As per UGC guidelines an electronic bar code is provided to seure your paper

International Journal for Modern Trends in Science and Technology, 8(04):146-151, 2022

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0804027
Available online at: http://www.ijmtst.com/vol8issue04.html

Speaker Recognition using Radial Basis Function Neural Network with DWT coefficients

Krishnababu Yele | Naveena Shaik | Venkata PavanVellalacheruvu | Naga Raju Parikala

Departmentof Electronics and Communication Engineering, RVR&JC College of Engineering, Chowdavaram, Guntur, India.
*Corresponding Author Email ID: krishnaktsmcl@gmail.com

To Cite this Article
Krishnababu Yele, Naveena Shaik, Venkata PavanVellalacheruvu and Naga Raju Parikala. Speaker Recognition using Radial Basis Function Neural Network with DWT coefficients. International Journal for Modern Trends in Science and Technology 2022, 8(04), pp. 146-151. https://doi.org/10.46501/IJMTST0804027

Article Info
Received: 06 March 2022; Accepted: 5 April 2022; Published: 08 April 2022.

ABSTRACT

Speaker recognition is a means of automatically determining who is speaking by utilizing speaker specific information contained in speech waves to confirm identities asserted by users accessing systems in other words, it allows users to operate various services using their voices. The physiological and behavioral features of a speaker's speech production process are tied to their identity. In the classification and recognition process, neural networks play a significant role. The Radial Basis Function (RBF) is a mathematical function that For faster approximation and classification, neural networks are examined, whereas Discrete Wavelet Transform coefficients are employed instead of random weights. When compared to traditional Mel Frequency Cepstral Coefficients, the proposed Discrete Wavelet Transform technique (DWT) achieved almost the same accuracy with less iterations and less time for recognition.

KEYWORDS:Radial Basis Function Neural Network (RBFNN), Discrete Wavelet Transform (DWT), Mel-Frequency Cepstral Coefficients (MFCC), Speech Recognition, Back propagation, Harris Hawk Optimization.

1. INTRODUCTION

Speaker recognition refers to technologies that are designed to identify, verify, and distinguish individual speakers.It was identified by the regular frequency and flow of their speech pronunciation. In general, speaker recognition can be divided into two categories:Speaker verification and speaker identification.The purpose of Speaker Identification is to identify an unknown speaker from a series of recordings voices that are well-known DWT features deconstruct the speaker signals features.The input for the classification is the DWT coefficient characteristics. When compared to other models, RBFNs have a number of distinct advantages because their models can be scaled and their training is relatively quick updated.

The first step in the speaker recognition step is to extract features of a speech signal. It can be done by many methods like by extracting MFCC,DWT or GFCC features. Next step is to train a network which classifies the speakers. In our paper we took a database consisting of information about 3 speakers of 10 utterances of each.

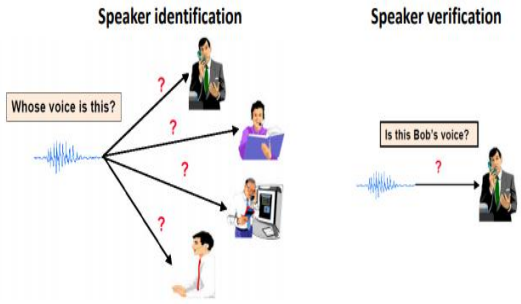

Fig.1.Speaker Recognition categories.

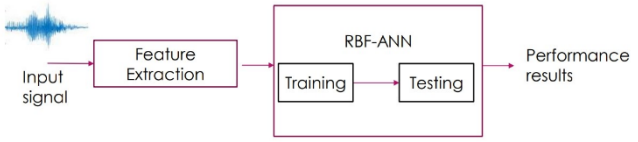

Fig.2 Speaker Recognition System.

After training is done a threshold for decision is fixed testing is done based on previously fixed threshold.

2.RADIAL BASIS FUNCTION NEURAL NETWORK

Radial basis function neural networks are also referred to as feed forward neural networks with a single hidden layer. RBFN networks have a wide range of applications including function approximation, time series prediction, system control and classification and so on. We use the RBFN network to convert linearly non-separable feature classes into linearly separable feature classes by applying linear transformation on the input vectors and increasing the dimensionality of the feature vectors.

A.Architecture of RBFN network

The RBFN network contains three layers: one input layer, one hidden layer and one output layer. In the hidden layer, each node performs a high-dimensional radial basis function. The network inputs are neuron parameters and the output is a linear combination of hidden layer outputs.

The distance between the network inputs and hidden layer centers is used to calculate the input layer outputs. The linear hidden layer is the second layer, and its outputs are weighted versions of input layer outputs. A parameter vector called center exists for each neuron in the hidden layer. The linear combination of hidden layer outputs and bias parameters is calculated by the output layer.

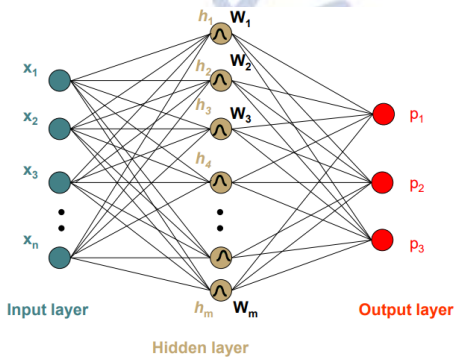

Fig 3. Radial Basis Function Neural Network

The following equation can be used to determine the output of the ith activation function ϕ i in the hidden layer of the network depending on the distance between the input pattern x and the center ci:

$$\varphi i = (||x - ci||) = e^{\left(-\frac{||x-ci||^2}{2.\sigma^2}\right)}$$

The Euclidean norm is represented by || ||, width of the hidden layer is represented by σi and the center of the hidden layer is represented by ci. The network's overall expression is given by:

$$yj = \sum_{i=1}^{m} wij\varphi(x - ci) + \beta j$$

RBFN involves two levels of training. Involves two levels of training. First level is Training of hidden layer nodes. For each of the nodes we have to find out what is the receptor and spread of the radial basis function. Second level is training of weight vectors connecting the outputs of hidden layer nodes to the output layer nodes. Because it is the linear combination of the output of thehidden layer nodes which decides to which class a sample will belong.In this paper we used a back propagation algorithm for training the RBF network.

B.Back Propagation Algorithm

In this paper for training of RBF neural network back propagation algorithm was used. Let us assume x1,x2,x3..xn are the input features and w1,w2,w3,.. are the random weights given for connecting input to activation node and bias b was applied. The activation function used was sigmoid which ranges from 0 to 1 finally connected this hidden node to the output node using some random weights. Then the actual calculated output will be

$$y = [x1w1 + x2w2 + \cdots + x3w3] + b$$

y′ is the predicted output and it is given as 0.

Then the Loss factor given by $L = (y - y')^2$

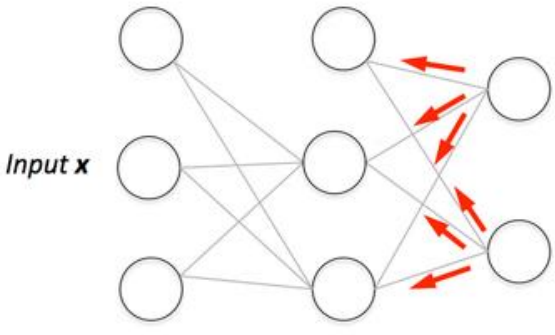

Fig 4. Back propagation network.

The updation of weights is done by the following equation.Let us assume Wn be the nth weight to be updated using a learning rate of μ then

$$Wn = Wn - \mu \frac{dE}{dWn}$$

Learning rate μ=Measures the amount of error that a node in the network's weights are accountable for during training.

3. FEATURE EXTRACTION

For decades, researchers have been researching automatic speech recognition by machines. Acoustic signals can be represented in a variety of ways using parametric models. The MFCC and DWT feature extraction algorithms are the most commonly used among them. Numerous studies have been conducted on MFCC with the goal of improving recognition accuracy. Even at low frequencies, DWT feature extraction delivers good resolution. We employed an efficient method to calculate both MFCC and DWT in this study. We then compared the outcomes of both techniques to determine which was superior.

The first step in extracting features is passing signals through a pre-emphasis filter. Using a pre-emphasis filter, the speech is first pre-emphasized. The filter's impulse response is given by $H(z) = H(z) = 1\text{-}bz^{-1}z^{-1}$ b controls the slope of the filter and is usually between 0.4 and 1.0.The next step is framing. It is done to perform our observation on a single frame so that our analysis can be easily carried out in our research. We used a frame of length 20ms. To assure stillness between frames, there is 10ms overlap between the two adjacent frames.

A.Feature extraction using MFCC coefficients

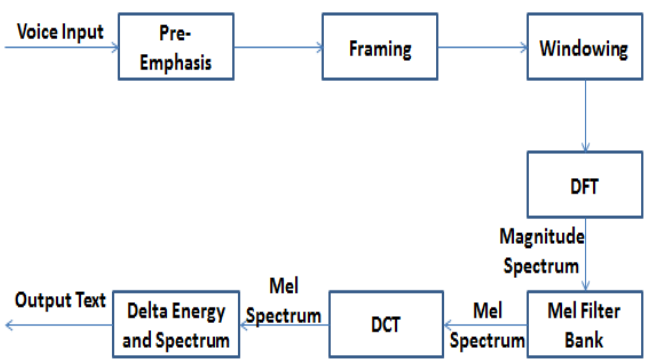

Fig 5.Block diagram for extracting MFCC.

Windowing is now applied to the signal frame. As the signal abruptly changes from high to low and low to high in a rectangular window, considerable high frequency noise is introduced at the starting and finish locations. As a result we opt for the Hamming window. The following is the mathematical expression for hamming window:

$$Ham\,(N) = 0.54 - 0.46cos\left(2\pi \frac{n-1}{N-1}\right)$$

The number of points in one frame is N and n ranges from 1 to N.After the FFT block, the spectrum of each windowed frame is filtered by a set of filters, and the power of each band is computed. We can translate or map the linear power spectrum onto a non-linear mel scale using mel-filter banks. Mel wrapping is the term for this. The conversion of linear frequency to Mel-Frequency is shown in the equation below:

$$Mel\,(f) = 1127ln(1 + \frac{f}{700})$$

Using the equation below, we can calculate the Mel-Frequency cepstrum from the filter bank's output power:

$$c_n = \sum_{k=1}^{K} (log\ S_k) cos[n(k - 0.5)\frac{\pi}{K}]$$

Where $S_k S_k$ is the $k^{th} k^{th}$ filter's output power. And n is a number from 1 to 12. As one of the coefficients, we can also determine each frame's logged energy.

$$E = log \sum_{n=1}^{N} {s_n}^2$$

13 coefficients are chosen for determining speaker since most of the human speech information are in the first 13 coefficients. There are also 13 delta and 13 double delta MFCC coefficients are also present. The use of these coefficients will increase complexity and also contain very less information, Disadvantages: Due to DCT employed in MFCC, if one of the signal's frequency components is destroyed by noise, the entire frequency spectrum will be substantially interfered. In the presence of additive noise, MFCC results are not robust.

B.Feature extraction using DWT coefficients

The wavelet transform is a signal representation that gives the signal a time-frequency representation. Wf(n,m) is a continuous signal frame with limited energy and function of two parameters scaling and shifting parameters m,n accordingly m=2^j,n=2^j. Wf(n,m) is the nth wavelet coefficient at level m, while ψ is the mother wavelet.

$$Wf(n,m) = \frac{1}{\sqrt{m}} \int_{-\infty}^{+\infty} S(t) \cdot \psi\left(\frac{t-n}{m}\right) dt$$

DWT requires two related functions namely scaling and wavelet functions. ψ(t) stands for scaling function and φ(t) stands for wavelet function. A low pass filter's impulse response is h[n], whereas a high pass filter's impulse response is g[n].

$$\psi(t) = \sum g[n] * \sqrt{2}\, \varphi(2t-n) \text{ for } n=0 \text{ to } N-1$$

$$\varphi(t) = \sum g[n] * \sqrt{2}\, \psi(2t-n) \text{ for } n=0 \text{ to } N-1$$

A pair of low-pass and high-pass wavelet filters are recreated from a selected mother wavelet and its related scaling function to create the DWT. The signal is decomposed into a low-frequency and a high-frequency component using those filters. To improve low-frequency resolution, the low-frequency portion might be further decomposed at the following decomposition level. The low-pass and high-pass filtering methods are used.

$$Am+1 = Am * h\,[2n]$$
$$Dm+1 = Am * g\,[2n]$$

g and h represent the low-pass and high-pass conjugate mirror filters, respectively. $*$ is the convolution operation, and Am and Dm are Approximation and Detail coefficients respectively. At most, DWT consists of log2N stages. The first stage generates two sets of coefficients: approximation coefficients CA1 and detail coefficients CD1. In wavelet analysis, the high-scale ,low frequency components of the signal are called the approximatssions, while the low-scale, high-frequency components are called the details. There are many wavelet families, but the Daubechies are the most used in voice detection. The surnames of the daubechies are written as dbN, with N denoting the family's order.

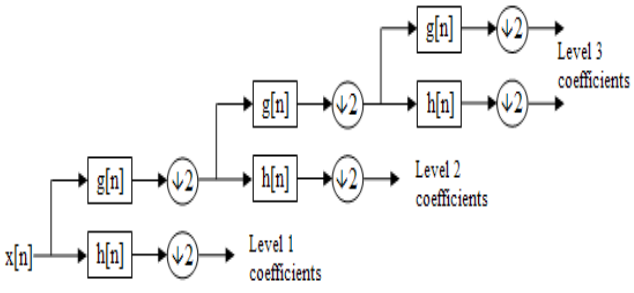

Fig 6. DWT coefficients at different levels.

6. DESIGN APPROACH

Network Creation:A RBF neural network was created to classify the speakers.The hidden nodes of the network can be varied as of our choice we designed in such a way.The network was trained with 10 utterances of each speaker and tested.

Preprocessing:This module tests the applied input signal using numerous factors before converting the analogue signal to a digital signal. This block procedure digitalizes the supplied input signal. This phase involves converting the signal to a discrete form. This translation facilitates data processing and reduces mathematical computations greatly.Pre-emphasis,framing and windowing was done during this step.s

Feature Extraction: Features are extracted using MFCC and DWT coefficients and results are compared.

Training and Testing:In this phase network was trained with 3 speakers' utterances and tested using different extraction techniques and results were compared.

7. RESULTS

In this paper we performed our training on 3 speakers with 10 utterances of each.

Results for MFCC extraction and classification with RBF network are as follows:

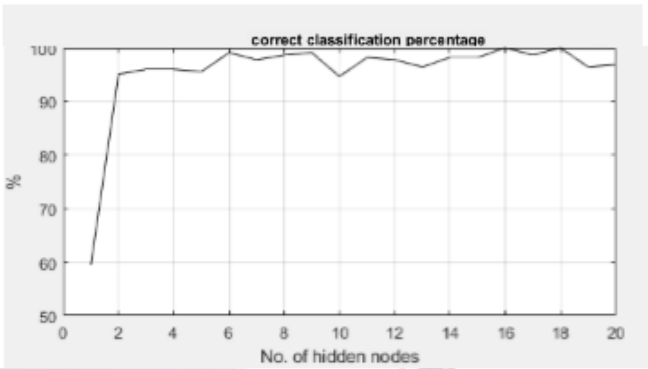

Fig 7. Percentage of correct classification at different nodes for MFCC coefficients.

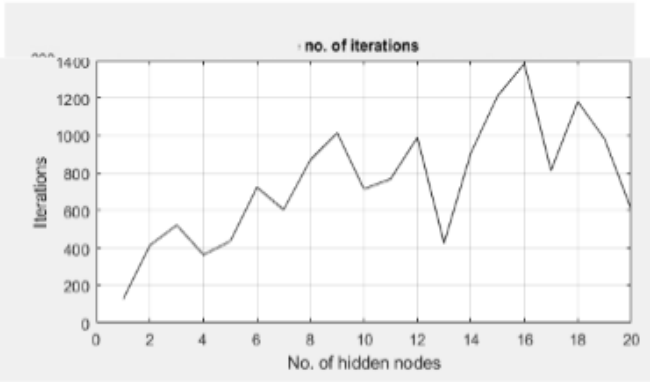

Fig 8. No. of iterations at every hidden node.

Results for MFCC coefficients

Column	Average Correct (%)	Average Iterations
1	59.5556	117
2	95.1111	562
3	96.8000	682
4	96.0000	340
5	95.5556	453
6	99.1111	720
7	97.7778	604
8	98.6667	868
9	99.1233	1014
10	94.4443	714
11	98.2222	763
12	97.7778	989
13	96.4444	425
14	98.2222	905
15	98.2222	1211
16	100.000	1301
17	98.4677	811
18	100.0000	1180
19	98.2300	921
20	95.2300	605

Table 1. Table shows Average correct recognition and average number of iterations.

Results for DWT coefficients

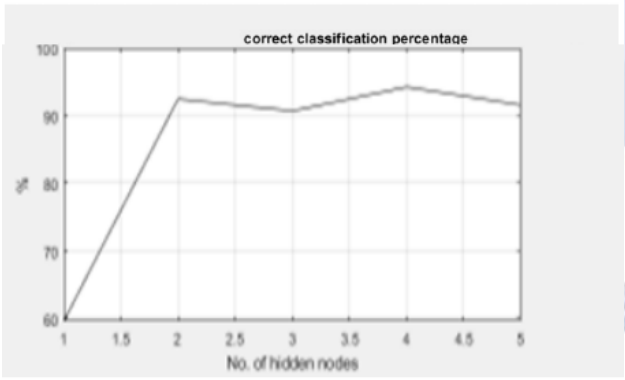

Fig 9.Percentage of correct classification at different nodes.

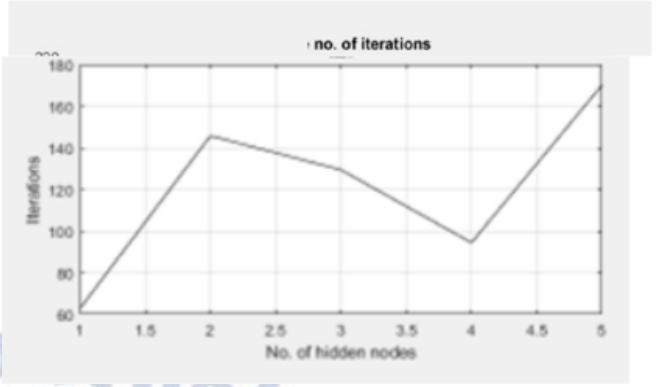

Fig 10. No . of iterations at every hidden node.

Column	Average Correct(%)	Average Iterations
1	59.5556	623
2	92.4444	145
3	90.6667	130
4	94.2222	95
5	91.5667	170

Table 2. Table shows Average correct recognition and average number of iterations.

- For training RBF with MFCC features time elapsed is 403 seconds.
- For training RBF with DWTfeatures time elapsed is 143 seconds.

Comparing results of both speaker systems

Speaker recognition system	Classification percentage	No.of iterations
MFCC-RBFNN	100	1180
DWT-RBFNN	94.22	95

Table3.Node at which maximum accuracy obtained.

8. CONCLUSION

From this paper it was concluded that with less number of iterations and with less amount of time for training RBF network with DWT coefficients achieved almost same accuracy as RBF network with MFCC coefficients which takes more number of iterations and training time.

FUTURE SCOPE

In this paper we carried out results by training our RBF network with Back propagation Algorithm. We can also increase our accuracy by introducing optimization algorithms for training our network. Optimization means where we will train our network iteratively until an

optimum or satisfying one is identified. We would like to continue further research on this project by employing one of the optimization algorithms called Harris Hawk Optimization Algorithm. HHO is a well known swarm based gradient-free optimization technique with many active time varying exploration and exploitation phases. Our results are expected to increase by using this algorithm.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Radial Basis Function Networks for Speaker Recognition by J.Oglesby and J.S.Mason Department of Electrical and Electronic Engineering University University college, SWANSEA,UK.

[2] SPEECH RECOGNITION USING RADIAL BASIS FUNCTION NEURAL NETWORK Dr.R.L.K.Venkateswarlu Professor and Head, Department of Information Technology Sasi Institute of Technology and Engineering, Tadepalligudem.

[3] An Efficient MFCC Extraction Method in Speech Recognition Wei HAN, Cheong-Fat CHAN, Chiu-Sing CHOY and Kong-Pang PUN Department of Electronic Engineering, The Chinese University of Hong Kong Hong Kong.

[4] DWT and MFCCs based Feature Extraction Methods for Isolated Word Recognition Mahmoud I. Abdalla Department of Electronics and Communications, Zagazig University, Egypt.

[5] Radial Basis Function Network Learning with Modified Backpropagation Algorithm Article in TELKOMNIKA Indonesian Journal of Electrical Engineering · February 2015.

[6] A. Mohamed, G. E. Dahl, and G. E. Hinton, "Acoustic Modeling using deep belief networks," IEEE Transactions on Audio, Speech and Language Processing, vol. 20, no. 1, pp. 504–507, 2012

[7] Wilson Burgos-GAMMATONE AND MFCC FEATURESIN SPEAKER RECOGNITION, Florida Institute of Technology, Melbourne, Florida ,November 2014.

[8] Speaker Recognition System Based on Wavelet Features and Gaussian Mixture Modelsby K.SajeerPaul Rodrigues International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-1, October 2019.

[9] Text-Independent Speaker Recognition Using Radial Basis Function Network Anton A. Yakovenko, Galina F. Malykhina Peter the Great Saint-Petersburg Polytechnic University, Polytechnicheskaya 21, 194021 St.Petersburg, Russia

[10] Improved Hidden Markov Model Speech Recognition Using Radial Basis Function Networks Elliot Singer and Richard P. Lippmann Lincoln Laboratory, MIT Lexington, MA 02173-9108, USA