Speaker Identification Using Radial Basis Function Neural Network

Ahmed K. Hassan

Received on : 14 / 1 / 2007
Accepted on :30 / 6 /2008

Abstract

The main objective of this paper is to apply radial basis function neural networks (RBFNN) and evaluated its performance by comparing the results with other method. In this paper two feature vectors are used separately to address speaker identification problem. The features are linear predictive code (LPC) and Mel-frequency cepstral coefficient (MFCC). The radial basis function neural network (RBFNN) approach is used for matching purpose.

This work proposes can be summarized into three steps. The first step is to frame and windowing the input speech signal using hamming window. The second step is to extract the reference and test speech signal using LPC or MFCC as feature extraction. Finally, in the third step, radial basis function neural network has been used to perform the similarity between the test and reference templates. The results show that speaker identification using MFCC and RBFNN gives (100%) identification rate and higher identification rate compared with other method.

الخلاصة

ان الغرض الرئيسي للبحث هو تقيم اداء منظومة تعريف الشخص باستخدام (RBFNN)ومقارنة النتائج مع الطرق الاخرى.في هذا البحث تم استخدام مشفرة التخمين الخطي(LPC)و(MFCC)لاستخراج الصفات المميزة للصوت وكذلك تم استخدام (RBFNN)لقياس التشابه بين الاشارة المرجعيه والاشارة المختبرة. يمكن ايجازالطريقة المقترحة بثلاثة خطوات الخطوة الاولى تم تقطيع اشارة الصوت الى عدد من المقاطع باستخدام(Hamming window) . الخطوة الثانية تم استخراج الصفات المميزة للصوت باستخدام مشفرة التخمين الخطي او(MFCC). الخطوة الثالثة ولغرض قياس التشابة بين الاشارة الاصلية والاشارة المختبرة تم استخدام(RBFNN).وضحت نتائج الاختبار ان طريقة تعريف الشخص باستخدام (MFCC)و(RBFNN) اعطت معدل تعريف بمقدار(100%) وكذلك معدل تعريف اعلى مقارنة مع الطرق الاخرى.

Electromechanical Eng.Dept Technology University

1-Introcution

There has been a growing interest in the use of voice as a means of recognizing or confirming a person's identity. The reasons for this is that a person's voice is considered a biometric identifier, as are finger prints, retinal pattern and DNA. It is a characteristic that is supposed to be intrinsic and unique to a person and, as such, should not be reproducible by anyone else. Furthermore, it benefits from the fact that the person to be identified does not have to carry a card or a key that can be stolen. Also a biometric identifier does not have to be remembered like the personal identification number (PIN) for an automatic machine (ATM) card [1].

Speaker recognition can be classified into identification and verification. Speaker verification refers to the process of determining whether or not the speech samples belong some specific speaker. On the other hand, speaker identification is the process of determining which registered speaker provides a given utterance (word or phrase) [2].

Speaker recognition methods can also be divided into text independent and text dependent methods. In a text independent system, speaker models capture characteristics of what one is saying, while in text dependent system it is assumed that the speaker is cooperative, and wishes to be recognized. This is most often the case in the security applications where a person may identify themselves using their voice to gain restricted access to premises or sensitive information. Some common examples of security applications are voice activated locks, access to be restricted computer data and voice verification for telephone banking and ATM transactions [3].

2-Feature Extraction

2.1-Linear Predictive Coding (LPC)

One of the most powerful speech analysis techniques is the method of linear predictive analysis. This method has become the predominant technique for estimating the basic speech parameters, e.g., pitch, formants, spectra, vocal tract area functions and for representing speech for low bit rate transmission or storage. The importance of this method lies both in its ability to provide the speed and extremely accurate estimates of the computation. The basic idea behind LPC analysis is that a speech sample can be approximated as a linear combination of past speech samples. By minimizing the sum of the squared differences (over a finite interval) between the actual speech samples and the linearly predicted ones.

It is assumed that the variations with time of the vocal tract shape can be approximated with sufficient accuracy by a secession of stationary shapes. It is possible to define an all-pole transfer function H(z) that produces the output speech s(n) given the input excitation u(n) (either an impulse or random noise) is given by [4]:

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \quad (1)$$

Thus, the linear filter is completely specified by scale factor G (gain factor) and p predictor coefficients $a_1,\ldots,a_p$. The number of coefficients p required to represent any speech segment adequately is determined by many factors, such as the length of the vocal tract, the coupling of the

nasal cavities, the place of the excitation and the nature of the glottal flow function.

A major advantage of the all-pole model of the speech production is that it allows one to determine the filter parameters in a straight-forward manner by solving a set of linear equations. In the all-pole model, the speech sample s(n) at n^{th} sampling instant is related to the excitation, u(n) by the following equation[4]:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \quad (2)$$

where u(n) is the n^{th} sampling of the excitation and G is the gain factor. Equation (2) represents the LPC difference equation, which shows that the value of the present output may be determined by summing the weighted present input, Gu(n),and the weighted sum of the post output samples. If the excitation u(n) is white noise, the best estimate of the n^{th} speech sample based on speech samples is given by[4]:

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \quad (3)$$

where $\hat{s}(n)$ is called the predicted value of s(n) and a_k is the predictor coefficient. The prediction error between the actual speech sample and the predicted sample is defined as [4]:

$$e(n) = s(n) - \hat{s}(n) \quad (4)$$

$$= s(n) - \sum_{k=1}^{p} a_k s(n-k) \quad (5)$$

which is the output of a system whose transfer function is[4]:

$$A(z) = \frac{e(z)}{S(z)} = 1 - \sum_{k=1}^{p} a_k z^{-k} \quad (6)$$

where A(z) is the transfer function of the predictor error filter or the inverse filter for the system H(z) and e(z) is the prediction error. To determine the filter coefficients, a_k, the mean squared prediction error is minimized over a short-segment of speech (N). The average square of the prediction error becomes [4]:

$$E_m = \sum_{n=0}^{N-1} e^2(n) = \sum_{n=0}^{N-1} \left[s(n) - \sum_{k=1}^{p} a_k s(n-k) \right]^2 \quad (7)$$

The values of the estimated predictor coefficients can be determined by minimizing the partial derivatives of E_m with respect to a_k.

$$\frac{\partial E_m}{\partial a_k} = 0 \qquad (k = 1,2,...,p) \quad (8)$$

This yields p linear equations:

$$\sum_{n=0}^{N-1} s(n-i)s(n) = \sum_{k=1}^{p} a_k \sum_{n=0}^{N-1-k} s(n-i)s(n-k) \quad (9)$$

where i=0,1,…,p and k=1,2,……,p.

Defining

$$R(i) = \sum_{n=0}^{N-1-i} s(n)s(n+i) \quad (10)$$

where R(i) is the autocorrelation for the speech sample s(n). Then, Equation (10) can be expressed by matrix representation as:

$$\begin{bmatrix} R(0) & R(1) & .... & R(p-1) \\ R(1) & R(0) & .... & R(p-2) \\ & & & \\ R(p-1) & R(p-2) & .... & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \\ a_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \\ R(p) \end{bmatrix} \quad (11)$$

The p×p autocorrelation matrix of the term has the form of a Toeplitz matrix, which is symmetrical and has the same values along the lines parallel to the main diagonal. This type of equation is called a Yule-Walker equation. Since the positive definition of the autocorrelation matrix is guaranteed by the definition of the autocorrelation function, an inverse matrix exists for the autocorrelation matrix. Solving the equation permits obtaining a_k.

The equation for the autocorrelation method can be effectively solved by the Durbin's recursive solution method [5].

2.2 Mel-Frequency Cepstral Coefficient (MFCC)

The main purpose of the MFCC processor is to minimize the behavior of the human ears. In addition, rather than the speech waveforms themselves, MFCC are seen to be less susceptible to the mentioned variations.

Cepstral parameters whose frequency scale is approximated by Mel-scale are considered very powerful in reprocessing speech signal [6].The block diagram of the structure of an MFCC processor is given in figure (1)

The Mel-frequency wrapping is simulated by using a filter bank. The method for obtaining Mel-scale spectrum is by calculating the IDFT of the logarithmic Mel-scale spectrum as follows:

Suppose that S(m) is the FFT of the input signal s(n), and $H_k(m)$ is the frequency response of the k^{th} filter bank. Then the spectral components for each output filters are [6]:

$$Y_k(m)=S(m)\,H_k(m)\,,\ 0\le m\le N-1,\quad \text{and } 0\le k\le N_f \qquad (12)$$

Where N is the number of coefficients of the signal S, and Nf is the number of filter bank. Then, the spectrum energy coefficients are given by [6]:

$$E_k=\frac{1}{N}\sum_{n=0}^{N-1}\left|Y_k(n)\right|^2,\ (1\le k\le N_f) \qquad (13)$$

The Mel-frequency cepstral coefficients (MFCC) can by found by converting the log Mel spectrum back to the time [6].

$$c_{mfcc}(i)=\frac{1}{N_f}\sum_{k=1}^{N_f-1}\log_{10}(E_k)\,e^{j\frac{2\pi}{N_f}ik},\quad 1\le i\le N_c \qquad (14)$$

3- Radial Basis Function Neural Network

Artificial Neural Networks (ANNS) are information paradigms inspired by the way biological nervous system, such as human brain, process information. An ANN consists of a large number of densely connected processing units, which are often referred to as artificial neurons or nodes. These nodes are interconnected through links. Each link is associated with a connection strength, which is often called weight. The first computational model for ANN single-layer perceptron network was proposed by McCulloch and Pitts. A single-layer perceptron network consists of only a single layer of output nodes. The inputs are fed directly to the output via a series of weights. Such simple perceptron networks are only capable of learning linearly separable patterns. The architectures of ANN evolve from the perceptron to complex structures as multi-layer feed-forward networks (MFN) and recurrent networks. Radial Basis Function (RBF) is a special case of multi-layer feed-forward neural networks. A RBF and a general multi-layer feed-forward neural network differ on the node characteristics. Sigmoidal functions are usually used as node characteristics for multi-layer feed-forward network, while radial basis

functions are employed as node characteristics for RBF networks [7].

Radial basis function (RBF) surfaced as a possible variant of artificial neural networks (ANNS) in the late 80s and have been used in basically two areas- functional approximation for the time series modeling and pattern classification. In the area of pattern classification they have been used for tasks such as speech recognition, speech prediction, phone recognition and face recognition [8].

The basic architecture of RBF networks is shown in figure (2).The input data is fed into the input layer and the input layer passes it to the hidden neurons, and the output layer combines the output linearly from the hidden neurons [9].

Figure (2) shows basic architecture of RBF networks.

From Fig. (2) each layer is fully connected to the next one with simple first order connection. The output of ith neuron of the output layer is [10]:

$$y_i(x)=\sum_{j=1}^{N} w_{ij}\,\Phi\,(\|x-x^j\|) \qquad (15)$$

where Φ (.) is a function from R^+ to R, generally decreasing, x is the input vector, x^j are the input examples of the learning database and w_{ij} are the weights between RBF and output unit. The index (i) is omitted and Equation (15) becomes [11]:

$$y(x)=\sum_{j=1}^{N} w_j\,\Phi\,(\|x-x^j\|) \qquad (16)$$

4- Speaker Identification Using Radial Basis Function Neural Network (RBFNN) Model

Figure(3) shows speaker identification block diagram using RBFNN model.

Procedure:

1- Framing the input speech signal.
2- Windowing the input speech signal using Hamming window.
3- Extracting the features (either LPC or MFCC) from the reference speech signal.
4- Extracting the features (either LPC or MFCC) from the test speech signal.
5- Feature matching performs the similarity measure between test and reference using RBFNN.

Figure (4) shows flowchart for the speaker identification using RBFNN

5- Experimental Results

There are two inputs to the speaker identification system; the first is the identity claim which may be provided be a keyed-in identification number that gives a reference data corresponding to the claim to be retrieved. The second is actived by a request to speak the sample utterance. All the experiments were performed using 20 speakers. The speech signal is sampled at 16 KHZ using computer blaster (in normal room conditions). The speech samples are quantized into 16 bit. The next step is to normalize the utterance with respect to identity claim. The continuous speech signal is blocked into frames of N samples with adjacent frames overlapping of M samples (M<N)

$$F(k,n)=s(n+M(k-1)),\ n=0,1,.....,N-1 \quad k=1,.....,L \qquad (17)$$

where L is the number of frame in the speech signal. The typical chosen value of N and M are 280 samples (about 17.5 msec) and 100 samples (about 6 msec) respectively. The frame windowing used to minimize the signal discontinuities at the beginning and end of each frame is defined as:

$$f_w(k,n)=f(k,n)w_h(n),n=0,1,\ldots\ldots,N-1 \quad (18)$$

where $w_h(n)$ is the hamming window is used in this work. Then, the utterance is converted into effective parametric representation for speaker identification done by feature extraction step (LPC or MFCC). Feature matching performs the similarity measure between the unknown utterance and reference template. RBFNN is used for matching purpose The RBFNN have two output nodes, one indicating the likelihood that the input vectors belongs to the true speaker and the other likelihood that it belong to an impostor although only the first of these was actually used in experimental results. Target values during training were [0,1] for true speaker frame. The number of training pattern used to train network was typically 1100 (depending upon utterance length) The RBFNN used 275 nodes in hidden layer. The decision rule is then made by selecting the test speech signal with maximum similarity to reference speech signal. The previous procedures are repeated for all unknown speakers and the system is checked to be accessed for identifying speaker or not, then the system is tested to find the identification rate which is defined as:

$$Identification\ Rate\,(IR) = \frac{NO.of\ correct\ identification\ speakers}{total\ NO.of\ speakers} \times 100\% \quad (19)$$

Table (1) shows the results of identification rate for different methods. And table (2) shows the identification rate for different number of speakers and various identification methods.

Table (1) Results of identification rate

Model	Identification Rate (%)
LPC	70
MFCC	85
LPC + RBFNN	95
MFCC + RBFNN	100

Table (2) Identification rate for different NO of speakers and various identification methods

No of speaker / Model	5	10	15	20
	Identification rate (%)			
LPC	100	80	73.33	70
MFCC	100	100	93.33	85
LPC+RBFNN	100	100	100	95
MFCC+RBFNN	100	100	100	100

Conclusion

The following points are concluded from the simulation results:

1-Speaker identification using Mel-frequency cepstral coefficient (MFCC) and radial basis function neural network (RBFNN) model gives the highest identification rate compared with other method and it can be seen that the identification rate of this method is (100%).

2-Speaker identification using LPC and RBFNN model gives lower identification rate compared with MFCC and RBFNN.

3-The MFCC is better method as feature vector com pared with LPC and can be used to increase the robustness of speaker identification system. MFCC's are shown to be less susceptible to the mentioned variations. Also it can be seen that the identification rate of MFCC is about (85%) and about (70%) when LPC is used as feature extraction.

4- Increasing number of speakers gives more reliable for the speaker identification system.

Reference

1-S. Fredrickson and L. Tarassebko, "Radial Basis Function for Speaker Identification", Proc. ESCA workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland, PP. 107-110, 1994.

2- Richard J. Mammone et al., "Robust Speaker Recognition", IEEE Signal Processing Magazine, September, 1996.

3- Ceoffrey G. Zweig, "Speech Recognition with Dynamic Bayesian Network", PhD Thesis, University of Galifornia, Barkeley, Spring, 1998.

4- L. R. Rabiner and Ronald W. Schafer, "Digital Processing of Speech Signal", Prentice Hall New Jercy, 1978.

5- Fadel S. Hassen, "Cepstral Based Speaker Recognition System", MSc. Thesis, Mustansiria University, 2003.

6- Manal Hassan Mouhammed, "Stochastic Modeling and Quantization Applied for Arabic Speech Recognition ", Ph.D.thesis, University of Technology, 1992.

7-Ghaida'a Wajeh Ahmed, "Speaker Recognition Using Hybrid Transform", MSC.Thesis, Informatics Institute for Postgraduate Studies Iraqi Commission for Computers and Informatics, 2006.

8- Mesbahi Larbi and Benyettou Abdelkader, " A New Look to Adaptive Temporal Radial Basis Function Applied in Speech Recognition", Department of Computer Science, Saida University, Department of Computer Science, Algeria, Journal of Computer Science 1(1): 1-6, 2005.

9- Tae Hang Park, "Toward Automatic Musical Instrument Timbre Recognition ", Ph.D.thesis, University of Princeton, November, 2004.

10- Christian Jutten, "Supervised Composite Networks", IOP Publishing Ltd and Oxford University Press, 1997.

11- Humphrey K.K Tung, Pascal Baup and Michael C.S.Wong," A Radial Basis Function Approach to Credit Barrier Model", City University of Hong Kong, August, 2007.

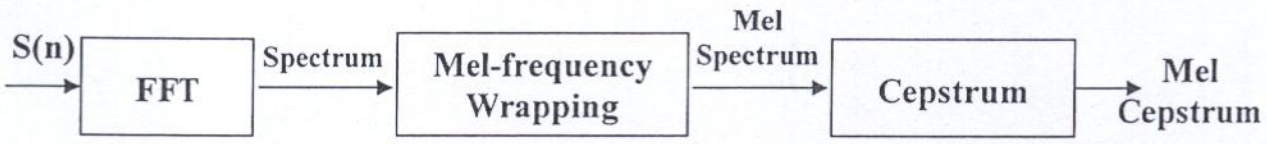

Fig.(1) Block diagram of the MFCC processor

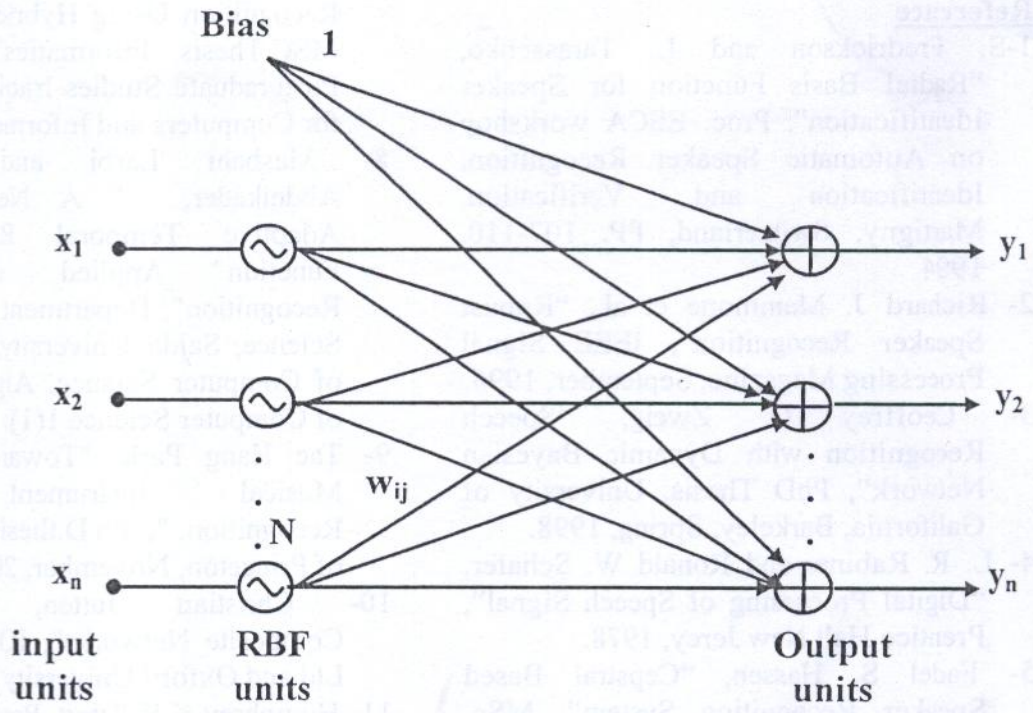

Fig. (2) Basic architecture of RBF networks

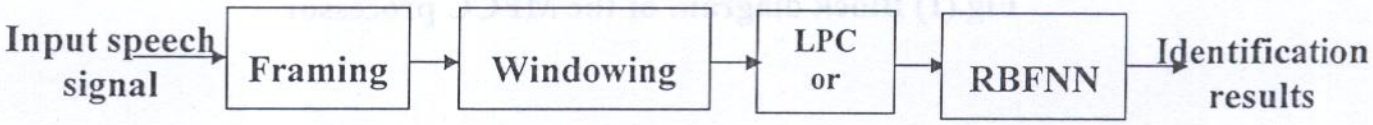

Fig. (3) Speaker identification using radial basis function neural network (RBFNN) model

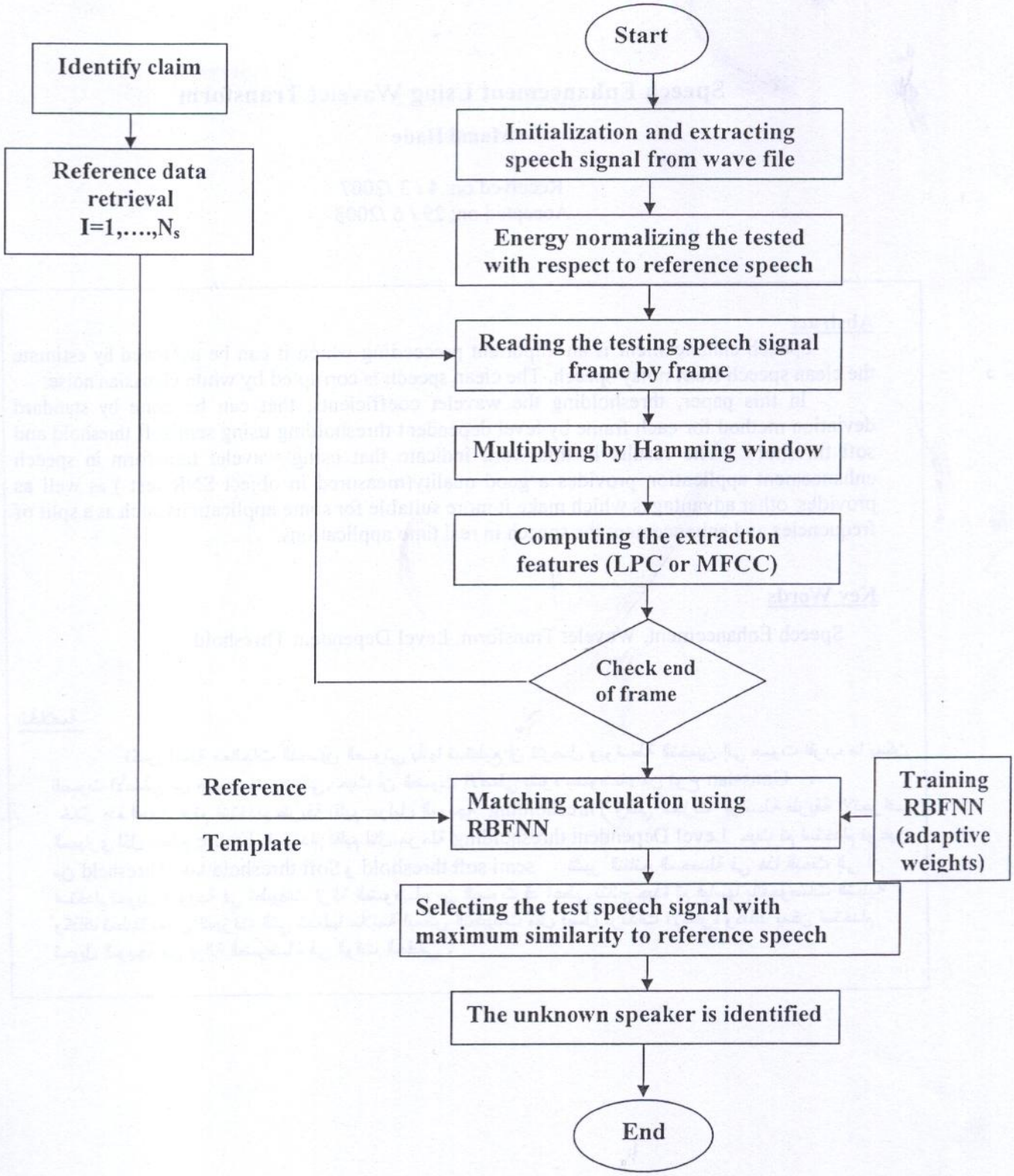

Fig. (4) Flowchart for the speaker identification using RBFNN