data gathered from these microphones.
- b) Signal distortions due to the telephone channel.
- c) Inadequate control over speaker/speaking conditions.

Consequently, speaker identification and verification systems have not yet reached acceptable levels of performance over the telephone network. Several feature extraction techniques are explored but only the Power Spectral Densities (PSDs) and Linear Prediction Coding (LPC) techniques are reported in this paper. The discrete Fourier transform of the telephone voice samples is obtained and the PSDs are computed. The PSDs of two different speakers A and uttering the same phrase is shown in figures 3 and 4 respectively.

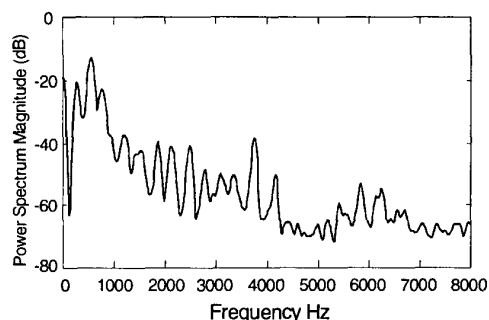


Figure 3: PSD of Speaker A

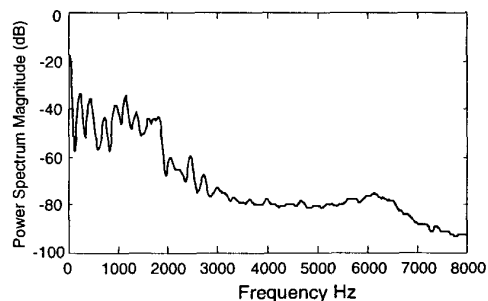


Figure 4: PSD of Speaker B

It can be seen from these figures that the PSDs of the speakers differ from each other. Ref. [5] has reported success on speaker identification up to 66% and 90% with PSDs as input vectors to multilayer feedforward neural networks and Self-Organizing Maps (SOMs) respectively.

Figures 5 below represents the LPC plots of two speakers, A and B. This diagrams show a marked difference in the feature characteristics obtained for speaker A and B unlike that of figures 3 and 4. It is reported in section 6 of this paper that LPCs as inputs to ANNs yield a higher success rate in speaker identification.

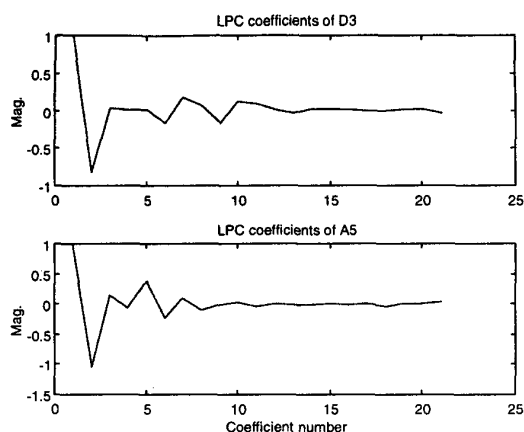


Figure 5: LPC of two different Speakers

Linear prediction models a given signal  $x(nT)$  as the impulse response of an all-pole filter. It assumes that each output sample of a signal,  $x(k)$ , is a linear combination of the past  $n$  outputs (that is, it can be “linearly predicted” from these outputs), and that the coefficients are constant from sample to sample.

#### 4. PATTERN MATCHING USING ARTIFICIAL 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 by some multipliers called the connection weights or strengths. Neural networks are trained by adjusting values of these connection weights between the neurons.

Neural networks have a self learning capability, are fault tolerant and noise immune, and have applications in system identification, pattern recognition, classification, speech recognition, image processing, etc. In this paper, ANNs are used for pattern matching. The performance of different neural network architectures are investigated for this application. This paper presents results for the MLP feedforward network and the self-organizing feature map. Descriptions of these networks are given below.

##### 4.1. MLP Feedforward Network

A three layer feedforward neural network with a sigmoidal hidden layer followed by a linear output layer is used in this application for pattern matching. The neural network is trained using the conventional backpropagation algorithm. In this application, an adaptive learning rate is used; that is, the learning rate is adjusted during the training to enhance faster global convergence. Also, a momentum term is used in the backpropagation algorithm to achieve a faster global convergence.

The MLP network in figure 6 is constructed in the MATLAB environment [6]. The input to the MLP network is vector containing the PSDs. The hidden layer consists of thirty neurons for four speakers. The number of neurons in the output layer depends on the number of speakers and in this paper it is seven.

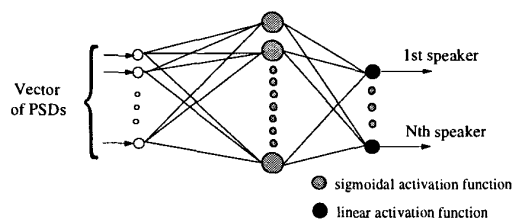


Figure 6: MLP Network

An initial learning rate, an allowable error and the maximum number of training cycles/epochs are the parameter that are specified during the training phase to the MATLAB neural network program.

#### 4.2. Self-Organizing Feature Maps

The second type of neural network selected for this investigation is the self-organizing feature map [7]. This neural network is selected because of its ability to learn a topological mapping of an input data space into a pattern space that defines discrimination or decision surfaces. The operation of this network resembles the classical vector-quantization method called the k-means clustering. Self-organizing feature maps are more general because topologically close nodes are sensitive to inputs that are physically similar. Output nodes will be ordered in a natural manner.

Typically, the Kohonen feature map consists of a two dimensional array of linear neurons. During the training phase the same pattern is presented to the inputs of each neuron, the neuron with the greatest output value is selected as the winner, and its weights are updated according to the following rule:

$$w_i(t+1) = w_i(t) + \alpha[x(t) - w_i(t)] \quad (1)$$

where  $w_i(t)$  is the weight vector of neuron  $i$  at time  $t$ ,  $\alpha$  is the learning rate and  $x(t)$  is the training vector.

Those neurons within a given distance, the neighborhood, of the winning neuron also have their weights adjusted according to the same rule. This procedure is repeated for each pattern in the training set to complete a training cycle or an epoch. The size of the neighborhood is reduced as the training progresses. In this way the network generates over many cycles an ordered map of the input space, neurons tending to cluster together where input vectors are clustered, similar input patterns tending to excite neurons in similar areas of the network.

### 5. IMPLEMENTATION OF THE SPEAKER IDENTIFICATION AND VERIFICATION SYSTEM

The work that is being reported in this paper is implemented in software. The telephone speech is captured and processed on a Pentium II 233 MHz computer with a 16 bit sound card. The telephone receiver is interfaced to the sound card. Telephone speech is captured over signals transmitted within 10 kilometres of transmission network. Digital signal processing and neural network implementations are carried out using the MATLAB signal processing and neural network toolboxes respectively. This work is currently undergoing and an implementation of a real-time speaker identification and verification system over telephone lines on a digital signal processor is envisaged.

### 6. EXPERIMENTAL RESULTS

The MLP network is trained with the PSDs of eight voice samples recorded at different instants of time under controlled and uncontrolled speaking conditions of seven different speakers uttering the same phrase at all times. Controlled speaking conditions refer to noise and distortion free conditions unlike uncontrolled speaking conditions which have noise and distortion on the transmission lines. The number of PSD points for each voice sample is about 500. As mentioned in section 4.1, an adaptive learning rate is used for the MLP network. The initial learning rate is 0.01. The allowable sum squared error and maximum number of epochs specified to the MATLAB neural network program is 0.01 and 10000 respectively. It is found that the sum squared error goal is reached within 1000 epochs.

A success rate of 100% is achieved when the trained MLP network is tested with the same samples used in the training phase. However, when untrained samples are used, only a 63% success rate is obtained. This is due to the inconsistency in the PSDs of the input samples with those used in the training phase. The MLP network is also tested with unseen voice samples of people who are not included in the training set and the network successfully classified these voice samples as unidentified.

Seven speakers are identified using the self-organizing feature map like in the case of the MLP network. An initial learning rate of 0.01, an allowable sum squared error of 0.01 and a maximum of 15000 epochs are specified at the start of the training process to the MATLAB neural network program. The results with the self-organizing feature

map shows a drastic change in the success rate in identifying the speakers as reported in [5]. With PSDs as inputs, a success rate of 85% and 90% is achieved under uncontrolled and controlled speaking conditions respectively.

With LPCs as inputs to SOMs the success rate increased to 98% under uncontrolled speaking conditions. The drawbacks with PSDs as inputs to neural networks are that it is computationally intensive with MLPs and it takes a lot of time to train the SOMs.

## 7. CONCLUSIONS

This paper has reported on the feasibility of using neural networks for speaker identification and verification over short distance telephone lines and has shown that performance with the self-organizing map is higher compared to that with the multilayer feedforward neural network. It also shows that the LPCs as inputs yield better results than with PSD as inputs. This paper has shown that speaker identification is possible over the telephone lines and therefore telephonic bank and other transactions can be authenticated. Hence a technique to combat and/or reduce white collar crime.

## 8. REFERENCES

- [1] Reynolds DA, "Large population speaker identification using clean and telephone speech", *IEEE Signal Processing Letters*, vol. 2 no. 3 March 1995, pp. 46 - 48.
- [2] Naik JM, Netsch LP, Doddington GR, "Speaker verification over long distance telephone lines", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 23-26 May 1989, pp. 524 - 527.
- [3] Mcilraith AL, Card HC, "Birdsong Recognition Using Backpropagation and Multivariate Statistics", *Proceedings of IEEE Transactions on Signal Processing*, vol. 45, no. 11, November 1997.
- [4] Venayagamoorthy GK, Moonasar V, Sandrasegaran K, "Voice Recognition Using Neural Networks", *Proceedings of IEEE South African Symposium on Communications and Signal Processing (COMSIG 98)*, 7-8 September 1998, pp. 29 - 32.
- [5] Moonasar V, Venayagamoorthy GK, "Speaker identification using a combination of different parameters as feature inputs to an artificial neural network classifier", *Proceedings of IEEE Africon 99 conference*, Cape Town, 29 September - 2 October 99, pp. 189 - 194.
- [6] Demuth H, Beale M, *MATLAB Neural Network Toolbox User's Guide*, *The Maths Works Inc.*, 1996.
- [7] Kohonen T, *Self-organizing and associate memory*, Springer Verlag, Berlin, third edition, 1989.