Mansoura Engineering Journal,(MEJ),Vol.20, No.4, December, 1995. E.61

FEATURE SELECTION FOR SPEAKER IDENTIFICATION AND ARABIC DIGITS RECOGNITION

M. I. A. ABDALLA* and A. E ELMALLAWANY**

* Lect. Faculty of Eng. Zagagig Univ.
** Prof. Of Acoustic Eng. Building Research Center.

اختيار خصائص الصوت للتعرف على المتحدث و التعرف على ا رقام العربية

الملخص العربى

هذا البحث يقدم دراسة مقارنة بين ثلاثة خصائص مختلفة للكلام و الهدف من هذه الدراسة هو ايجاد أكثر الخصائص تأثيرا فى التعرف على الحروف العربية و التعرف على المتحدث نفسه. ، فتم دراسة خاصية معاملات الكبسترم وخاصية الطاقة فى المربع الواحد و الخاصية الثالثة كانت معاملات LPC، ثم تطبيق هذه الخواص الثلاثة على الاعداد العربية للتعرف على المتكلم و التعرف على الرقم و ذلك بأستخدام الدوائر العصبية.

ثم التوصل الى دائرة عصبية لكل خاصية وتم تعليمها للتعرف على المتحدث و التعرف على الاعداد العربية و ذلك باستخدام هذه الخواص الثلاثة. و بعد اختبار هذه الدوائر وجد ان دقة الدائرة فى حالة استخدام معاملات الكبسترم كانت 96٪ للتعرف على المتحدث بينما كانت 94٪ للتعرف على الارقام. وكانت دقة الدائرة المعتمدة على خاصية الطاقة 94٪ للتعرف الارقام بينما كانت 60٪ للتعرف على المتحدث، و كانت دقة الدائرة المعتمدة على خاصية LPC 95٪ فى حالة التعرف على المتحدث ، و البحث يقدم دراسة تحليلية لنظام يتعرف على الارقام ويتعرف على الاشخاص باستخدام الدوائر العصبية.

ABSTRACT

This article introduces a comparison between three different processing techniques for the selection of speech features. These features can be used for speaker recognition or speech recognition. A comparison between the performance of a system based on the linear prediction code, a system based on the cepstrum and a system based on the short time energy is introduced. Feature selection is very effective for recognition accuracy. This work illustrates where each of these features are more efficient for speaker recognition or for speech recognition.

The results show that the short time energy in time domain is very effective for speech recognition where its accuracy is found to be 92%. In speaker identification, the accuracy of identification for the features depending on energy in each frame is found to be 60%. It may be recommended that the features based on the energy per frame may be used for speech recognition. The features based on the Cepstrum give accuracy of 94% and 96% for speech recognition and speaker identification respectively. The accuracy of linear prediction code feature is found to be 95% for speaker identification.So the features depend on cepstrum may be recommended for speaker identifier or speech recognition.

A recognition system for spoken digits are given using the above features with neural networks. The neural network has been used as a tool in this comparison.

1. INTRODUCTION

All the application areas of speech technology, including speech recognition, speech synthesis, speech coding and speaker recognition require some form of preliminary analysis of the speech. Speech analysis techniques may be broadly classified as either frequency domain or time domain approacnes.

The major goal in speech analysis is to estimate the frequency response of the vocal tract. The techniques of processing the speech signal using a multiple bandpass filter, discrete Fourier transformation (DFT) and cepstrum can be used to achieve this goal. Time domain measures such as the auto correlation function , zero -crossing rate, and signal energy can also be used to extract useful information about the speech signal. The parameters considered by Walf[1] were pitch at selected point in words, spectral characteristics of nasal consonants, spectral characteristics of selected vowels, estimated slope of the excitation spectrum and duration of a selected vowel. Recognition error rates of 1.5 percent were obtained. The features considered by Sambur[2] were vowel formant frequency and bandwidth and glottal source poles, location of pole frequencies in nasal consonant, pitch contours and timing characteristics. Sambur found that the most important features were timing characteristics, pitch and low -order formants. He obtained an error of 3 percent. Li and Wrench[3] extended Wakita's technique by comparing all the vowels in the unknown utterance with all the vowels in each reference set. They obtained recognition accuracy ranging from 79 percent to 96 percent.

Markel and Davis[4] used a data of approximately 36 hrs. of speech, taken from interviews. The features used were pitch, amplitude, and 10 reflection coefficients. Means and standard deviations for all these parameters were computed for varying number of voiced frames.

2. EXPERIMENTS AND TECHNIQUE

2.1 Linear Prediction Method

The basic idea behind the linear prediction method is that sample value of speech, X[n] can be approximated as a linear combination of the past p speech samples. Mathematically, the linear predictor is described by the equation[5]:

$$X[n] = a_1 X[n-1] + a_2 X[n-2] + \ldots\ldots\ldots\ldots\ldots + a_p X[n-p]$$
$$= \sum a_k X[n-k] \quad ;k=1,2,\ldots\ldots P \qquad (1)$$

where

X[n] is the predicted sample at instant n and $a_1, a_2, \ldots a_p$ are the predictor coefficients. It is impossible to predict each signal sample exactly and this leads to a prediction error E[n] at each sample instant.

$$E[n] = X[n] - X`[n] \qquad (2)$$

Where X'[n] is the actual speech sample.

By minimizing the mean squared error between the actual speech samples and the linearly predicted ones, the predictor coefficients can be determined by solving a set of linear equations. A set of predictor coefficients can predict the speech reasonably accurately over stationary portions. In order to match the time varying properties of speech, a new set of predictor coefficients are calculated every 10 - 30 ms[1]. The problem in linear prediction is to determine the a_k coefficients so as to minimize the mean square error, E over a specified number of samples.

$$E = \sum E^2[n] = \sum \{ X'[n] - X[n] \}^2$$

$$\text{Or } dE/da_j = -2 \sum X[n-j] \{ X[n] - \sum a_n X[n-k] \} = 0$$

So,

$$\sum a_k \sum X[n-j]X[n-k] = \sum X[n] X[n-j];\ k=j=1,2,\ldots p \qquad (3)$$

Two different efficient methods exit for finding the solution of this system of equations. These are known as the auto correlation and covariance methods[6]. The multipliers of the au coefficients and the right-hand sides of the system of equation (3) can be put in the form of auto-correlation values of the speech signal for specific sample shifts. These are computed by first multiplying the speech signal X[n] by a soft window function of duration N samples, and the auto-correlation values are calculated from:

$$R(k) = \sum \{ w[n]X[n] * W[n+k]X[n+k],\ k=0,1,2\ldots\ldots p$$

The auto correlation function gives a measure of the correlation of a signal with a delayed copy of itself. These predictor coefficients will be considered as the feature of the speech samples for each frame.

2.2 Short Time Energy Function

The short time energy function of speech may be computed by splitting the speech signal into frames of N samples and computing the total squared values of the signal samples in each frame. Splitting the signal into frames can be achieved by multiplying the signal by a suitable window W[n], n=0,1,2,..N-1, which is zero for n outside the range (0,n). A simple rectangular window of duration 10- 20 ms is suitable for this purpose. For a window starting at sample m, the short- time energy function E_m may be written as[7]:

$$E_m = \sum \{ X[n] . W[n-m] \}^2 \qquad (5)$$

The energy per frame will be considered as a feature for the speech signal.

2.3 The Cepstrum

The cepstrum of a signal is the Fourier transform of the logarithm of its power spectrum. Let $X(\omega)$ denote the specrtum

of the voiced speech signal, P(ω) the spectrum of the pitch impulses and H(ω) the spectrum of the vocal tract which includes the effects of the glottal wave form. The relationship between the magnitude of these three spectra can be expressed simply as follows[8]:

$$|X(\omega)| = |P(\omega)| \cdot |H(\omega)|$$

Taking the logarithm of this equation gives :

$$\text{Log}\{|X(\omega)|\} = \text{Log}|P(\omega)| + \text{Log}|H(\omega)| \qquad (6)$$

Thus, in the logarithm of $|X(\omega)|$ the contributions due to $p(\omega)$ and $H(\omega)$ are added. The contribution from $H(\omega)$, which is essentially determined by the properties of the vocal tract itself, tends to vary slowly with frequency, while the contribution from $p(\omega)$ tends to vary more rapidly and periodically with frequency. These two component should be separable by means of a linear filtering operation. Removing pitch ripple from equation (6)leaves only the vocal tract transfer function. Taking the Fourier transform of it, cepstrum can be obtained.

2.4 TIME NORMALIZATION

By time normalization, is meant the process where time-varying features within the words are brought into line. The classical technique is simply to stretch or compress the unknown word uniformly until it attains the same length as the reference. This process depends for its accuracy on accurate end point identification. Time normalization is now frequently done by a process known as time warping[9]. Time warping technique is used in this work for time normalization for spoken digits.

2.5 ENDPOINT ALIGNMENT

The first step in recognition or identification is the problem of endpoint alignment(determine the location of the spoken word during the recording interval). The algorithm used to determine the endpoint has been described by Rabiner and Sambur[10]. The end region is defined as the region from the end of the word to the point at which the energy first exceeds 10 percent of the maximum energy. Equivalently, an initial region is defined from the beginning of the word to the point at which the energy first exceeds 10 percent of the maximum. The remaining section is termed the middle region. The middle region is the desired signal.

2.6 EXPERIMENTS

Speaker identification becomes more difficult as the size of the speaker population increases, the input template must be compared each with each stored template. Using neural network for speaker identification may solve this problem.

Five persons have recorded 2250 files containing Arabic spoken digits. These files are divided into two sets. One set is used for learning the neural network and the other for testing it. Improving the accuracy of the performance of the neural

network can be achieved by using a tree structured network[11] So for speaker identification, there are nine networks, neural network for every digit. Fig. 1 shows the block diagram of the hardware implementation of the neural network speaker/speech recognition system. After endpoint alignment and time normalization, these files are used for preparing the features of each digit. This is done using a PC with speech card and programs in C-language have been written for this propose. The features which have been calculated are energy density per frame in time domain and cepstrum in time for the speech signal. The frame length is taken as 30 ms.
After selection of features, the learning of the neural network is done in a supervised fashion to compare the performance of each feature. For a given collection of input/output pairs of data $(x_1,t_1),\ldots,(x_n,t_n)$, the parametric learning modifies the parameters of the network minimizing the given performance index Q. The general scheme of learning can be concisely expressed as:

$$\Delta\_parameter = -\alpha\, \partial Q/\partial parameters$$

where α denotes the learning rate. The parameters of network are adjusted following these increments.

$$new\_parameters = actual\_parameters + \Delta\_parameters$$

The reader can refer to [12,13] for more details.

3. RESULTS AND DISCUSSION

In pattern recognition, a comparison is made between a test pattern, representing the unknown to be recognized or identified and one or more reference patterns which exist in the pattern library. This comparison takes time depending on the contents of the library. A neural network learns the features of the patterns. So, it takes a very short time for recognition or identification.
The normalized features are applied as inputs to the neural network. For the energy feature, the frame length is taken as 30 ms. After time normalization the longest ward has 40 frames, so the neural network has 40 inputs. Fig 2 shows samples of energy-features of the different spoken arabic digits by the same speaker. The energy-features of digit 'wahed' for different speakers is given in Fig. 3. Any speech contains information about the word being spoken and about the identity of the speaker. In speech recognition we wish to select the first type of features and reduce the effect of the second. The information about the word being spoken can not be isolated from the information about the identity of the speaker. The neural network used for speech recognition has 40 inputs with 4 outputs and one hidden layer with 20 nodes. After learning the network and adjusting the neural network parameters, a recognition accuracy of 92% is obtained. A neural network with the same inputs and 3 outputs and one hidden layer with 20 nodes is used to identify the speaker.

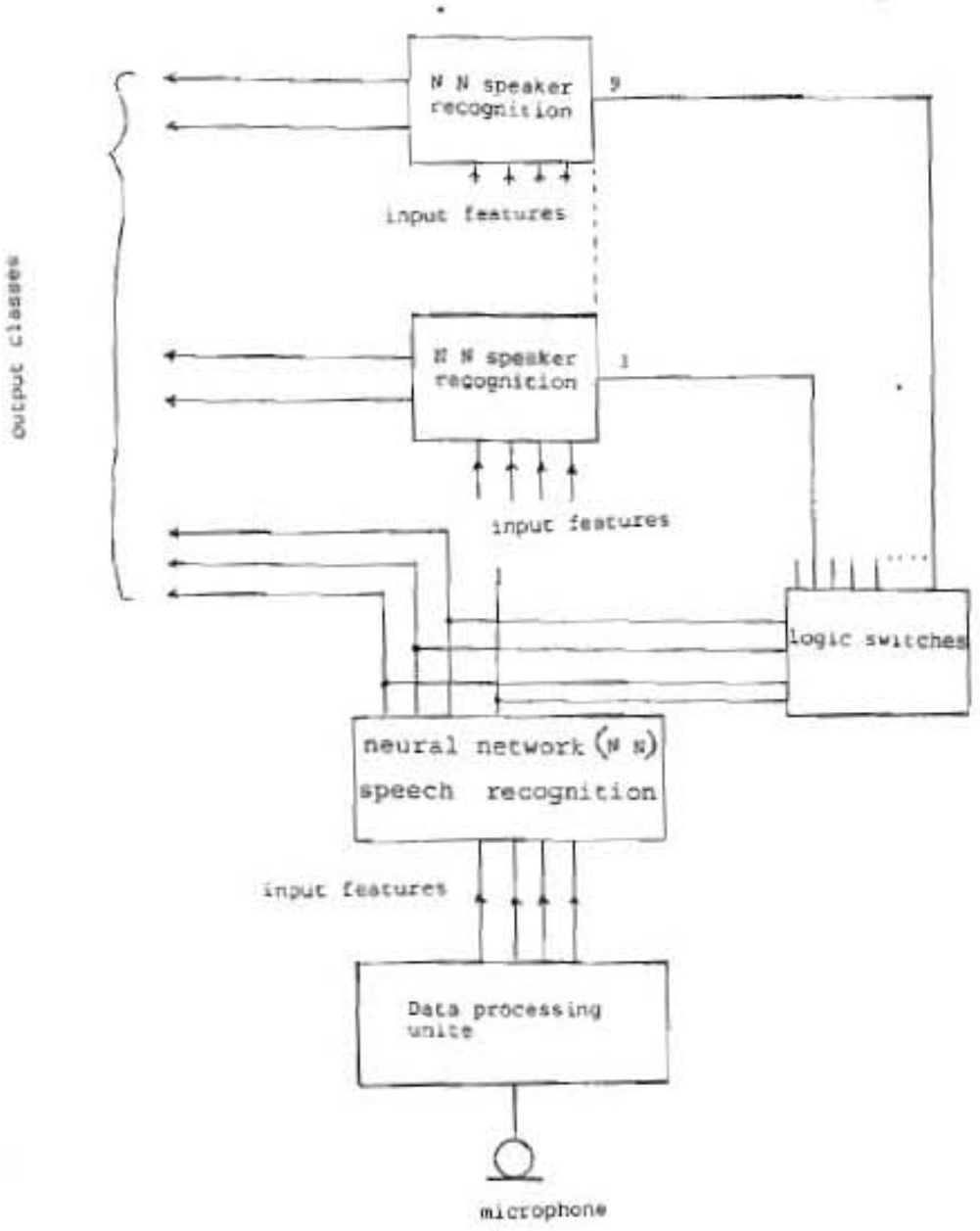

Fig. 1 The hardware implementation of the neural network speaker - speech recognition system

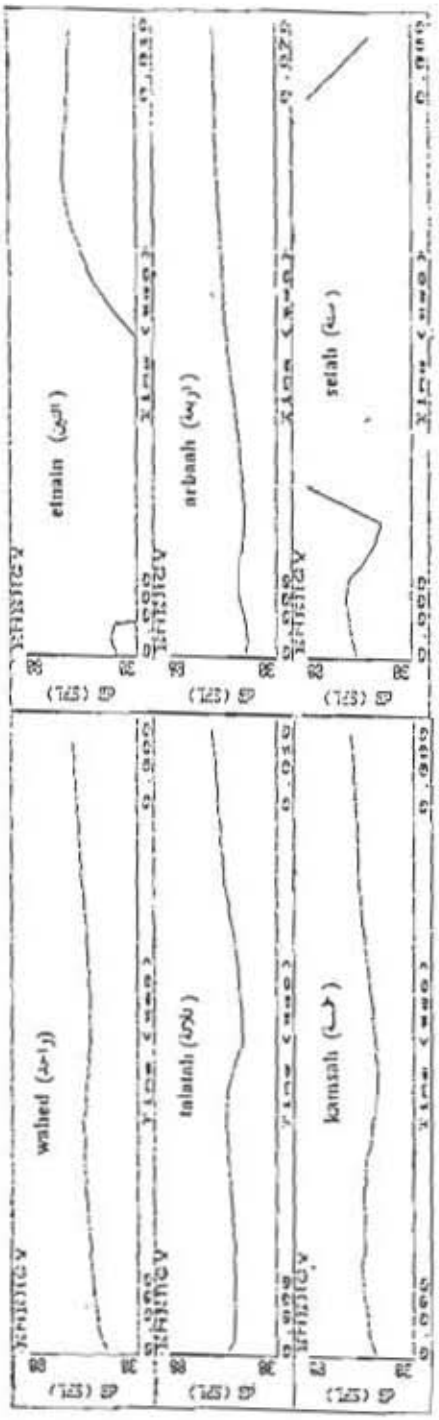

Fig. 2 Samples from energy features for different arabic spoken digits.

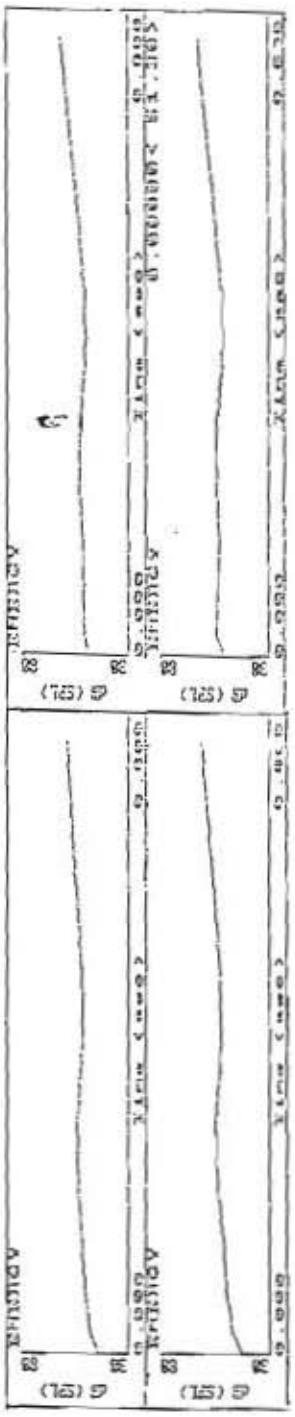

Fig. 3 Samples from energy features of digit "wahed" for the same speaker.

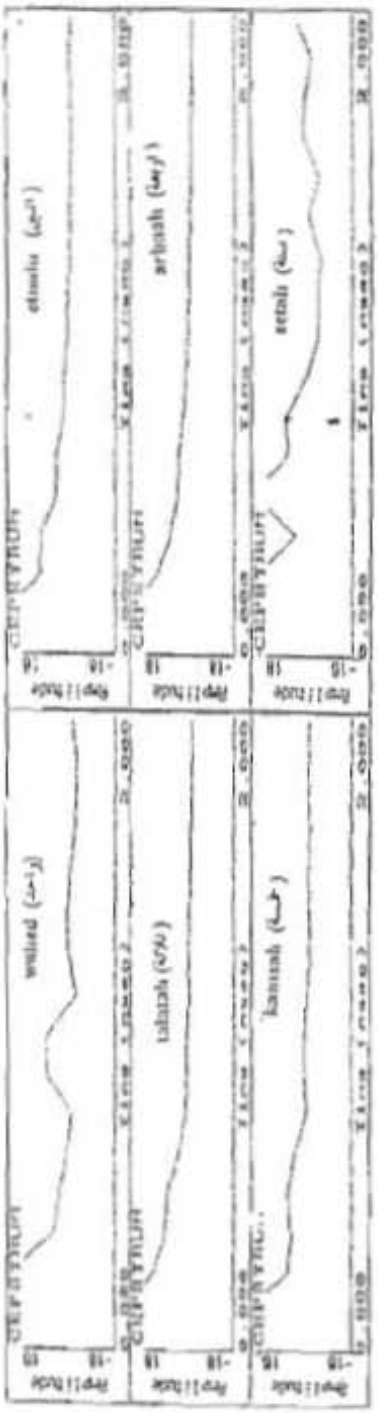

Fig. 4 Samples from cepestrum features for different arabic spoken digits.

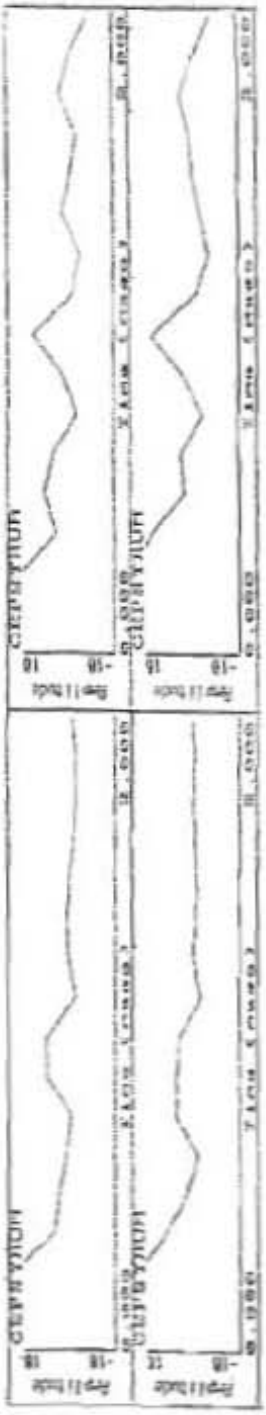

Fig. 5 Samples from cepestrum features of digit "wahed" for the same speaker.

This network may be connected to the recognition network through an activated logic switch. The first network recognizes the digit while the second identifies the speaker. Tables 1 and 2 summarize the learning results. From these results it is recommended that energy per frame may be used for speech recognition. Also, the response of the vocal cord can be expressed by cepstrum analysis which is considered as an efficient feature. The cepstrum coefficients are obtained in 12 frame. So the neural network for speech recognition has 12 inputs with 4 outputs and one hidden layer with 16 nodes while for speaker identifier a neural network with 12 inputs, 3 outputs, and one hidden layer with 16 nodes. is used. The accuracy of the feature using the cepstrum algorithm is found to be 94% and 96% for recognition and identification, respectively.

Table 1 Test results using energy features for speech recognition

Digit	NO. of errors	No. of tested files
1	3	10
2	0	10
3	0	10
4	0	10
5	0	10
6	0	10
7	0	10
8	4	10
9	0	10

It is clear that the feature of the cepstrum analysis is more accurate than that of energy per frame especially for speaker identification. Tables 3 and 4 summarize the testing results of cepstrum features for speech recognition and speaker identification. Figs. 4 and 5 show samples from the cepstrum features of different spoken arabic digits of the same speaker and the cepstrum of digit 'wahad' of different speakers

Table 2 Test results using energy features for speaker identification

Speaker	No. of errors	No. of tested files
speaker1	1	5
speaker2	3	5
speaker3	2	5
speaker4	3	5
speaker5	0	5

Table 3 Test results using cepstrum features for speech recognition

Digit	No. of errors	No. of tested files
1	1	20
2	0	19
3	0	17
4	3	14
5	1	21
6	1	21
7	3	22
8	0	20
9	0	21

Table 4 Test results using cepstrum features for speaker identification

speaker files	No. of errors	No of tested
speaker1	1	5
speaker2	0	5
speaker3	0	5
speaker4	0	5
speaker5	0	5

Mostafa [14] used the LPC coefficients only as a feature to investigate the speaker identifier system using neural network. His results show that the accuracy of LPC was found to be 95%.

4. CONCLUSION

The objective of this article is to investigate the speech features and a comparison between these features is given. It is found that the cepstrum is a powerful feature selection technique especially for speaker identification. A neural network speaker -speech recognition system is given with complete analysis for the arabic spoken digits

REFERENCES

1. Wolf, J. J., "Efficient Acoustic Parameters For Speaker Recognition", JASA, vol.51, June, 1972.
2. Sambur, M. R., "Selection of Acoustic Features For Speaker Identification", IEEE-Trans., vol.ASSP-23, no.2, April, 1975.
3. Li, K. and Wrench, E. H., "An approach to text-independent speaker recognition with short utterances"ICASSP-83, 1983.

4. Markel, J. D. .and David, S. B., "Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base", IEEE Trans.,vol.ASSP-27,No.1, FEB.T,1979.
5. Parsons, T. W., "Voice and Speech Processing", McGraw-Hill Book Company,New York, 1987.
6. Owens,F. J., "Signal Processing of Speech", Macmillan Press LTD, Hong Kong 1993.
7. Schafer,R. W. and Rabiner L., "Parametric Representation of Speech", IEEE Symposium in 1974, ACADEMIC PRESS, New York 1975.
8. Furui, S., "Cepstrum Analysis For Automatic Speaker Verification", IEEE Trans.,vol. ASSP,No.2,April 1981.
9. Sakoe, H. and Chiba, S., "Dynamic Programming Alogrith Optimization for Spoken Word Recignition", IEEE Transaction on Acoustic, Speech and Signal Processing, vol.ASSP-26,Feb. 1975
10. Rabiner L. and Sambur M., "An algorith for Determining the Endpoints of Isolated Utterances", The Bell System Tech. Journal, vol.64, No.2, Feb.1975.
11. Rashwan, M. A. A. and others, "Improving Classification Using a Tree Structured Neural Network", Journal of Intelligent and Fuzzy Systems, Vol.2, 1994.
12. Pao,Y., "Adaptive Pattern Recognition and Neural Networks", Addison-Wesley Publishing Company, INC, New York, 1989.
13. Nguyen, D and Widrow, B., "Neural Networks for Self-Learning Control System", IEEE Contr. Syst. Mag., Vol.10, No.3,April 1990.
14. Mostafa W., "Application of Neural Network on Speaker Recognition", M.Sc.thesis, Faculty of Eng. Cairo univ.,1995