01 Jan 1999

Speaker Identification Using a Combination of Different Parameters as Feature Inputs to an Artificial Neural Network Classifier

Viresh Moonasar

Ganesh K. Venayagamoorthy
Missouri University of Science and Technology

Recommended Citation

V. Moonasar and G. K. Venayagamoorthy, "Speaker Identification Using a Combination of Different Parameters as Feature Inputs to an Artificial Neural Network Classifier," *Proceedings of IEEE Africon, 1999*, Institute of Electrical and Electronics Engineers (IEEE), Jan 1999.
The definitive version is available at https://doi.org/10.1109/AFRCON.1999.820791

Speaker Identification using a Combination of Different Parameters as Feature Inputs to an Artificial Neural Network Classifier

Viresh Moonasar and Ganesh K Venayagamoorthy, *Member, IEEE*

Electronics Engineering Department, ML Sultan Technikon, Durban, South Africa
moonasv@telkom.co.za *gkumar@saiee.org.co.za*

Abstract: This paper presents a technique using Artificial Neural Networks (ANNs) for speaker identification that results in a better success rate compared to other techniques. The technique used in this paper uses both Power Spectral densities (PSDs) and Linear Prediction Coefficients (LPCs) as feature inputs to a self organizing feature map to achieve a better identification performance. Results for speaker identification with different methods are presented and compared.

1. Introduction

Biometrics are methods for recognizing a user based on his/her unique physiological and/or behavioural characteristics. These characteristics include finger prints, 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. The task performed by the system described in this paper can be classified into identification and verification of people using their voice signals. 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.

Speaker identification 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 phrase) during the training and the identification phases. The identification phase is when the trained ANN is tested with unseen voice samples at different moments in time. A text independent system does not necessarily require the training *text* during the identification phase to identify speakers. Both systems consist of 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. In similarity analysis, a voice sample is compared for similarity with another sample by computing the regression between the coefficients.

The most significant factor affecting automatic speaker recognition performance is variation in the signal characteristics from moment to moment (inter-session variability and variability over time). Variations arise from the different states under which the speakers can be, from the differences in the recordings and the transmission conditions, and from the background noise.

Speakers cannot repeat an utterance precisely the same way at different moments in time. It is well known that samples of the same utterance recorded in one short period are much more highly correlated than samples recorded in different periods of time. Duration of a period can vary from a day to weeks. A long-term period (months to years) can cause major changes in voices. For example, especially with males, in the process of growing up, their voices' change indicating adulthood. The short or long period of time is not only the factor causing variations in the speaker's signal characteristics. The emotional state, under which the speaker is, can also be a major factor causing variations in a speaker's signal characteristic. This paper has taken into account all these factors in speaker identification. It is important for robust speaker recognition systems to accommodate these variations.

Two types of normalization techniques have been tried namely: one in the parameter domain, and the other in the distance/similarity domain. A number of normalization techniques have been developed to account for variation of the speech signals.

The Multi-Layer Perceptron (MLP) network trained using backpropagation algorithm with PSDs as feature inputs for speaker identification have a low success rate during the identification phase [1]. The Self-Organizing Maps (SOMs) compared with the MLP networks showed a higher identification success rate [1]. In this paper, PSDs, LPCs, and both PSDs and LPCs are used as feature inputs to the SOMs and results are presented to show a higher speaker identification success rate with artificial neural networks.

2. Speaker Identification System

A block diagram of a typical speaker identification 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

0-7803-5546-6/99/$10.00 © 1999 IEEE

carried out to create a template for the voice pattern and this is stored in memory.

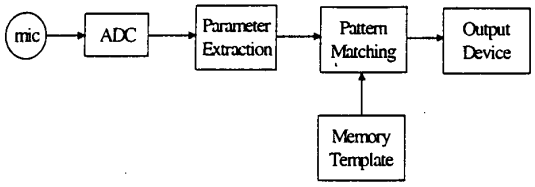

Figure 1: Block Diagram of a Typical Speaker Identification System.

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.

The speaker identification system investigated in this paper using artificial neural networks is shown in figure 2.

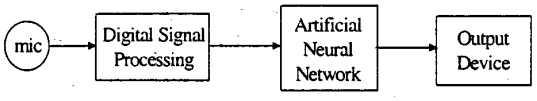

Figure 2: Block Diagram of the Speaker identification System using ANN.

In this paper, the speaker identification system 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 using a directional microphone. 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. 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 the speaker in the group.

3. Characteristic Feature Extraction

Feature extraction plays a very important role in speaker identification. Human speech can be sensibly interpreted using frequency-time interpretations such as spectrograms. Frequency-energy interpretations and power spectral densities can be used to differentiate between speakers. Other methods used for this purpose are the linear predictive coding and cepstral analysis.

The first technique used in the feature extraction in this paper is to obtain the discrete Fourier transform of the voice samples and then compute the PSDs by taking the square magnitude of the Fourier transform. These periodograms are then averaged and scaled. The PSD of a voice sample contains unique features attributed to an individual that are used in the speaker identification. These PSD values are presented in a vector form to the neural network. The PSD of two different speakers are shown in figures 3 and 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 PSDs. These PSDs are computed using the MATLAB signal processing toolbox [2].

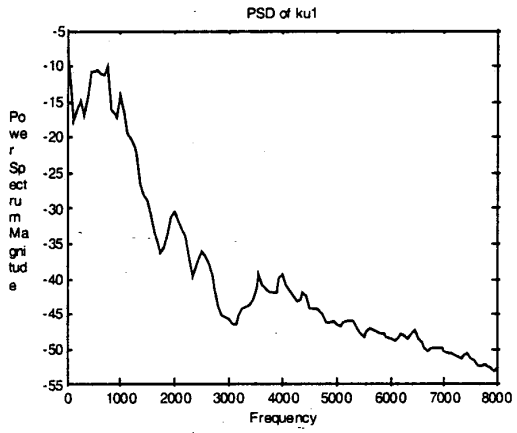

Figure 3: PSD of Speaker A

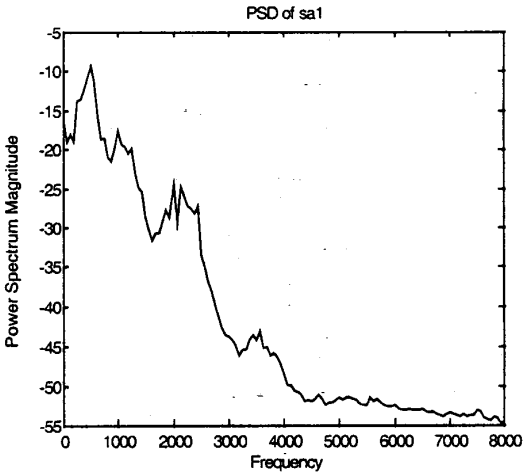

Figure 4: PSD of Speaker B

A second digital signal processing technique to extract unique features for speaker identification is to take the

time-dependent Fourier transform of voice signals. The time-dependent Fourier transform is the windowed discrete-time Fourier transform of a sequence computed using a sliding window. The spectrogram of the sequence is then the magnitude of the time-dependent Fourier transform versus time.

The spectrograms of the same two speakers, A and B, of figures 3 and 4 are shown in figures 5 and 6 respectively. Spectrograms of speaker A consists of more peaks than that of speaker B. The spectrograms give a clear distinction between speaker A and speaker B than just taking the PSDs as shown in figures 3 and 4.

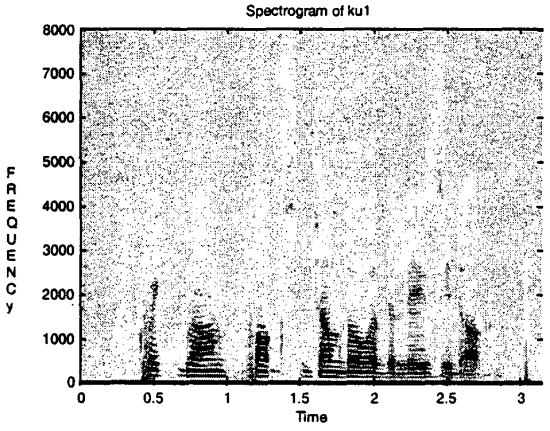

Figure 5: Spectrogram of Speaker A

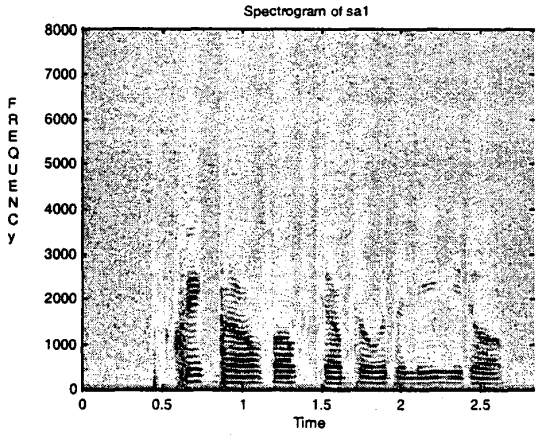

Figure 6: Spectrogram of Speaker B

A third technique for feature extraction is based on Linear Prediction Coefficients (LPCs). A number of parametric descriptions of speech based on linear predictive coding [3] have been used in an attempt to reduce the sensitivity to the inevitable recall error. Line Spectral Pair representation (LSP) [4] has been found to produce the best results, but large training times are required to sufficiently reduce the error on the training set, and generalization remains poor [5].

Modern speech devices only use linear terms, hence the expression 'linear prediction', a special case of Kolmogorov's polynomial:

$$s(n) = \sum_{i=1}^{M} a_i s(n-i) \qquad (1)$$

where s(n) is the predicted value of the nth sample based on the previous M samples. The a_i are the constant coefficients of a filter to be determined.

In linear prediction representation, the speech signal is modeled as the output of an all pole filter *H(z)*, that is, excited by a sequence of pulses separated by the pitch period for voiced sounds, or pseudo random noise for unvoiced sounds [6]. The filter models the combined modulation properties of the glottal source and the vocal tract.

The vector space described by the LPCs obtained from the analysis of human speech is not populated with uniform density, but instead vectors tend to be concentrated in clusters. Vector quantization [7] is a technique widely used in low bit rate speech coding by which any vector within a cluster is replaced by a reference vector representing the centroid of the cluster (in practice large clusters may be assigned with more than one reference vector). A codebook containing the centroids of cluster is stored in memory. During pattern matching, each vector is compared with the codebook to find the most similar reference vector; the index of this codebook entry is then the identity of the speaker.

The first 20 LPC's from two separate recordings at time T_1 and T_2 of a speaker, C, are shown in figures 7 and 8. There is consistency between the two samples recorded at different times.

In this paper, results using the PSD, LPC and a combination of PSD and LPC as feature inputs (described above) to an ANN are presented.

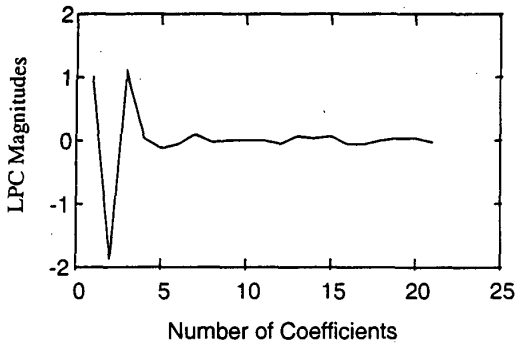

Figure 7: LPCs' of Speaker C at Time T_1

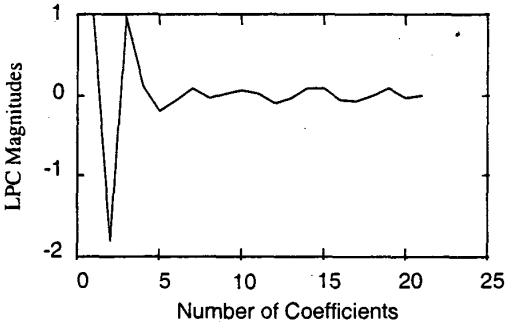

Figure 8: LPCs' of Speaker C at Time T_2

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. MLP feedforward neural networks trained with the backpropagation algorithm have been applied to identify bird species using recordings of birdsong [8].

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 9 is constructed in the MATLAB environment [9]. The inputs to the MLP network is a vector containing the PSDs. The hidden layer consists of thirty neurons for five speakers. The number of neurons in the output layer depends on the number of speakers and in this paper, it is five.

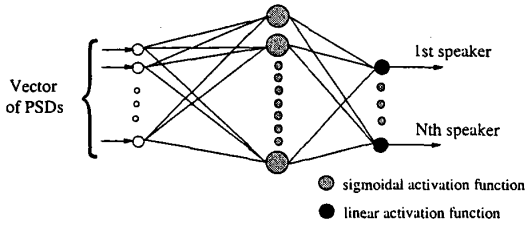

Figure 9: MLP Network

An initial learning rate, an allowable error and the maximum number of training cycles/epochs are the parameters that 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 [10]. 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. This process has been used by Kohonen in a system for phonetic recognition of Finnish and Japanese [11].

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 training phase the same pattern is presented to the inputs of each neuron, the neuron with the greatest output value

0-7803-5546-6/99/$10.00 © 1999 IEEE

is selected as the winner, and its weights updated according to the following rule:

$$w_i(t+1) = w_i(t) + \alpha[x(t) - w_i(t)] \qquad (2)$$

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

Those neurons within a given distance, the neighborhood, of 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 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 a similar areas of the network.

When trained, the weight matrix of the network forms a codebook of the vector space described by the training set, which is then used for vector quantization.

5. Implementation of the Speaker Identification System

The work that is being reported in this paper is implemented in software. The data is captured and processed on a Pentium 133 MHz computer with a 16 bit sound card. 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 system using a digital signal processor is envisaged.

6. Results with the MLP network and the Self-Organizing Feature Maps

The MLP network is trained with PSDs of seven voice samples recorded at different instants of time of five different speakers uttering the same phrase at all times. The number of PSDs points for each voice sample is about 500. As mentioned in the section 4.1, for the MLP network an adaptive learning rate is used. The initial learning rate is 0.01. Figure 10 shows a plot of learning rate versus the number of training epochs.

The allowable sum squared error and maximum number of epochs specified to the MATLAB neural network program is 0.01 and 10000 respectively. Figure 11 shows that in 174 epochs the sum squared error goal is reached.

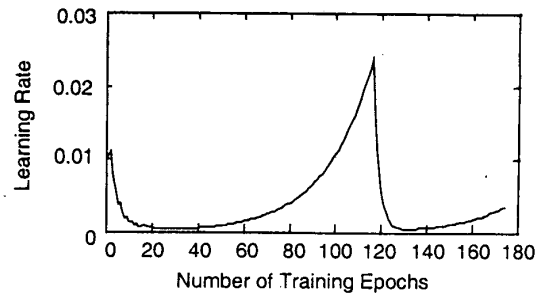

Figure 10: Learning Rate versus Number of Training Epochs

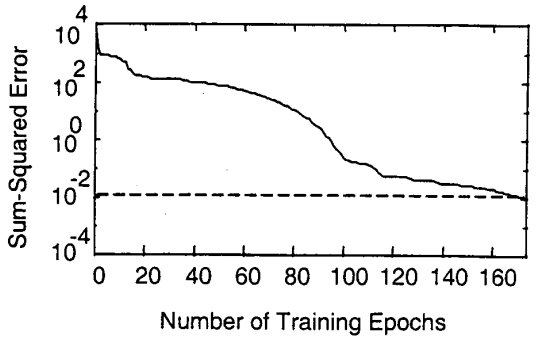

Figure 11: Sum Squared Error versus Number of Training Epochs

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 66% success rate is obtained. This is due to the inconsistency in PSDs of the input samples with those samples 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 successful classified these voice samples as unidentified.

With the self-organizing feature maps, three different feature inputs are used namely: PSDs, LPCs, and both PSDs and LPCs. Five speakers are identified using the self-organizing feature 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 10000 epochs are specified at the start of the training process to the MATLAB neural network program. The number of inputs are 500 and 20 with the PSDs and LPCs methods respectively.

The results with self-organizing feature map shows a drastic change in the success rate in identifying the speakers. However, this is at the expense of more training time or training epochs. With PSDs as inputs, the success rate is 90%, and with LPCs as inputs, the success rate further improved to 98%. The complexity of the computations is reduced drastically with the

LPCs. LPC's are better feature extraction parameters than the PSD's. With a combination of PSDs and LPCs, the success rate further improved by 1 % at the expense of more computations.

Figure 12 shows a comparison of the different methods reported in this paper for speaker identification.

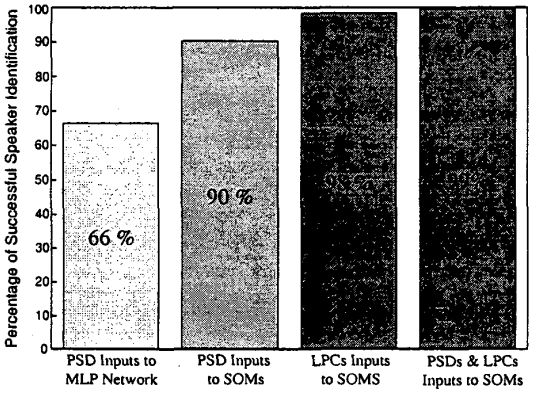

Figure 12: A Comparison of the Success Rate with the Different Methods

6. Conclusions

This paper examined different methods for speaker identification using artificial neural networks as classifiers. The MLP feedforward neural network and the self-organizing feature map with vector quantization shows different success rates in speaker identification. The self-organizing feature maps are excellent pattern classifiers compared to the MLP networks and therefore a relatively high success rate.

Cepstrums and spectrograms as feature inputs to an ANN classifier for speaker identification are currently being investigated. There is no one fixed, robust characteristic feature that can be extracted from a voice sample that can uniquely identify a speaker, and provide a 100% success rate. This paper has presented results with a method that uses a combination of different parameters as features inputs to a neural network and shown improved performance to the level of achieving 99% success in speaker identification. A real-time speaker identification system using an artificial neural network implementation on a digital signal processor is envisaged soon.

7. References

[1] Venayagamoorthy GK, Moonasar V, "Voice Recognition Using Neural Networks", *Proceedings of IEEE South African Symposium on Communications and Signal Processing (COMSIG 98)*, pp. 29-32, September 1998.

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

[3] Rabiner LR, Schafer RW, "Digital processing of speech signals", Chapter 8, *Prentice Hall*, Eaglewood Cliffs, New Jersey, USA, 1978.

[4] Sugamura N, Itakura F, "Speech analysis and synthesis methods developed at ECL in NTT – from LPC to LSP", *Speech Communication 5*, 1986, pp 199-215.

[5] Cawley GC, Noakes PD, "LSP speech synthesis backpropagation network", *IEE-ANN-93*, May 1993.

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

[7] Gray R, "Vector quantization", *IEEE ASAP Magazine*, April 1984, pp 4-29.

[8] Mcilraith AL, Card HC, "Birdsong Recognition Using Backpropagation and Multivariate Statistics", *Proceedings of* IEEE *Trans on Signal Processing*, vol. 45, no. 11, November 1997.

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

[10] Kohonen T, "Self-organizing and associate memory", *Spring Verlag*, Berlin, third edition, 1989.

[11] Kohonen T, "The neural phonetic typewriter, *IEEE Computer Magazine*, Vol 21, 1988, pp 11-22.