Neural Networks based Classification for Speaker Identification System

Prof. Dr. K. Soliman

Prof. Dr. M. Alkasasy

Eng. R. Orban

Dept. of Computers and Control Systems.
Faculty of Engineering, Mansoura University
2001

Abstract

Speaker identification is the process of automatically recognizing the unknown speaker by extracting the speaker specific information included in her/his speech wave [11]. This paper presents speaker identification system using neural network techniques similar to that reported in [8] but, with different type of neural networks. The results have shown that using a feed-forward neural network in classification stage has improved the percentage of correct classification. It reaches 97.5% compared to 96% correct classification obtained in [8].

Keywords:
Speaker identification, neural network.

1. Introduction

The Problem of resolving the identity of a person can be categorized into two fundamentally distinct types of problems with different inherent complexities: (i) verification and (ii) identification. Verification (authentication) refers to the problem of confirming or denying a person's claimed identity (Am I who I claim to be?). Identification (Who am I?) refers to the problem of establishing a subjects identity. A reliable personal identification is critical in many daily transactions. For example, access control to physical facilities and computer privileges are be-coming increasingly important to prevent their abuse.

There is an increasing interest in inexpensive and reliable personal identification in many emerging civilian, commercial, and financial applications.

Typically, a person could be identified based on (i) a person’s possession (“something that you possess”), e.g., permit physical access to a building to all persons whose identity could be authenticated by possession of a key; (ii) a person’s knowledge of a piece of information (“something that you know”), e.g., permit login access to a system to a person who knows the user-id and a password associated with it. Another approach to positive identification is based on identifying physical characteristics of the person. The characteristics could be either a person’s physiological traits, e.g., voice and signature or his physiological traits, e.g., fingerprints, hand geometry, etc. This method of identification of a person based on her/his distinctive physiological / behavioral characteristics (see Fig. 1) is called biometrics. Since the biological characteristics can not be forgotten (like passwords) and can not be easily shared or misplaced (like keys), they are generally considered to be a more reliable approach to solving the personal identification problem [2].

Although biometrics can not be used to establish an absolute "yes/no" personal identification like some of the traditional techniques, it can be used to achieve a "positive identification" with very high level of confidence. Recently, biometrics technology has achieved a great deal of attention. It is claimed to be the ultimate technology for automatic personal identification [3]

Voice identification is considered as one of the most recently important biometric identification methods. Most automatic speaker identification systems [ASIS] have the basic structure shown in Figure 1

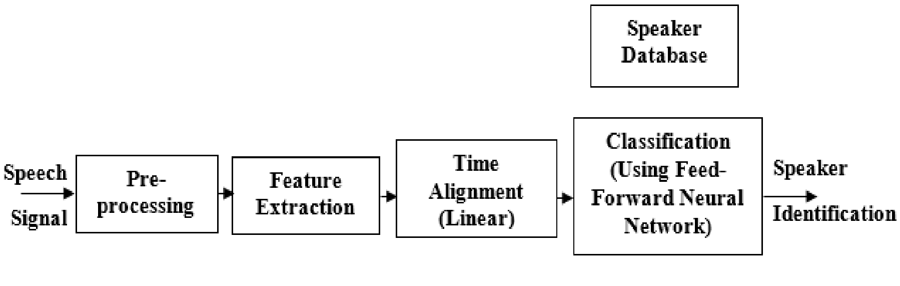

Figure 1: Block diagram of the identification algorithm

Firstly, this paper presents the pre-processing stage (sec. 2) using quantization followed by spectral preemphasis, framing and windowing. Then, the feature extraction stage (sec. 3) is done using two techniques. These are time domain analysis and frequency domain analysis. Then, the time aliqnment stage (sec.4) is done using linear time alignment algorithm. Finally, the classification stage (sec. 5) is presented using feed-forward neural networks and then, the results were compared with those obtained in [8] with different type of neural network.

2. Preprocessing module

This stage consists of the following parts:

2.1. Quantization

In its original meaning, quantization is the step from a continuous to discrete variable, like in its original meaning, quantization is the step of passing from a continuous to a discrete variable, like in analogue-to-digital signal conversion. More generally, the term can be used for any method decreasing the precision of representation by eliminating part of the information [9].

2.2.Spectral Preemphasis:

Preemphasis is used to spectral flatten the speech signal to reduce the computational instability associatrd with finit precision arithmetic [12]. If S(n) is the speech signal, then the spectrally flattened signal SP(n) is given by

$$SP(n)=S(n)-A.S(n-1) \quad (1)$$

Where A is a pre-emphasis coefficient lying usually in an interval of (0.95 to 0.99).

2.3.Framing

The speech is non-stationary process over time as it is generated by time varying movements of the articulators and vector tract. To extract feature vectors. The speech signal is segmented into small frames that can be assumed to be stationary. Consecutive frames are overlapped to provide smoothing [7], in the present work the speech is segmented into 60 ms lengths with 50% overlapping, see Figure 2.

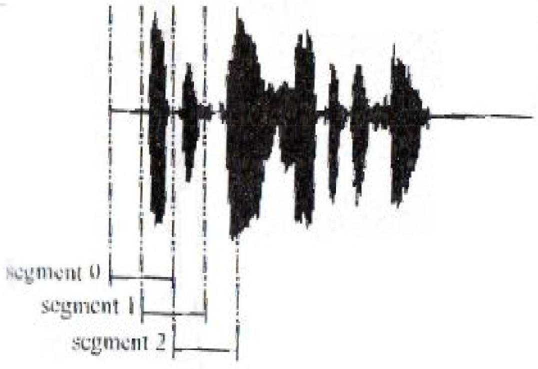

Figure 2: Framing with 50% overlapping

2.4. Windowing

In order to minimize the adverse effect of chopping samples section out of the running speech signal, a smoothing window W(n) is used [7]. A typical smoothing window is the hamming window [1] defined as:

$$W(n)=0.54 - 0.46 . \cos((2 * pi * n)/(N - 1)) \quad \mathbf{(2)}$$

3. Feature Measurement Module

Feature extraction is done using two analysis:

3.1. Time Domain Analysis

In this subsection, there is a set of useful features that are turned as timer domain features.

3.1.1. Short-Time Average Magnitude:

It was observed that the amplitude of the speech signal varies appreciably with time. In particular, the amplitude of unvoiced segments is generally much lower than that of voiced segments. The short-time energy of the speech signal provides a convenient representation that reflects these amplitude variations [7]. The AM is generally defined as:

$$AM = \frac{1}{n} sigma|x(n)| \quad \mathbf{(3)}$$

$$AM = \frac{1}{N}\sum_{m=1}^{N} |x(m)|$$

Where x(m) is the mth speech samples, and N is the number of speech samples per frame.

3.1.2. Zero Crossing Rate

The rate, which zero crossing occur, is a simple measure of the frequency content of a signal (specially narrow-band signals) [6]. Ito and Donaldson summarize some of the previous trials that used ZCRin speech analysis. The ZCR is generally defined as:

$$\boldsymbol{ZCR = \frac{1}{2.N}\sum_{m=1}^{N} |sgn[x(m)] - sgn[x(m-1)]} \quad \textbf{(4)}$$

Where

$$\boldsymbol{sgn[x]} \quad \begin{matrix} =1 & x \geq \\ =-1 & x<o \end{matrix}$$

x(m) is the mth speech samples, and *N* is the number of speech samples per frame.

3.2.Frequency Domain Analysis:

Most useful parameters in speech processing are found in the frequency domain representation of the signal. The vocal tract produces signals that are more consistently and easily analyzed spectrally than time domain. Most of the speech analysis algorithms have been done in the frequency domain such as linear prediction coefficient analysis.

3.2.1. Linear Predictive Coding:

Linear predictive coding (LPC) provides an alternative method to processing speech by calculating spectral energy peaks. LPC uses a linear combination of the previous P data to predict the value of current sample [7]. That is

$$\boldsymbol{x(n) = \sum_{i=1}^{P} a_i . x(n-i) + e(n)} \quad \textbf{(5)}$$

Where

P is the order of the predictor.

e(n) is the prediction error in the nth speech sample.

$[a_1,a_2,\dots,a_p]$ are the prediction coefficients.

In the present work 6 Linear Predictive coefficients are used.

4. Time Alignment Techniques:

Two of the major problems in speaker identification systems have been due to the fluctuations in the speech pattern time axis and spectral pattern variation. Speech is greatly affected by differences in the speaker such as age and sex as well their physical and psychological condition.

The length of the input pattern to the neural network in question is constrained by the number of input neurons to the neural network since this type of network architecture cannot be varied once it's trained. The input pattern vectors must be modified to fit the neural network while still retraining all their discriminating features.

Several techniques have been proposed for determining the alignment path. Including Linear Time Alignment. Time event matching, correlation maximization, and Dynamic Time Warping. This paper applies Dynamic Time Warping technique [4].

The purpose of Dynamic Time Warping is to compute a non-linear mapping of one signal onto another by minimizing the distances between the two [6]. See Figure 3.

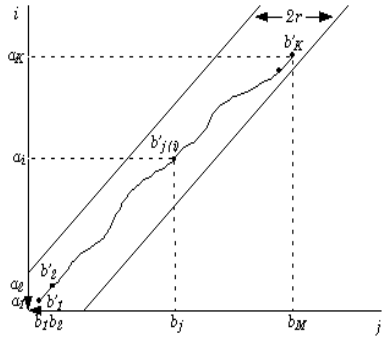

Figure 3: Dynamic time warping between two signals; A and B

The Dynamic Time Warping algorithm consists of the following steps:

Assume *A(i)* where *i*=1,2,3,..N sample and *B(j)* where *j*=1,2,3,…,M sample are the reference and input signals respectively.

1. Constructing the Local Distance Matrix (LDM). The value in LDM would be LMD(I,j) where

$$LDM(j,i)=|B(j)-A(i)| \qquad (6)$$

Where i=1,2,3,.., N and j=1,2,3,…, M

2. Constructing the Accumulated Distance Matrix (*ADM*). The values in *ADM* would be *ADM(j,i)* where

$$ADM(1,1)=LDM(1,1) \qquad (7)$$

$$ADM(j,i)=LDM(j,i)+min\{ADM(j,i-1), ADM(j-1,i-1), ADM(j-2,i-1)\} \qquad (8)$$

Where *(j,i-1)*, *(j-1,i-1)*, and *(j-2,i-1)* are neighboring points of the point at *(j,i)* as defined in. See Figure 4.

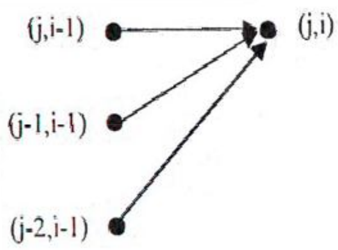

Figure 4: Neighboring points as defined in Itakura method

3. Derive best path, w, by travelling through the cells with the lowest accumulated distances in *ADM*, starting at *w(N)=M* and working back to *w(1)=1. w(i)* will be the indices of input signal to use to shrink/stretch it to reference length.
4. Once the path has been tracked out, the signals can be mapped onto each other in the following way:

$$Time\_warped\_B(1:M)=B((w(i)) \qquad (9)$$

5. Classification

One of the most challenging, powerful, and robust systems introduced in the past few years, are neural networks. The term neural network originally referred to a network of interconnected neurons. The motivation for using the neural networks in so many applications is mainly due to high degree of parallelism associated with them due to their arrangement and structure of neurons.

Neural networks with different architectures have been successfully used in recent years for the identification and control of a wide class of non-linear systems [5, 10].

Using multi-element feed-forward neural networks, a proper choice of the weights, the separating boundary in pattern space can be established to satisfy more combinations of input/output relations and hence, more capacity, see Figure 5.

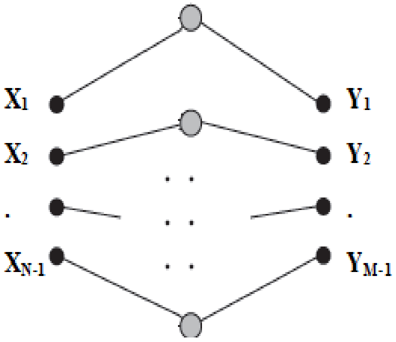

Figure 5: An example for a two-layer feed-forward network with N inputs, M outputs, and a hidden layer

The designed feed-forward neural network has tree layes; an input layer, an output layer, and a hidden layer. The input layer consists of 8 neurons corresponding to the number of features. Instead of using 4 neurons in the output layer for 4 different speakers (the target activations were 0.0 for all output nodes except for a 1.0 on the node representing the given class), this paper uses 2 neurons. This can be accomplished using binary numbers (00,01,10,11). The hidden layer is thought to consist of 10 neurons to obtain best result. The initial learning rate was 0.1. See Figure 6.