Eng. & Tech. Journal ,Vol.29 ,No.15, 2011

Speaker Identification Using Wavelet Transform and Artificial Neural Network

Manal Hadi Jaber*

Received on: 2/5/2011
Accepted on: 3/11/2011

Abstract

This paper presents an effective method for improving the performance of speaker identification system based on schemes combines the multi resolution properly of the wavelet transform and radial basis function neural net works (RBFNN), evaluated its performance by comparing the results with other method. The input speech signal is decomposed into L sub band. To capture the characteristic of the vocal tract, the liner prediction code of each (including the linear predictive code (LPC) for full band) are calculated. The radial basis function neural network (RBFNN) approach is used for matching purpose.

Experimental results shows that the speaker identification using the methods achieve (combines the wavelet and RBFNN) give (100%) identification rate and higher identification rate compared with multi band liner predictive code, in this paper used Matlab program to prove the results.

Key words: speaker identification, wavelet transform, linear predictive code, radial basis function artificial neural network.

تعريف الاشخاص باستخدام تحويل المويجه المتعدده والشبكة العصبية الاصطناعيه

الخلاصه

يقدم هذا البحث طريقة فعالة لتحسين اداء منظومة تعريف الاشخاص اعتمادا على دمج خصائص تحويل المويجة المتعددالتحليل وشبكة الخلايا العصبية المعتمده على دالة الاساس القطري (شعاعي) (RBFNN).) ومقارنة النتائج مع الطرق الاخرى, للحصول على خصائص الحبال الصوتية تم استخدام مشفرة التخمين الخطي LPC والمتضمنة التخمين الخطي للحزمة الكاملة . و تم استخدام (RBFNN) لقياس التشابه بين الاشارة المرجعيه والاشارة المختبرة. وضحت نتائج الاختبار ان تعريف الاشخاص بطريقة الدمج المنجزه اعطت معدل تعريف مقداره 100% وكذلك معدل تعريف اعلى مقارنة مع multi band LPC . تم استخدام برنامج matlab لاثبات النتائج .

1- Introduction

Speaker identification is the process of determining which resisted speaker provides a given utterance by feature extraction of a small amount of data from the voice signal that can later be used to represent each

speaker. Feature matching involves the actual procedure to identify the unknown speaker by comparing extracted features from his/her

* Electromechanical Engineering Department, University of Technology/Baghdad

https://doi.org/10.30684/etj.29.15.17
2412-0758/University of Technology-Iraq, Baghdad, Iraq

voice input with the ones from a set of known speakers [1].

The identification may close set, where it is assumed that the unknown is in the set of known speakers, or open set where the unknown speaker may or may not be in the set of known speakers. Open set identification is more difficult. It is equivalent to performing a closed set identification followed by verification [2]. speaker identification system was used in this paper. Figure (1) shows the block diagram of speaker identification system. When the feature extraction component is performed using discrete wavelet transform and LPC, while the speaker matching components performed using RBFNN.

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 [3]:

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

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[3]:

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

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[3]:

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

Where $\hat{s}(n)$ is called the predicted value of s(n) and a_k is the predictor coefficient.

The values of the estimated predictor coefficients can be determined by minimizing the partial derivatives of E_m with respect to ak (k=1,2,……,p)

$$\frac{\partial E_m}{\partial a_k} = 0 \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

This yields p linear equations:

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

$$\ldots\ldots\ldots\ldots (5)$$

where i=0,1,…,p and k=1,2,……,p.

The autocorrelation for the speech sample s (n) is R (i)

$$R(i) = \sum_{n=0}^{N-1-i} s(n)s(n+i) \ldots\ldots\ldots\ldots\ldots (6)$$

Then, Equation (6) can be expressed by matrix representation as [4]:

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

$$\ldots\ldots\ldots\ldots\ldots (7)$$

The pxp 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 [4].

2.2 - Discrete Wavelet Transform

The Discrete Wavelet Transform (DWT) is a special case of the WT that provides a compact representation of a signal in time and frequency that can be computed efficiently. The DWT analysis can be performed using a fast, pyramidal algorithm related to multi rate filter banks.

The main process performed by this algorithm is a number of successive high pass and low pass filtering of the time domain signal.

Consider wavelet function [5]:

$$y_{jk}(t) = 2^{-j/2} y\left(2^{-j} t - K\right)$$

$$\ldots\ldots\ldots\ldots\ldots\ldots (8)$$

Note that function $y\left(2^{-j}t-K\right)$ was taken from the wavelet function as a result of binominal dilation and two parameter shift .that will do for representing arbitrary signal x(t) from $L^2(R)$ in the form of

$$x(t)=\sum_{j}\sum_{k}d_{jk}y_{jk}(t) \quad \text{.......................(9)}$$

Where d_{jk} wavelet coefficients which are defined by equation:

$$d_{jk}=\int x(t)y_{jk}(t)dt \quad \text{..................(10)}$$

if view the sequence of scaling function:

$$f_{jk}(t)=2^{-j/2}f\left(2^{-j}t-k\right) \quad \text{..............(11)}$$

Then we get alternative wavelet transform given a discrete signal $x(n), n\in Z$, its discrete wavelet transform up to a level J of depth (its multi resolution decomposition on J octaves)is defined [6]:

$$x(n)=\sum_{j=1}^{J}\sum_{k=Z}d_j(K)y_{2^j}\left(n-2^jK\right)+\sum_{k=Z}a_j(K)f_{2^j}\left(n-2^jK\right) \quad \text{........(12)}$$

Where $y_{2^j}\left(n-2^jK\right)$ are the analysis wavelets and $f_{2^j}\left(n-2^jK\right)$ are the scaling sequences. These are the discreet versions of the continuous wavelet and scaling function $d_j(K)$ are the wavelet coefficients, or the detailed signal at scale $2^j; a_j(K)$ are the scaling coefficients, or the approximated at scale 2^j; .Note that the wavelet coefficients represent the details of the original signal at different of resolution. The scaling coefficients represent the approximation of the original signal x (n).

The coefficients h (n) and g (n) ,used to construct the set of scaling and wavelet basis, are low pass (H) and high pass (G) FIR filter coefficients respectively. Where H= {h (n)} and G = {g (n)}.According to the equation (17), G is the reverse of H.

$$g(n)=(-1)^n\ h\,(N-n) \quad \text{..................(13)}$$

Figure (2) shows filter band of discrete wavelet transform .The symbol ↓2 is down-sampler (decimator) that it takes a signal x(n) as input and produces an output of y(n) = x(2n) ,which mean that is discarded [5].

3- Radial Basis Function Neural Network

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 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 (3) shows basic architecture of RBF networks.

From Figure (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]:

$$y_i(x) = \sum_{j=1}^{N} w_{ij} f\left(\left\| x - x^j \right\|\right) \quad \text{.................(14)}$$

where f (.) 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 (19) becomes [11]:

$$y(x) = \sum_{j=1}^{N} w_j f\left(\left\| x - x^j \right\|.\right) \quad \text{................(15)}$$

4- Distance Measure

For the speaker identification task, the unknown speech is compared with all reference speech. This can be done through a distance measure. A simple geometric distance measure can be used. That is the Euclidean distance measure. The Euclidean distance can be defined as [12]:

$$D(x\text{-}y) = (a_x - a_y)^T (a_x - a_y) \quad \text{.................... (16)}$$

Where a_x and a_y are prediction coefficients for reference and tested speech respectively. The decision rule is to select the Pattern that best matches the unknown. In this approach, the minimum distance classifier is used. This classifier assigns the unknown speech pattern to the nearest reference speech pattern.

5- Multi band Linear Predictive Code (MBLPC) Speaker Identification Model

Figure (4) shows speaker identification using multi band combination feature model.

6- Simulation Results for Multi band Linear Predictive Code (MBLPC)

Simulations of speaker identification using Multi band Linear Predictive Code (MBLPC) is carried out. The speech signal is sampled at 16 KHz using a computer sound blaster (in normal room conditions). The speech samples are quantized into 16 bit. The continuous speech signal is sectioned into frame of N with adjacent frames overlapping of M samples. Typically chosen values of N and M are 320 samples (about 20 ms) and 128 samples (about 8 ms) respectively. All the experiments were performed using section of speech from 15 speakers.

* Table (1) shows identification rate using MBLPC model with two bands and different types of wavelet family (db2, db4, db6, db8 and db10).

* Table (2) shows identification rate using MBLPC model with three bands and different types of wavelet family.

* Table (3) shows identification rate using MBLPC model with four

bands and different types of wavelet family.

7 - Speaker identification using multi band and radial basis function (MBLPC and RBFNN) MODEL

The proposed model for speaker identification using multi band and radial basis network is shown in figure (5).

8- Simulation Procedure for MBLPC and RBFNN (MODEL)

a- Framing the input speech signal.
b- Windowing the input speech signal using Hamming window.
c- Obtaining discrete wavelet transform decomposition of the input speech signal using different types of wavelet family.
d- Obtaining the approximate coefficients from the wavelet transform.
e- Extracting the features LPC from each band (including full band)
f- Extracting the features LPC from the test speech signal.
g- Recombining the LPC from each band and full band in a signal feature vector.
h- Feature matching performs the similarity measure between test and reference using RBFNN.

9- Experimental Results for MBLPC and RBFNN (MODEL)

All the experiments were performed using 15 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 utterance is converted into effective parametric representation for speaker identification done by feature extraction step discrete wavelet transform and (LPC). 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:

$$[identification\ Rate(IR)=] \frac{NO\,of\,correct\,identification\,speakers}{total\,NO\,of\,speakers}\times 100\% \quad .(21)$$

Table (1) shows identification rate using wavelet transform and RBFNN with two bands and different types of wavelet family (db2, db4, db6, db8 and db10).

Table (2) shows identification rate using wavelet transform and RBFNN with three bands and different types of wavelet family

Table (3) shows identification rate using wavelet transform and RBFNN with four bands and different types of wavelet family.

From these tables speaker identification rate using wavelet and RBFNN with two band gives higher identification rate compared with other bands.

10- Conclusions

The following points are concluded from the simulation result.

1- Speaker identification using wavelet transform and (RBFNN) model gives the highest identification rate compared with MBLPC and it can be seen that the identification rate of this method is (100%) for wavelet family (db2,db4,db6,db8) and two bands.

2- Speaker identification using wavelet transform and (RBFNN) model gives the highest identification rate with different band in db2, db6 and db8.

3- Speaker identification using this method with two bands gives the highest identification rate compared with three band and four bands.

4- Speaker identification using this method with four bands gives bad results for identification rate compared three bands.

11- References

[1]Gyanendra Kr. Verma,., “Speaker Identification System using Wavelet Transform”, MSC. Thesis, Indian institute of information technology Deemed University, 2009.

[2]K. Daqrouq1, Emad Khalaf1, and others, “Wavelet Formants Speaker Identification Based System via Neural Network ”, International Journal of Recent Trends in Engineering, Vol 2, No. 5, November 2009.

[3]L. R. Rabiner and Ronald W. Schafer, “Digital Processing of Speech Signal”, Prentice Hall New Jercy, 1978.

[4]Fadel S. Hassen, “Cepstral Based Speaker Recognition System”, MSc. Thesis, Mustansiria University, 2003.

[5]Wael AL-sawalmeh and others "The Use of Wavelets in Speaker Feature tracking identification System Using Neural Network" MSC, Thesis, Philadelphia University, issue 5, volume 5, 2009.

[6]P. Shanmugapriya, Y. venkataramani, "Implementation of Speaker Verification System Using Fuzzy Wavelet Network", Saeanathan College of engineering, Trichy, Tamilnadu. 978-1-4244-9799-20 11 IEEE.

[7]Musab Tahseen Salah Al-Deen," Speech compression Using Linear Prediction Coding in Wavelet Domain", M.Sc. Thesis, Al-Mustansiria University, 2004.

[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.

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

Table (1) shows identification rate using MBLPC model with two bands and different types of wavelet family (db2, db4, db6, db8 and db10).

Table (1) Identification rate results using MBLPC model

Wavelet family	Identification rate %
db2	93.333
db4	80
db6	86.667
db8	80
db10	86.667

Table (2) shows identification rate using MBLPC model with three bands and different types of wavelet family (db2

Table (2) Identification rate results using MBLPC model

Wavelet family	Identification rate %
db2	93.333
db4	80
db6	86.667
db8	80
db10	93.333

Table (3) shows identification rate using MBLPC model with four bands and different types of wavelet family (db2

Table (3) Identification rate results using MBLPC model

Wavelet family	Identification rate %
db2	86.667
db4	80
db6	86.667
db8	80
db10	86.667

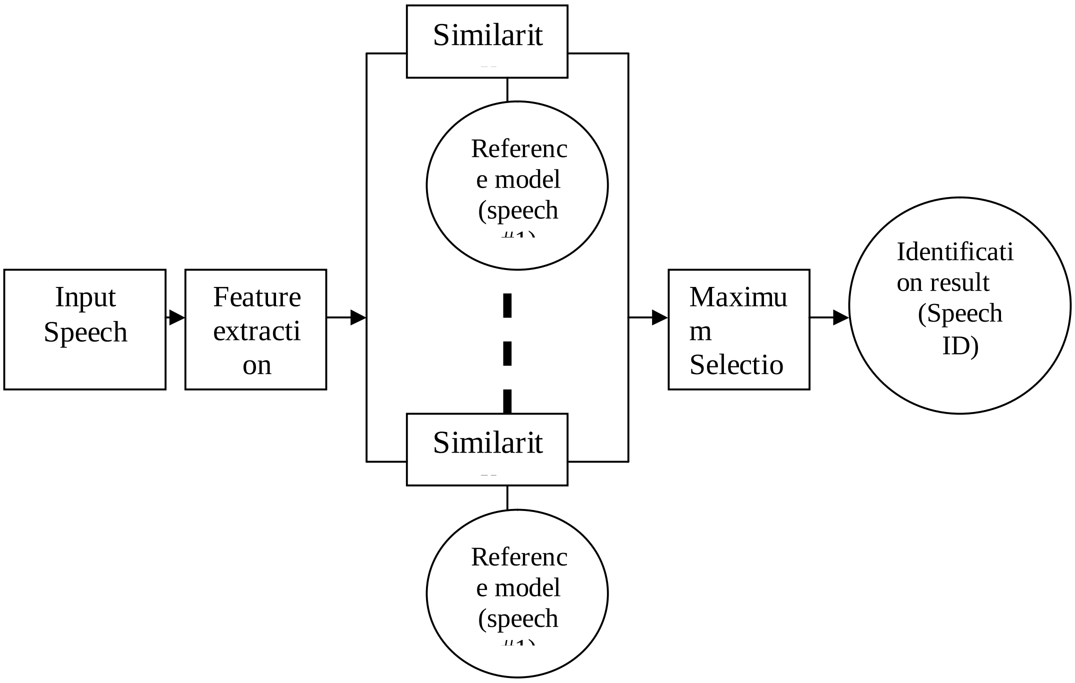

Figure (1) The block diagram of speaker identification

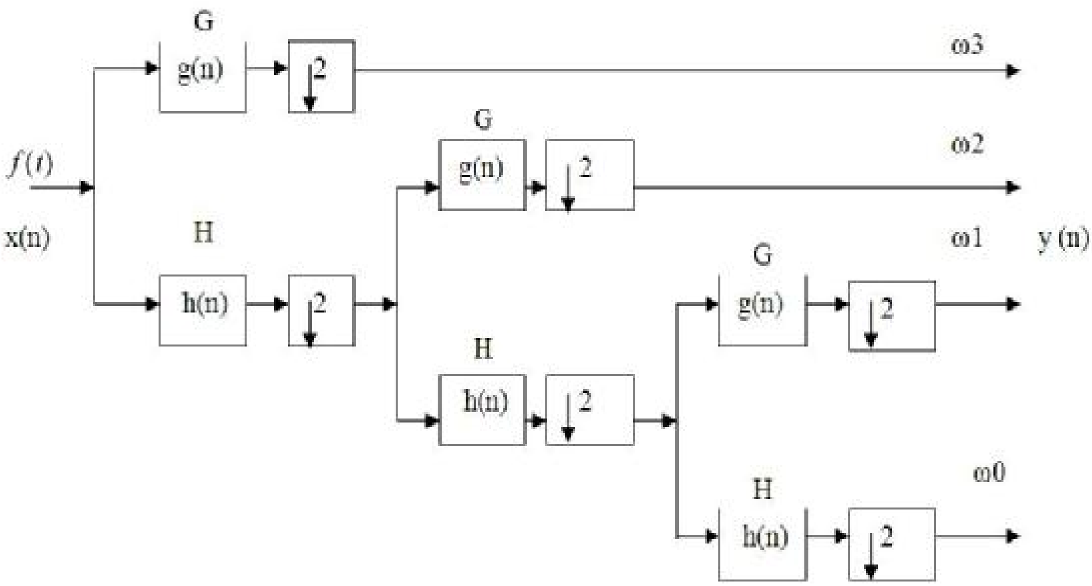

Figure (2) Filter band of discrete wavelet transform [7]

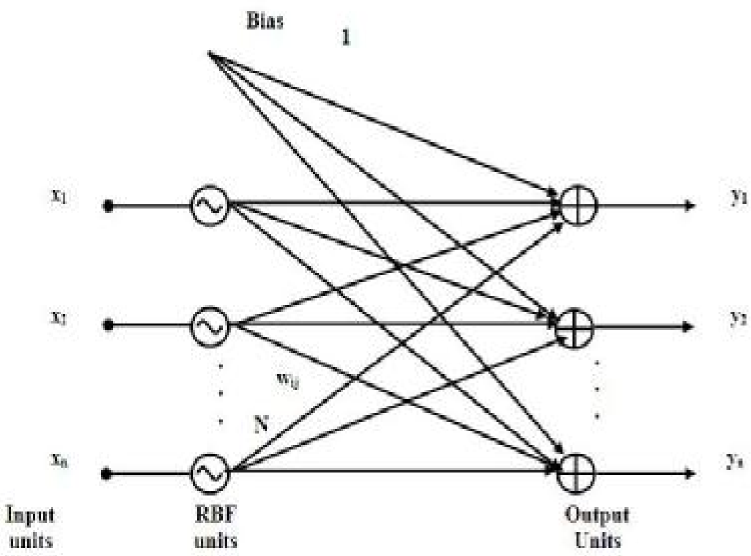

Figure (3) Basic architecture of RBF networks

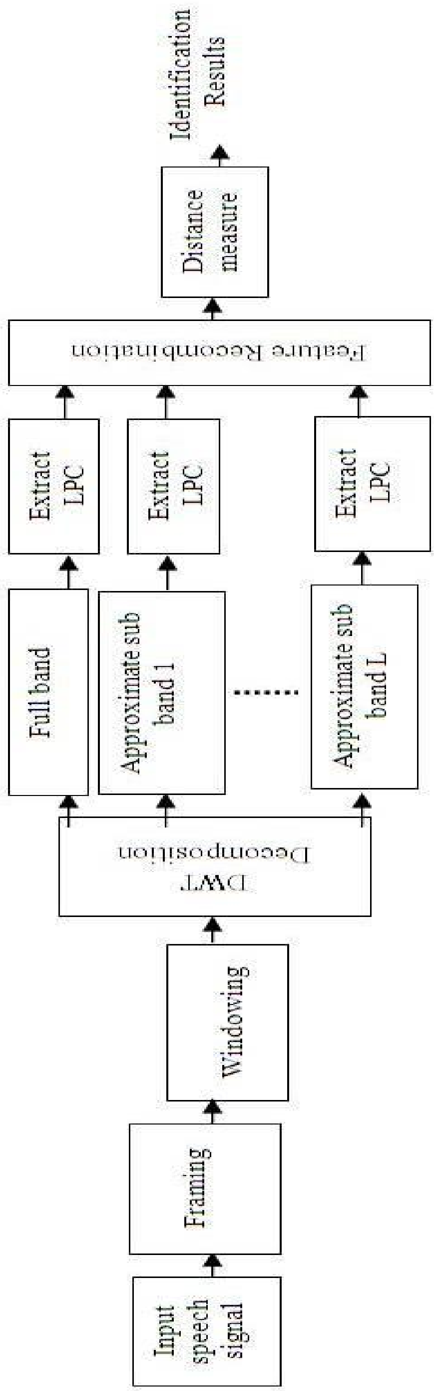

Figure (4) Block diagram of speaker identification using Multi band combination feature model

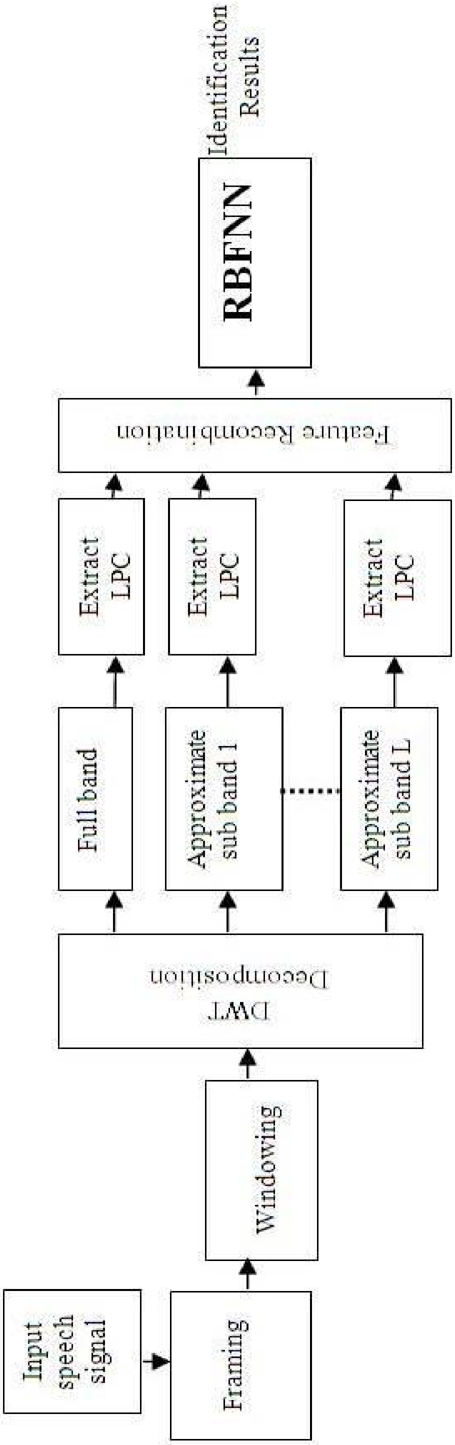

Figure (5) Block diagram of speaker identification using Multi band and radial basis function

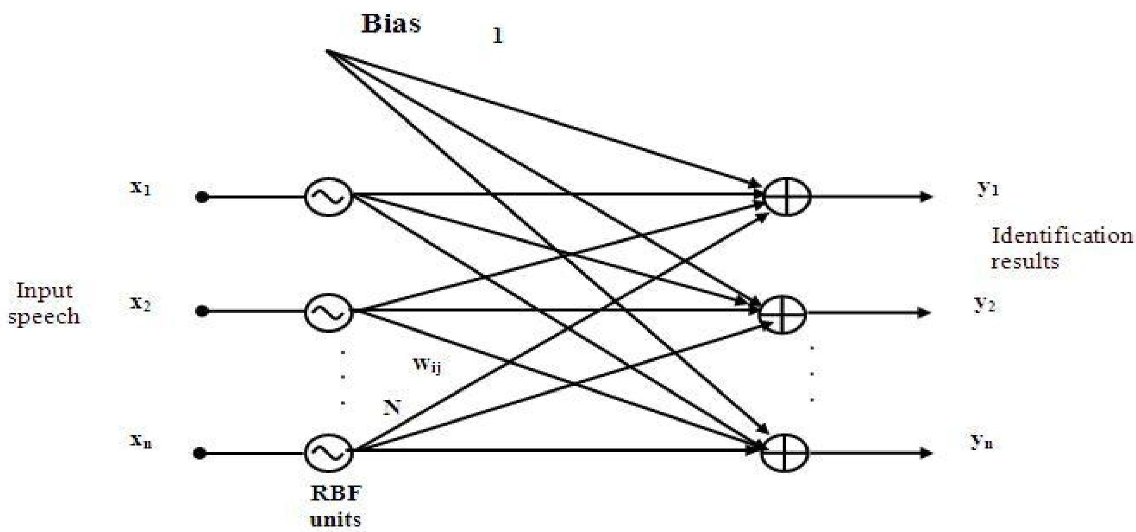

Figure (6) shows the architecture of RBF arterial neural network