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):194-199, 2022

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0804037
Available online at: http://www.ijmtst.com/vol8issue04.html

Speaker Recognition using RBF Neural Networks with Gammatone Frequency Cepstral Coefficients

Bairisetti Yaswanth Krishna | Chitipothu Sai Keerthana | Devalla Anusha | Bodepudi Hanuma Harshith

Department of Electronics and Communication Engineering, V. R & J. C College of Engineering, Chowdavaram, Guntur, India.
*Corresponding Author Email ID: yaswanth.bairisetti123@gmail.com

To Cite this Article
Bairisetti Yaswanth Krishna | Chitipothu Sai Keerthana | Devalla Anusha | Bodepudi Hanuma Harshith. Speaker Recognition using RBF Neural Networks with Gammatone Frequency Cepstral Coefficients. International Journal for Modern Trends in Science and Technology 2022, 8(04), pp. 194-199. https://doi.org/10.46501/IJMTST0804037

Article Info
Received: 12 March 2022; Accepted: 06 April 2022; Published: 10 April 2022.

ABSTRACT

The process of identifying the person by characteristics of their voice is called Speaker Recognition. Theadvancementof speakerrecognitionautomationprocessimprovestheinterfacebetweenmanandmachineinnumerous applications. It can be used to authenticate or verify the identity of a speaker as apart of security process. In fact, speech contains a great deal of information that allows todeterminegender,emotional stateandage ofthespeaker. Text-independent systems are most often used for speaker identification as theyrequire very little if any cooperation by the speaker. The speech features are extracted byusing Gamma tone Frequency cepstral coefficients (GFCC) technique. The proposed featureshave shown strong robustness in these challenging situations and they consistently performbetterthanthe well-knownMFCC. Feature vectors of spoken words wereappliedtoradialbasisneuralnetwork.The results show that the best accuracy willbe achieved byusingthe proposed SpeakerRecognitionsystemRBFNN-GFCC.

KEYWORDS:MFCC, GFCC, radial basis neural networks.

1. INTRODUCTION

There has been an increasing interest in using biometric traits to recognize or validate a person's identification in recent years. One of the biometric identifiers is a person's voice, which is believed to be unique to that person and not reproduced by others. As a result, a speaker recognition system can provide a better level of non-intrusive security than traditional security processes by using identifying elements in an individual's voice. Aside from physical distinctions, each speaker has a distinct way of speaking, which includes the usage of a unique accent, rhythm, intonation style, pronunciation pattern, word choice, and so on. Modern speaker identification systems employ a combination of these factors to attempt to recognize a speaker.

Speaker Recognition is divided into Feature Extraction & Feature Matching.Features of speech samples of different speakers are extracted by using feature extraction techniques such as Mel Frequency Cepstral Coefficients(MFCC) and Gamma tone frequency cepstral coefficients(GFCC). Feature Matching involves Training and Testing phase.

The text-dependency or text-independence of speaker recognition systems is an essential distinction This speaker identification system is said to be text-dependent if a person is needed to use the same text during the training and recognition sessions. The test speaker in text-independent speaker recognition has no prior knowledge of the contents of the training phase and can talk anything.

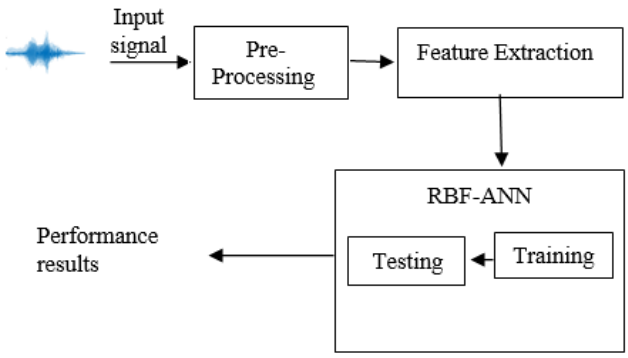

Fig. 1:Block diagram of speaker recognition system

Block diagram of speaker recognition system is shown in Figure 1. The features which are used for speaker recognition are MFCC and GFCC which contains speaker specific information. For training and testing RBFNN(Radial Basis Function Neural Networks) are used.

2. FEATURE EXTRACTION

The practice of extracting a small quantity of data from the speaker's voice signal that can later be utilized to represent the speaker is known as feature extraction. MFCC and GFCC are two feature extraction approaches employed in this paper.

A. MFCC

The features of speech samples are extracted by using MFCC. Mel Frequency Cepstral Coefficients (MFCC) are widely used features in speaker recognition. MFCC feature extraction process is shown in Figure 2.

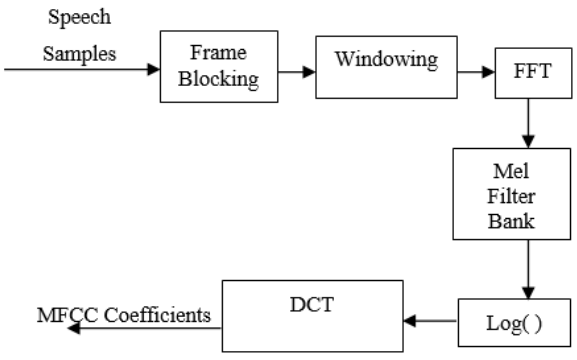

Fig.2:MFCC Block Diagram

1. Frame Blocking

Speech signals are processed in short time periods because they are reasonably steady in short intervals of time. They are divided into frames ranging in size from 20 to 30 milliseconds. Each frame has a predetermined overlap with the previous frame. The purpose of the overlapping technique is to make the transition from one frame to the next as seamless as possible.

2.Windowing

The second step is to window all of the signal's individual frames to reduce unnatural signal discontinuities at the start and end of each frame. This is done to find a segment of the signal that can be presumed to be stationary.

The window function used here is Hamming Window and it is represented as

$$\boldsymbol{w}[\boldsymbol{n}] = 0.54 - 0.46\cos(2\pi n/N\text{-}1);\ 0 \le \boldsymbol{n} \le N - 1 \qquad (1)$$

where w[n] is hamming window, N is number of frames.

3. FFT

A better procedure known as the fast Fourier transform, or FFT, can be used to calculate DFT. An input signal with a length of 2M for some M power of two is a condition for efficient FFT. In practice, the input signal is first zero-padded to the next largest power of two, and the zero-padded signal is used as the FFT's input. For example, if the signal is 230 samples long, it is zero padded to N = 256 samples long, allowing the FFT to be computed. The addition of zeros to the beginning or end of the signal has no effect on the DFT result.

4. Mel Filter Bank

A filter bank spaced uniformly on the Mel scale, is one method of simulating the subjective spectrum The Mel frequency scale is made up of a series of linearly spaced filter banks up to a cutoff frequency (about 1 kHz), followed by logarithmic filter bank spacing from there to the maximum frequency. This filter bank features a triangle band pass frequency response, with a constant mel frequency interval determining the spacing and bandwidth. Mel filter banks are not evenly placed along the frequency axis, therefore there are more filters in the low frequency areas and fewer in the high frequency areas. Formula to compute the mels for a given frequency f in Hz:

$$\text{Mel}(f)=2595*\log10(1+f/700) \qquad (2)$$

5. DCT

The log of the mel spectrum must then be converted back to time in the final step. The mel frequency cepstrum coefficients are the outcome (MFCCs). For the current frame analysis, the cepstral representation of the speech

spectrum provides a good representation of the signal's local spectral features. The Discrete Cosine Transform is used to transform the mel spectrum coefficients.

B. GFCC

To overcome the noise robustness problem in ASR, GFCC technique is used as sensitivity to additive noise is one of the major disadvantages of MFCC. Features of speech samples of different speakers are extracted by using Gamma tone frequency cepstral coefficients(GFCC). The Gammatone Frequency Cepstral Coefficients are auditory feature based on a set of Gammatone Filter banks. GFCC feature extraction process is shown in Figure 3.

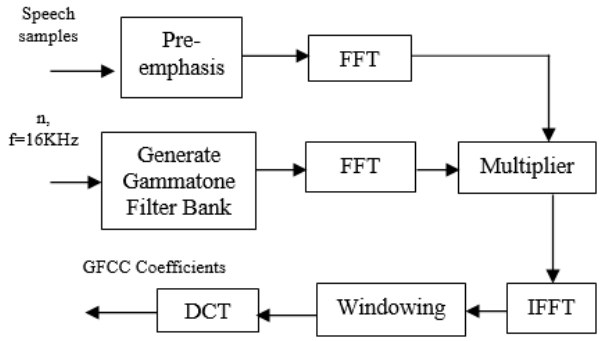

Fig.3:GFCC Block Diagram

1. Pre-emphasis

The speech will loss the information at high frequency, thus it needs the pre- emphasis process in order to compensate the high frequency loss. In the speech signal model, the pre-emphasis is a first order high pass filter. The transform function of pre-emphasis can be defined as:

$$H(z) = 1 - \alpha z^{-1} \quad (1)$$

where parameter $\alpha$is usually between 0 .94 and 0 .97

2. Gammatone Filter Bank

The Gammatone filters are intended to mimic the human auditory system's processes. The following is the definition of a Gammatone filter using fc as the centre frequency:

$$g(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi f_c + \varphi) \quad (2)$$

where φ is the phase but is usually set to zero, the constant a control the gain and the order of the filter is defined by the value n which is typically set to a value less than 4. The factor b is defined as:

$$b = 25.17\left(\frac{4.37 f_c}{1000} + 1\right) \quad (3)$$

The different stages ofits GFCC computation have similarities with those MFCC section such as FFT, Windowing and DCT.

3. SPEAKER MODELLING

The necessity for speaker recognition in the scientific and engineering fields is a much bigger problem known as feature matching. The true goal of feature matching is to classify objects into a variety of classes and categories. The approaches are used to classify patterns that are essentially items of interest from sequences of audio vectors. The speaker with the lowest matching score is an unknown speaker. The speaker modelling technique Radial Basis Function Neural Network is utilized for categorization.

A. RBF

Radial Basis Functions (RBF) are a type of feed-forward artificial neural network that has at least three layers of neurons: an input layer, a hidden layer, and an output layer, each of which implements a radially activated function. An RBF network's input is nonlinear, while the output is linear. Figure 4 depicts an RBF network with D inputs, M hidden units, and K outputs. The output layer acts as a linear combiner, calculating the weighted sum of the hidden units outputs. The k output of an RBF neural network has the form:

$$f_k(Y) = w_{0k} + \sum_{j=1}^{M} w_{jk}\,\phi_j(Y) \quad (1)$$

$j=1,2,3,\ldots,M$ and $k=1,2,\ldots,K$where w_{jk} - weights of the network. For an RBF network the activation function is:

$$\phi_j(Y) = exp\left\{-\frac{1}{2.\sigma_j^2}\|Y - c_j\|^2\right\} \quad (2)$$

where $\|.\|$denotes the Euclidean distance. In $\Phi_j$is activation function, $Y = \{y_t, t = 1, \ldots, T\}$ is the input vector of length T and dimension D, c_j - function centers, σ_j- the function width.

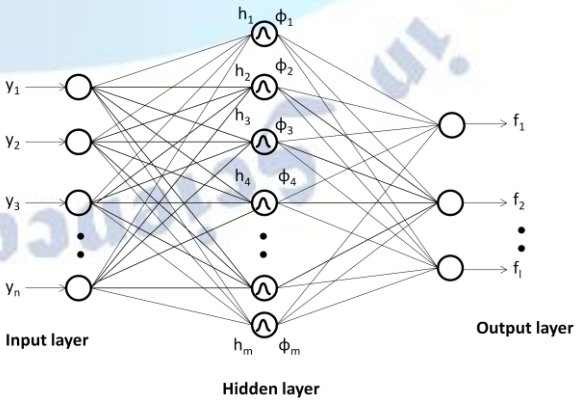

Fig.4:Radial Basis Function Neural Network

The neural network models the underlying function of a particular mapping through training. In the input data space, the buried layer neurons represent a sequence of centres. Each of these centres has a Gaussian-like activation function. The distance between the provided input vector and the centre determines the activation. The lower the activation, the farther the vector is from the centre, and vice versa. An unsupervised k-means clustering technique is used to generate the centres and their widths. The weights and biases of the hidden layer are formed by the weights and biases established by this procedure, which stay unaltered until the clustering is completed.Although RBF networks have both a supervised and unsupervised learning component, they are mostly employed in supervised applications. To discover neuron centres, widths, and amplitude, fully supervised training is required. Used a training set of data samples for which the appropriate network outputs are known in a supervised application. The network parameters are chosen in this example to minimize a cost function.

Gradient Descent Algorithm

The RBF centres and other network parameters go through a supervised learning procedure. A gradient descent approach, which is a generalization of the Least Mean Square (LMS) algorithm, is the most practical for RBF network learning. The family of RBF networks is broad enough to uniformly approximate any continuous function on a compact set and consists of functions represented by

$$F(x) = \sum_{i=1}^{m} a_i \, \varphi(w_i^T x) \qquad (3)$$

where m - the number of neurons in the first layer, a_i, $\mathbf{w}_i$- coefficients of neural network, $\phi(.)$ - activation function. As the activation function in the expression a family of exponential distributions with the shape parameter α is proposed. Calculating the mean square error of approximation of the mixture of multidimensional sampling distributions:

$$\varepsilon = \frac{1}{2}\sum_{j=1}^{N} e_j^2 \qquad (4)$$

where N is the size of the training sample. Error signal defined by:

$$e_j = d_j - \sum_{i=1}^{M} w_i f(x_j - m_i) \qquad (5)$$

where d_j- data. Neural network training procedure is performed incrementally using gradient descent algorithm.Changing weights on the next step:

$$w_i(n+1) = w_i(n) - \eta_1 \frac{\partial \varepsilon(n)}{\partial w_i(n)}, i = 1, \ldots, m_i \qquad (6)$$

$$\frac{\partial \varepsilon(n)}{\partial w_i(n)} = \sum_{j=1}^{N} e_j(n) f(x_j - m_i(n)) \qquad (7)$$

Adjustment of the position of the centers:

$$t_i(n+1) = t_i(n) - \eta_2 \frac{\partial \varepsilon(n)}{\partial t_i(n)}, i = 1, \ldots, m_i (8)$$

$$\frac{\partial \varepsilon(n)}{\partial t_i(n)} = \alpha \omega_i(n) \sum_{j=1}^{N} e_j(n) f'\left(x_j - m_i(n)\right) \times \Sigma^{-1}(x_j - ti(n))\alpha - 1 \quad (9)$$

Adjustment of distribution width:

$$\Sigma_i^{-1}(n+1) = \Sigma_i^{-1}(n) - \eta_3 \frac{\partial \varepsilon(n)}{\partial \Sigma_i^{-1}(n)}, i = 1, \ldots, m_i (10)$$

$$\frac{\partial \varepsilon(n)}{\partial \Sigma_i^{-1}(n)} = -\alpha w_i(n) \sum_{j=1}^{n} e_j(n) f'\left(x_j - m_i(n)\right) Q_{ij}(n) \quad (11)$$

$$Q_{ij}(n) = \left(x_j - m_i(n)\right)^{\alpha - 1} (x_j - m_i(n))^T \qquad (12)$$

Adjustment of the PDF shape parameter:

$$\alpha_i(n+1) = \alpha_i(n) - \eta_2 \frac{\partial \varepsilon(n)}{\partial \alpha_i(n)}, i = 1, \ldots, m_i \qquad (13)$$

$$\frac{\partial \varepsilon(n)}{\partial \alpha_i(n)} = 2\omega_i(n) \sum_{j=1}^{N} e_j(n) f\left(x_j - m_i(n)\right) \alpha^{-1} + f'\left(x_j - minax - m(n)\lambda\Sigma\alpha - 1\right. \qquad (14)$$

4. RESULTS AND DISCUSSION

The speaker recognition is done for three speakers. The below figures shows the result when MFCC features are used for training and testing .

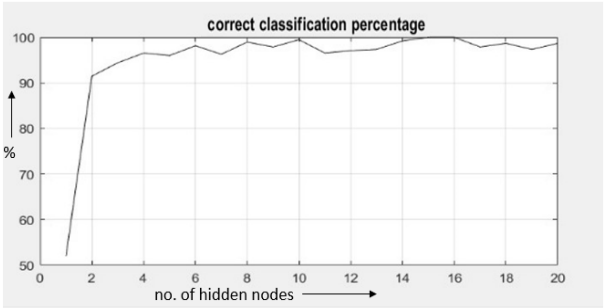

Fig.5 : Percentage of correct classification for MFCC-RBFNN Speaker Recognition System

Figure 5 shows the percentage of correct classification for RBFNN Speaker Recognition System when hidden nodes are varied.

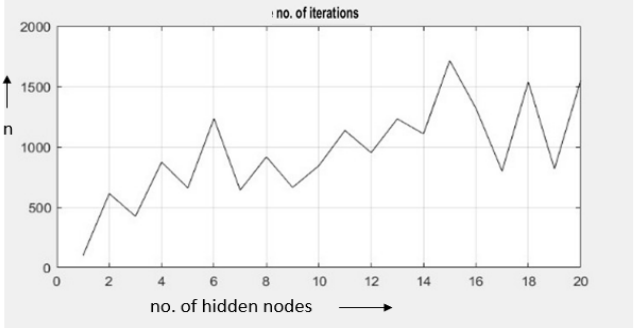

Fig.6 : No. of Iterations for MFCC-RBFNN Speaker Recognition System

Figure 6 shows the number of iterations required to recognize the speaker for RBFNN Speaker Recognition System when hidden nodes are varied.

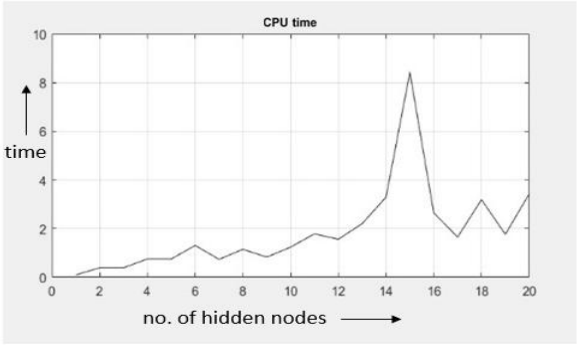

Fig.7 : CPU time graph for MFCC-RBFNN Speaker Recognition System

Figure 7 shows the CPU time required to recognize the speaker for RBFNN Speaker Recognition System when hidden nodes are varied.

Table 1. Correct classification percentage, CPU time and Number of iterations for MFCC-RBFNN Speaker Recognition System

Hidden nodes	Correct classification percentage(%)	No. of Iterations	CPU time (sec)
1	52.0000	100	0.1000
2	91.4667	612	0.3906
3	94.4000	424	0.3906
4	96.5333	873	0.7562
5	96.0000	658	0.7531
6	98.1333	234	1.3094
7	96.2667	641	0.7375
8	98.9333	917	1.1500
9	97.8667	663	0.8281
10	99.4667	843	1.2406
11	96.5333	1135	1.7844
12	97.0667	951	1.5625
13	97.3333	1232	2.2062
14	99.2000	1107	3.2398
15	100.0000	1714	8.4219
16	100.0000	1319	2.6469
17	97.8667	799	1.6531
18	98.6667	1538	3.1906
19	97.3333	817	1.7625
20	98.6667	1550	3.4281

The table shows the correct classification percentage, CPU time and Number of iterations for RBFNN Speaker Recognition System.

The below figures shows the result when GFCC features are used for training and testing .

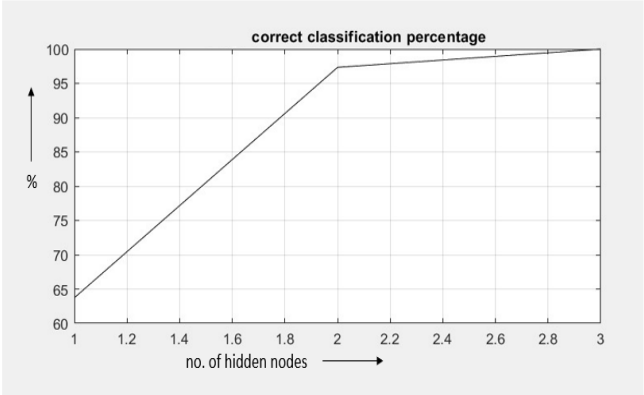

Fig.8 : Percentage of correct classification for GFCC-RBFNN Speaker Recognition System

Figure 8 shows the percentage of correct classification forRBFNN Speaker Recognition System when hidden nodes are varied.

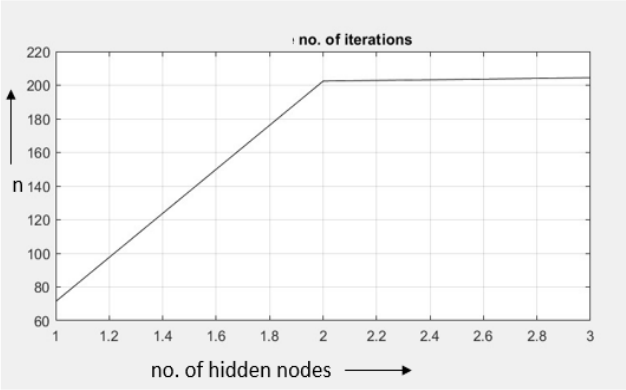

Fig.9 : No. of Iterations for GFCC-RBFNN Speaker Recognition System

Figure 9 shows the number of iterations required to recognize the speaker for RBFNN Speaker Recognition System when hidden nodes are varied.

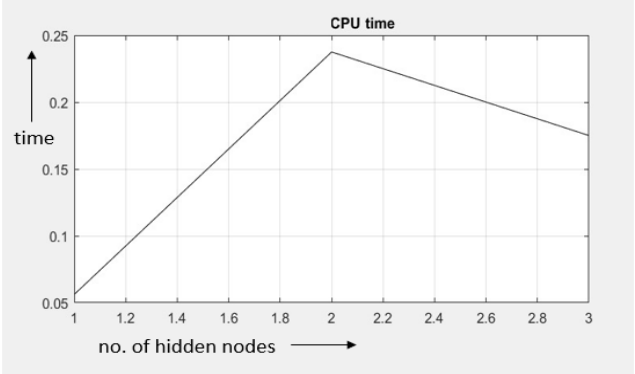

Fig.10 : CPU time graph for GFCC-RBFNN Speaker Recognition System

Figure 10 shows the CPU time required to recognize the speaker for RBFNN Speaker Recognition System when hidden nodes are varied.

Table 2. Correct classification percentage, CPU time and Number of iterations for RBFNN Speaker Recognition System

Hidden nodes	Correct classification percentage(%)	Number of Iterations	CPU time (sec)
1	63.7333	71	0.0563
2	97.3333	202	0.2375
3	100.00	204	0.1750

Table 3. Performance Measures for RBFNN Speaker Recognition System

Speaker recognition system	Correct Classification percentage	CPU time(sec)	no. of iterations
MFCC-RBFNN (hidden nodes=15)	100	8.42	1714
GFCC-RBFNN (hidden nodes=3)	100	0.175	204

5. CONCLUSION

It can be observed from above results when MFCC coefficients are used for training and testing the RBF network, the CPU time is 8.42 seconds, number of iterations are 714 and classification percentage is 100. When GFCC coefficients are used to train the RBF network the CPU time is 0.175seconds, number of iterations are 204 classification percentage is 100. So, we can conclude that system complexity is reduced by using the RBFNN-GFCC Speaker Recognition system.

Conflict of interest statement

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

REFERENCES

[1] Shaik Shafee and Dr.B.Anuradha- Speaker Identification and Spoken word Recognition In Noisy Background using Artificial Neural Networks, International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) – 2016.

[2] Murad Hossain, Mahmuda Asrafi, and Boshir Ahmed- A real time speaker identification neural network, Conference Paper · January 2008DOI: 10.1109/ICCITECHN.2007.4579414· Source:EEE Xplore

[3] Kritagya Bhattarai, P.W.C.Prasad , Abeer Alsadoon, L.Pham and A.Elchouemi – Experiments on the MFCC application in Speaker Recognition using Matlab, Seventh International Conference on Information Science and Technology, Da Nang, Vietnam; April 16-19,2017.

[4] D. O'Shaughnessy, Speech Communications: Human and Machine. IEEE Press, New York, 2000.

[5] X. Zhao, Y. Shao, and D. Wang, "Casa-based robust speaker identification," IEEE Transactions on Audio, Speech and Language Processing, vol. 20, no. 5 ,pp.1608–1616, 2012.

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

[7] Y. Shao, Z. Jin, D.Wang, and S. Srinivasan,"An auditorybased feature for robust speech recognition,"in Proc. ICASSP' 09, 2009, pp. 4625–4628.

[8] Wilson Burgos- GAMMATONE AND MFCC FEATU-ES IN SPEAKER RECOGNITION, Florida Institute of Technology, Melbourne,FloridaNovember 2014.