AUTOMATIC SPEAKER IDENTIFICATION USING NEURAL NETWORKS

التعـرف الآلـي على الأصـوات باسـتخدام الشبكـات العصـبية

Prof. Dr. K.Soliman

Dr. M.S.Al-Kasasy

Eng. Rasha Orban Mahmoud

Department of Computer and Control systems.

Faculty of Engineering, Mansoura University

2001

خلاصـة

هذا البحث يعرض طريقة آلية لتصنيف الأصوات. وقد استخدمت تقنية الشبكات العصبية ذات التغذية الامامية في مرحلة التصنيف. وقد اظهرت النتائج أن استخدام الشبكات العصبية ذات التغذية الأمامية قد زاد من نسبة صحة التصنيف. فقد وصلت صحة التصنيف إلى ٩٧.٥% مقارنة بالنتائج التي تم الحصول عليها في بحث سابق وهي ٩٥.٧٢% [٧].

ABSTRACT

This paper presents speaker identification system using neural network techniques similar to that reported in [7] 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 95.72% correct classification obtained in [7].

Accepted May 7, 2001

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 I am?). identification Who am I?) refers to the problem of establishing a subject's identity. A reliable personal dentification is critical in many daily transactions. For example, access control to physical facilities and computer privileges are becoming 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 hat you possess"), e.g., permit physical access to a building to all persons who identity could be authenticated by possession of a key: (ii) 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 assoeiated with it. Another approach to positive identification is based on identifying physical characteristics of a person. The characteristics could be either a person's behavioral characteristics, e.g., voice and signature or his physiologieal traits, e.g., ingerprints, hand geometry, etc. This method of identifieation of a person based on his/her behavioral/physiological characteristics is called *biometrics*. Since the biological haracteristics 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 he personal identification problem [5].

Although, biometrics ean not be used to establish an absolute "yes/no" personal dentification like some of the traditional technologies, it can be used to achieve a "positive dentification" with a very high level of confidence. Recently, biometrics technology has eceived a great deal of attention. It is claimed to be the ultimate technology for automatic ersonal identification [6].

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

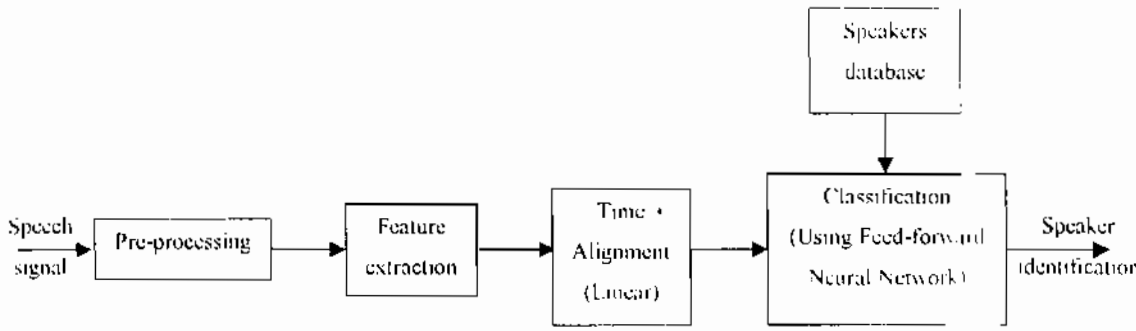

Figure 1: Block diagram of the identification algorithm

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 alignment 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 [7] 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 of passing from a continuous to a discrete variable, like in analogue-to-digital signal conversion. More generally, this term can be used to any method decreasing the precision of representation by eliminating part of the information [2].

2.2 Spectral Preemphasis:

Preemphasis is used to spectral flatten the speech signal to reduce the computational instability associated with finite precision arithmetic [11]. If S(n) is the speech signal, then the spectrally flattened signal SP(n) is given by

$$SP(n) = S(n) - A\ S(n\ \ 1) \qquad (1)$$

where A is the preemphasis coefficient and usually ranges from 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 vocal 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 [9]. In the present work the speech sample is segmented into 60 ms lengths with 50% overlapping. See Figure 2.

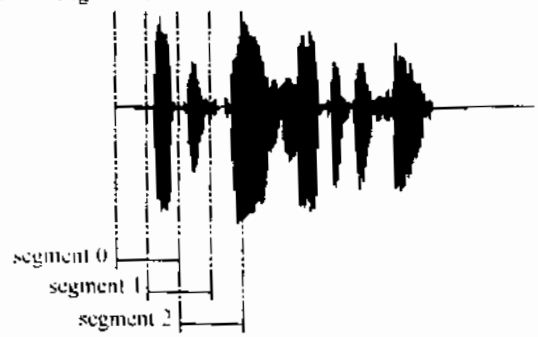

Figure 2: Framing with 50% overlapping

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 [9]. A typical smoothing window is the Hamming window [8] defined as:

$$W(n) = 0.54 - 0.46 \cdot \cos\left(\frac{2\pi n}{N-1}\right) \qquad (2)$$

where N is the number of speech samples per frame.

3 *FEATURE MEASUREMENT MODULE*

Feature extraction is done using two analyses:

3.1 Time domain analysis

In this subsection there is a set of useful features that are turned as time 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 average magnitude AM of the speech signal provides a convenient representation that reflects these amplitude variations [9]. The AM is generally defined as:

$$AM = \frac{1}{N}\sum_{m}^{N}|x(m)| \qquad (3)$$

where x(m) is the m^{th} speech samples, and N is the number nf speech samples per frame.

3.1.2 Zero Crossing Rate:

The rate, which zero crossings occur, is a simple measure of the frequency content of a signal (especially narrow-band signals) [1]. Ito and Donaldson summarize some nf the previous trials that used ZCR in speech analysis. The ZCR is generally defined as:

$$ZCR = \frac{1}{2 \cdot N}\sum_{m=1}^{N}|\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]| \qquad (4)$$

where,

$\mathrm{sgn}[x] = 1 \qquad x \geq 0$

$\quad = -1 \qquad x < 0$

x(m) is the m^{th} speech sample, and N is the nn. of samples per frame.

3.2 Frequency Domain Analysis:

Most useful parameters in speech processing are found in the frequency domain representation of the signals. The vocal tract produces signals that are mnre 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.

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 the current sample [9]. That is

$$x(n) = \sum_{i=1}^{p} a_i \; x(n-i) + e(n) \qquad (5)$$

where,

P is the order of the predictor.

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

$[a_1, a_2 \ldots, a_p]$ are the prediction coeffirients.

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 sexes 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 trained. The input pattern vectors must be modified to fit the neural network while still retaining 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 [10].

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 [1]. See Figure 2.

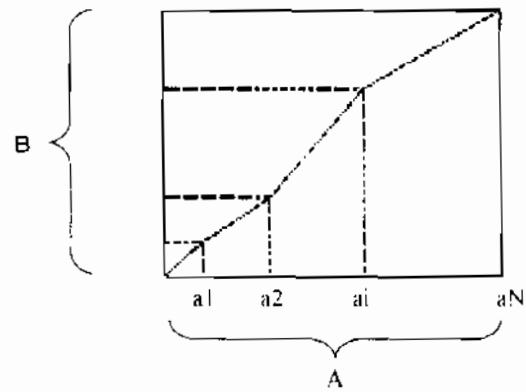

Figure 2 Dynamic time warping between two signals, A and B

The Dynamic time warping algorithm consists of the following steps:

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

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

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

where i=1,2…N and j=1,2…M

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

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

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

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 Itakura method. See Figure 3

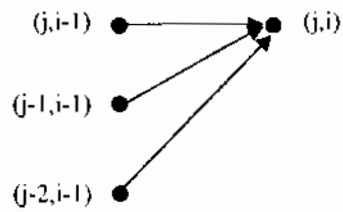

Figure 3 neighboring points as defined in Itakura method

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/streteh it to referenee length.
4. Once the path has been traced out, the signals can be mapped onto each other in the following way:

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

5 CLASSIFICATIONS

One of the most ehallenging, 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 arehitectures have been successfully used in recent years for the identification and control of a wide class of non-linear systems [3,4].

Using multi-element feed-forward neural networks, a proper choice of the weights, the separating boundary in pattern space can be establisbed to satisfy more combinations of input/output relations and hence, more capacity, see Figure 4.

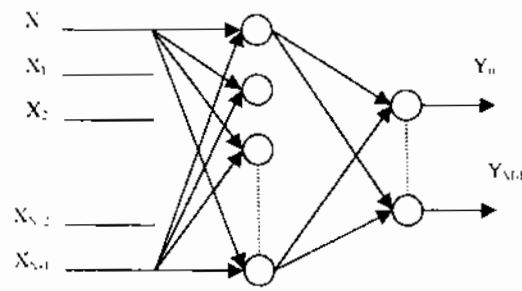

Figure 4. An example for a two-layer feed-forward network with N inputs, M outputs and a hidden layer

The designed feed-forward neural network has three layers; 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, and 11). The hidden layer is thought to consist of 10 neurons to obtain best results. The initial learning rate was 0.1. Figure 5 shows the learning curve of the neural network.

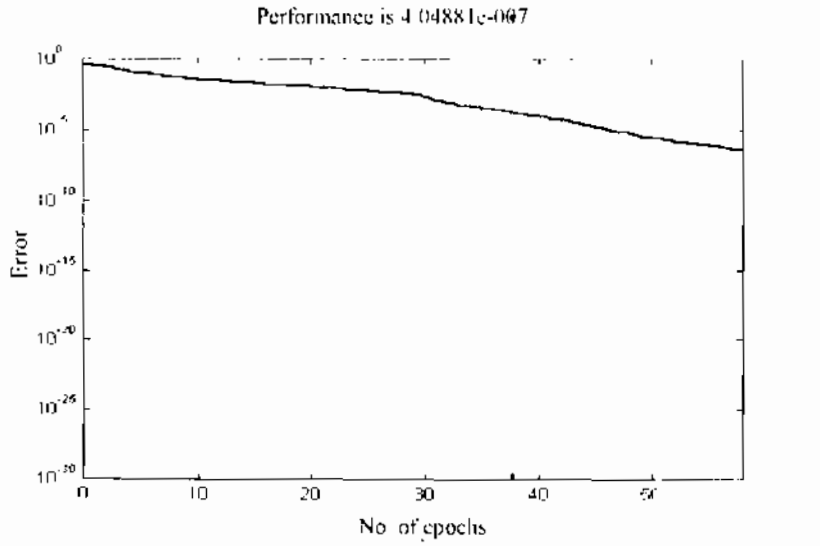

Figure 5 Learning curve of the NN

6 EXPERIMENTAL RESULTS

The speakers' database used in this experiment consists of 240 records for 4 different speakers (2 males, 2 females). Each speaker has 10 different records for the arabic digits

(واحـــد، اثـــنين، ثلاثـــة، أربعـــة، خمـــسة، ستـــة)

Table 1 and Figure 6 shows the percent of correct classification for each speaker. The results obtained show that the percentage of correct classification has been improved. Typical average correct classification accuracy reaches 97.5% compared to 95.72% obtained by [7].

Speaker / Speech	1st	2nd	3rd	4th
1	9	10	10	10
2	10	10	10	10
3	10	10	9	10
4	10	10	10	10
5	10	9	10	10
6	10	9	9	10
Total	59	58	58	60
Percent	98.3333	96.6667	96.6667	100.000

Table 1 Final Results

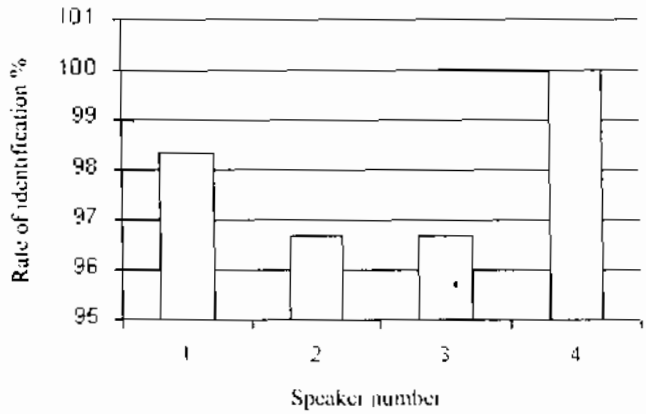

Figure 6 System chart of results

7 CONCLUSION

Automatic speaker identification is one of the most widely and accepted identification techniques nowadays. The proposed system uses feed-forward neural networks in the classification stage so; the correct classification accuracy reaches 97.5%. This result is better correct identification ratio compared to 95.72% obtained in [7]. The technique show promise and the finding can be considered as a guide for future studies.

REFERENCES

[1] Bharani, N., MacAlpino, S., Slavinsky, J., "Speaker Verification", Will Rice College, Rice University, 1999.

[2] Bock, R.K., http://www1.cern.ch/RD11/rkb/AN13pp/node227.html, 5 November 1997.

[3] Haykin, S., "Neural networks, a comprehensive foundation", Prentice Hall, 1999.

[4] Hwang, J.N., Kung S.Y., Niranjan, M., and Principe, J., "The past, present and Future of neural networks for signal processing", IEEE Signal Processing, Vol.14, No. 6, 1997.

[5] Jain, A.K., Biometrie lab., Michigan State University, Online :http://www.cse.msn.eud/, 2000.

[6] Jain, A.K., Hong, L., Parnkanit, S., and Bolle, R., "An Identity-Autheutication System Using Fingerprints", Proceeding of IEEE, Vol. 85 No. 9, pp.1364-1388, Sep. 1997.

[7] Mostafa, W.A., "Application Of Neural Networks On Speaker Recognition", An Msc thesis Electronics and Communications Department, Faculty of Engineering, Cairo University, 1995.

[8] Robinson, A.J., "Speech Analysis", Department Of Engineering, Cambridge University, 1998.

[9] Saber, N.S., "Speech Reeognition Using Neural Network And Wavelet Transform", An Msc thesis, Communication Department, Faculty of Engineering, Alex. University, 1997.

[10] Vignoli, F., and Lavagetto, F., "A Segmented Time-Alignment Technique For Text-Speech Synchronization", University of Genova, 1998.

[11] Yang, W., Benhouchta, M., and Yantorno, R., "Performance Of The Modified Spectral Distortion As An Objective Speech Quality Measure", Temple University, Philadelphia, 1998.