A SIMPLE SPEAKER IDENTIFICATION APPROACH BY USING MEL-FREQUENCY CEPSTRUM COEFFICIENTS

Dr. Kanaka Durga Returi[1]., Dr. Radhika, Y[2] and Dr.Vaka Murali Mohan[1*]

1. *Principal & Professor of CSE, Malla Reddy College of Engineering for Women, Medchal, Hyderabad, Telangana, India.*
Ph. No: +91-9966023477, 9885477515
E- Mail: murali_vaka@yahoo.com
2. *Director Academic Affairs & Professor of CSE, GITAM University, Rushikonda, Visakhapatnam, Andhra Pradesh, India.*

ABSTRACT

A simple speaker identification approach by using mel-frequency cepstrum coefficients have been reported in this paper. MFCC stay usually deliberate through Fourier transform of a space extracted signal and plotting effects of range achieved in mel scale by utilizing triangular intersecting windows. This paper also explained the process of frame blocking the constant speech gesture remains choked into N samples of frames and M samples of adjacent frames (M < N). The conclusions of every frame level to attach through every supplementary and this processing is named as *Windowing*. Generally *Hamming window* is utilized to achive good results. The ensuing treating phase exists “Fast Fourier Transform”, every structure of *N* illustrations transforms time into frequency field. Filter bank method simulating particular spectrum and equally moved on mel scale. The filter group takes three-sided group authorization frequency reply with respect to space and group is measured.

CHAPTER 1: INTRODUCTION

The main goal is identification of speaker, which contains corresponding speech signal of an indefinite speaker through recognized speaker named as ‘target’ and kept in a database. This method is trained through a number of speakers then it can identify the speaker with the utilization of the database. Speaker identification is method of defining with recorded speaker make available on a specified speech. Simultaniousley speaker authentication method declining or accepting distinctiveness privilege of presenter. Voice is utilized as an important applications to authorize the characteristics of a speaker.

Feature extraction and matching is the key method in the process of identification of a speaker. These transactions through extraction of significant features from a speech signal such as pitch or frequency. Speaker identification techniques are distributed as text-independent and text-dependent techniques. Text-independent method register features of speech which illustration of what speaker saying in this models. Text-dependent model recognition based on speaker's identity like specific phrases, passwords, card no., PIN No., etc.

The core objective of this effort is speaker identification, which contains identical speech indication of an unidentified speaker through that of a identified speaker named as 'target' deposited in a databank. The structure is accomplished through a quantity of utterers and it can distinguish the speaker constructed on the databank. The speaker identification exists the development of response which recorded speaker affords a specified speech.

2. LITERATURE REVIEW

The investigators projected so many works, such as Tobias Herbig et al [1] reported the models for identification and recognition of speaker. Masaki Naito et al [2] proposed models for identification speaker and speech by using vocal zone size. Vimala C., V. Radha [3] presented an isolated system for recognition of speaker independent speech for Tamil language through resources of "Hidden Markov Model (HMM)". Yongwon Jeong [4] presented a variation in "hidden Markov model" established system for identification of speech to obtain speaker and its noise surroundings. Çetingül, H, E et al [5] presented a model for recognition of speaker and speech to facilitate audio, texture and motion. Mats Blomberg [6] described synthetic generation system for speech recognition through prototypes and symbolic transformation. Sadaoki Furui [7] introduced latest developments in the speaker recognition with VQ and HMM model techniques. Mike Talbot [8] reported the interface design of the speech recognisers through matching utterances to stored voice data. Sadaoki Furui [9] introduced a model for utilizing pitch information, adaptation techniques, HMM model, neural networks training algorithms for speech recognition. Joseph Picone [10] introduced Hidden Markov Model based model system in view of spectral and duration information.

Howard, C, N et al [11] developed a model and analyzed the system performance by speech recognition devices through vocabulary, user's speech and algorithm. Renato, D, M et al [12] presented a network

based paradigm for speech recognition and description of speech properties by using Markov models. John, R, H et al [13] developed a model that recognizes the speech data which was recorded within a single channel. Ron, J, W et al [14] reported a model for segregation of single channel combination of speech model with hidden Markov models and factorial HMM model. Kai, Fu, Lee et al [15] introduced a technique and described the improvements of the Hidden Markov Model used in SPHINX for speaker appreciation.

Kanaka Durga Returi et al [16] presented a method of Speaker identification with respect to WA and SVM. Kanaka Durga Returi 17] reported a relative method of changed Lines designed for Presenter Appreciation. Kanaka Durga Returi and Y. Radhika [18] developed an ANN Model with WA.

3. MFCC

Mel-Frequency Cepstrum defined as symbol of minimum range sound influence. MFC is mutually created with coefficients and are named as MFCCs. These are derivative of cepstral illustration of the acoustic clip and named as nonlinear "spectrum-of-a-spectrum". The transformation in MFC among cepstrum and mel-frequency cepstrum, the occurrence posses exist similarly spread out on mel scale, which approaches the humanoid audio structure's answer thoroughly used in linearly-spaced occurrence groups in regular cepstrum.

The Mel scale deals by means of pitches and it is a qualified portion of frequency definite by auditors to differentiate one from the other. The Mel scale well-defined by associating 1000 Hz tone, 40 dB beyond auditor's edge through a pitch of 1000 mels. Mel scale is measured based on human ear by way of a intelligence observation to sound. It illustrates how the humanoid ear states the pitches used for various frequencies. The term *mel* derives as of the term *melody* to designate the scale through pitch differences.

The relation between fhertz and 'm' mel is

$$m = 1127.0105\, Log_e (1 + f/700)$$

and the inverse

$$f = 700\,(e^{m/1127.0105} - 1)$$

3.1 Cepstrum

Cepstrum exists as collective alter utilized towards information improvement from a individual's speech gesture and can exist utilized signal excitation and transfer function. Outcome of signal is considering Fourier transform of decibel spectrum.

$$Cepstrum\,of\ Signal = FT[Log\{FT(the\,windowed\,signal)\}]$$

The cepstral analysis is utilized as a simple speaker identification by using speech signal is named as "cepstral transform" and is represented as Figure 1.

Figure 1: 'Cepstrum' Flow Chart

Cepstrum can exist as data approximately degree of transformation in various spectrum groups.

3.2 Calculation of MFCCs:

MFCCs are usually deliberate through Fourier transformation of space extracted and plotting effects in mel scale by utilizing triangular intersecting windows. Logs influences in every mel occurrences exist as Shortest Cosine Transformation. The MFCCs stand as amplitudes of subsequent spectrum. The general step by step process for the calculation of MFCCs is represented in Figure 2.

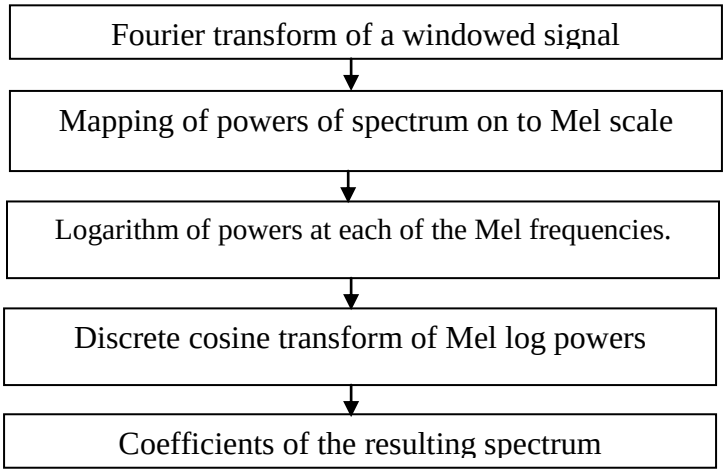

Figure 2. MFCCs Calculation

4. MEL-FREQUENCY CEPSTRUM COEFFICIENTS (MFCC)

In this process the speech waveform is converted into lower frequency dat for advanced investigation and named as front end of the signal processing. Generally speech signal is known as a gently timed variable and is termed as quasi stationary and represented as Figure 3.

The figure reveals that in a small period of time in between 5 to 10msec the signal is stationary. Though above extended periods of time the signal specific transformation towards replicate the dissimilar speech resonances existence pronounced. So, small period spectral investigation is the greatest method to describe the speech signal.

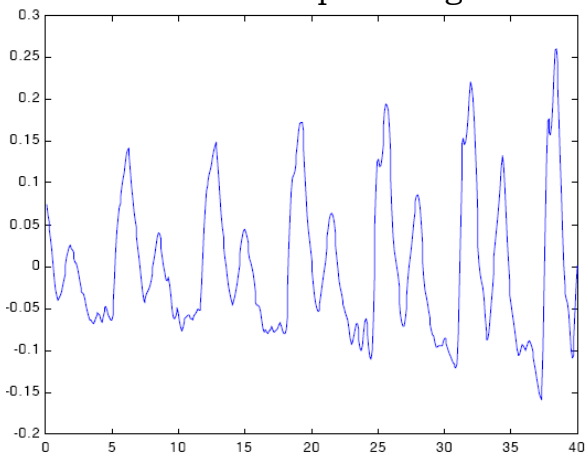

Figure 3: Speech zSignal

A extensive collection of potentials occur intended for parametrically on behalf of communication gesture recognition of speaker by way of Linear Prediction Coding, MFCC, etc.... . MFCC exists as feasibly the finest recognized and use friendly method.

MFCC's are established on the identified dissimilarity of the humanoid ear's acute group through incidence, sieves spread out directly at little rates and logs at maximum occurrences must existed to capture phonetically significant speech features. MFCCs are calculated by ustilizing the mel-frequency scale and is in the range of <1000Hz for a linear frequency and >1000Hz for logarithmic design.

4.1 MFCC Processing

The general block diagram of a MFCC mainframe represented as Figure 4. Speech response usually noted as selection frequency >10000 Hz. These collected rate existed properties by utilizing analog to

numerical translation. The captured signals occurrences towards 5 kHz and maximum strength produced by persons. MFCC processor defined as mimic performance.

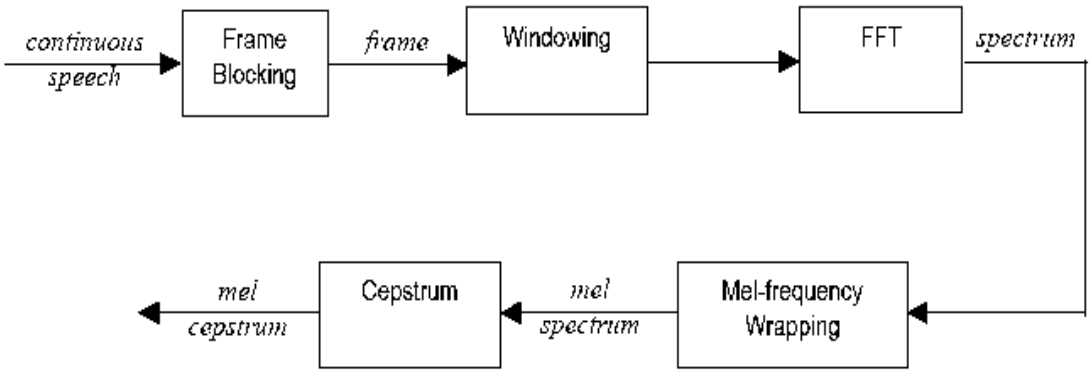

Figure 4: MFCC processor Block diagram

4.1.1 Frame Blocking

The hypothesis remains finished in long interval, waveform remains dynamic, concluded with a adequately short period pause approximately around 10-30msec and measured as static. The degree at which the range of speech wave modifications is openly dependant proceeding the frequency measure of speech articulators. It is incomplete through physical limitations, maximum speech investigation methods function at consistently spaced period breaks or structures of representative period 10 to 30 msec.

The process of frame blocking communication gesture remains choked into N samples of frames and M samples of adjacent frames (M < N). The first frame contains 1st N illustrations. The 2nd frame initiates M illustrations subsequently with 1st frame and correspondences is through N - M illustrations. Correspondingly 3rd frame creates 2M illustrations subsequently with 1st frame similarities is by N - 2M illustrations. The procedure remains pending altogether the communication remains accounted used within one or other frames. The standard characteristic at 30 msec windowing and simplify 2FFT are N = 256 and M = 100.

4.1.2 Windowing

In this following phase processing stands towards window every specific edge consequently reduce signal breaks starting to end of every frame. The idea now reduce spectral alteration utilizing space towards

match gesture to zero starting to end of every frame. In other phase, while implementing Fourier Transform signal duplications and end of one edge ensures efficiently through establishment of next unique. It presents specific faults consistent intermissions. The conclusions of every edge level attach through every supplementary and this processing is named as *Windowing*.

In this method, specified signal is multiplied with *Window Function*. Figure 5 and Figure 6 illuminate the perception of windowing with time and frequency domain respectively.

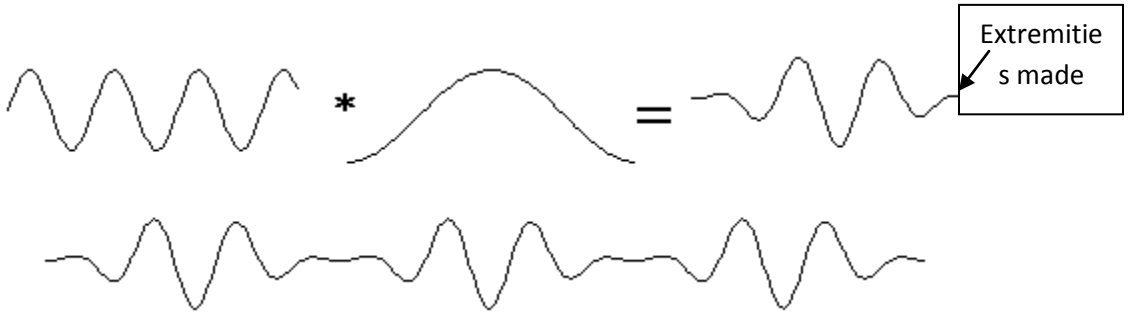

Figure 5: Windowing in time domain

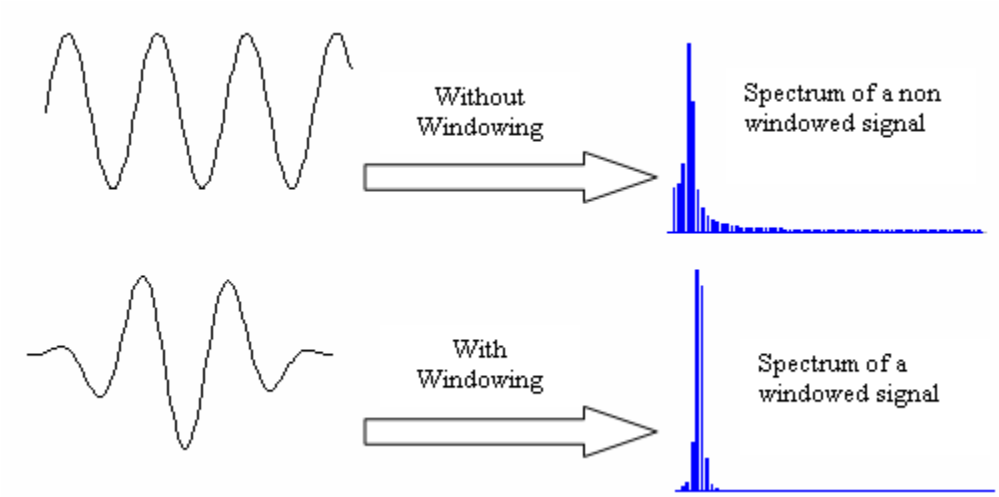

Figure 6: Windowing in frequency domain

If we define the window as

$w(n), 0 \le n \le N-1$

where $N = No.of\ samples\ in\ every\ frame$

The result of windowing is the signal

$y_1(n) = x_1(n)\, w(n), \quad 0 \le n \le N-1$

It remains chosen to utilize a 'soft windowing' procedure, efficiently matches ends of communication division Zero. Several 'soft windows' are utilized, generally *Hamming window* is utilized to achive good results and has formula..

$w(n) = 0.5384 - 0.4616\ Cos(2\pi n / N - 1)$

where

$0 \le n \le N-1$

The graphical illustrations of the hamming window are represented as Figure 7.

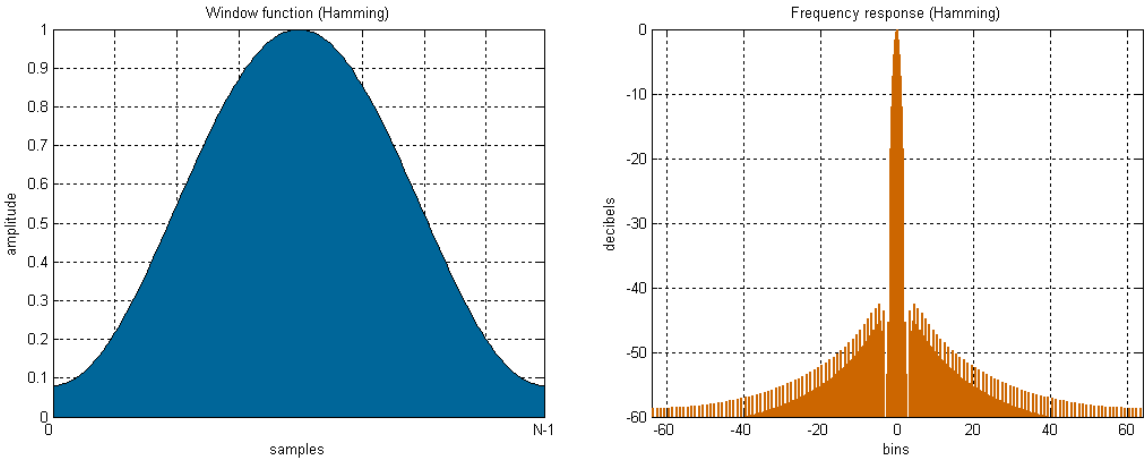

Figure 7: Hamming window

4.1.3 Fast Fourier Transform (FFT)

To ensuing treating phase Fast Fourier Transform with every frame of N illustrations transforms time to frequency field. The FFT is a fast procedure to apply Discrete Fourier Transform which is distinct arranged the set of N samples $\{x_n\}$, as follow:

$$X_n = \sum_{k=0}^{N-1} x_k \, e^{-2\pi jkn/N},$$

$n = 0,1,\ 2,.....N-1$

$j = \text{Im}\,aginary\,Unit$

$j = \sqrt{-1}$

$X_n = Complex\ Numbers$

$The\ resulting\ sequence\{X_n\}$

$the\ Zero\ \ frequency\ at\ n = 0$

$Positive\ \ frequencies\ \ 0 < f < F_s/2$

$Correspondingto\ values\ \ 1 \le n \le \frac{N}{2} - 1$

$While\ Negative\ \ frequencies\ -\frac{F_s}{2} < f < 0$

$Correspond\ to\ \frac{N}{2} + 1 \le n \le N - 1$

$F_s = Sampling\ Frequency$

The result after this step is referred as Spectrum or Periodogram.

4.1.4 Mel-frequency Wrapping

Psychophysical trainings must presented that human observation of the occurrence fillings of noises aimed at communication gestures organizes not in linear gauge.

On behalf of every quality through definite frequency, *f*, calculated in Hz, individual pitch calculated in 'mel' scale. *Mel-frequency* measure remains a direct rate rang < 1000 Hz and Log range >1000 Hz. Estimated formulation calculate mels on behalf of specified frequency:

$$mel(f) = 2595 * Log_{10}(1 + f/700)$$

Filter bank method is simulating the particular spectrum and equally moved arranged mel scale. Filter group takes a three-sided group authorization frequency reply with respect to space and group is measured in mel incidence. In general the quantity of mel spectrum constants *K* =12 or 20. Here the filter bank is functional in the frequency area; hence it basically quantities to attractive triangle-shape windows on spectrum and represented in Figure 8.

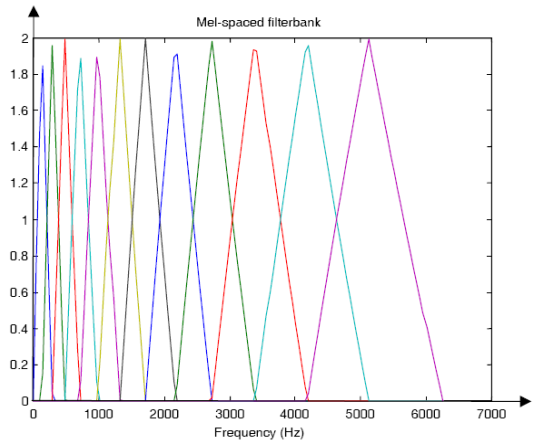

Figure 8: Mel Filter Bank (K=12)

5. CONCLUSIONS

A Method of MFCC for speaker identification have been presented in this paper. MFCC's are established on the identified dissimilarity of the humanoid ear's acute bandwidths through occurrence, sieves spread out directly at minimum rates and log at maximum rate must existed significant features of speech. MEL-FREQUENCY CEPSTRUM COEFFICIENTS (MFCC) are usually deliberate through Fourier transformation extract signal and plotting influences range achieved in mel scale by utilizing triangular intersecting windows. In the process of frame blocking the constant speech gesture remains choked into N samples of frames and M samples of adjacent frames (M < N). The conclusions of every frame level to attach through every supplementary and this processing is named as *Windowing*. Generally *Hamming window* is utilized to achive good results.

6. REFERENCES

1. Tobias Herbig., Franz Gerl., Wolfgang Minker "Self-learning speaker identification for enhanced speech recognition" Computer Speech & Language, Volume 26, Issue 3, 2012, pp 210–227.
2. Masaki Naito., Li Deng., Yoshinori Sagisak "Speaker clustering for speech recognition using vocal tract parameters" Speech Communication, Volume 36, Issues 3–4, Mar 2002, pp 305–315.
3. Vimala C., V. Radha "Speaker Independent Isolated Speech Recognition System for Tamil Language using HMM" Procedia Engineering, Volume 30, 2012, pp 1097–1102.

4. Yongwon Jeong “Joint speaker and environment adaptation using TensorVoice for robust speech recognition” Speech Communication, Volume 58, March 2014, pp 1–10.
5. H.E. Çetingül., E. Erzin., Y. Yemez., A.M. Tekalp “Multimodal speaker/speech recognition using lip motion, lip texture and audio” Signal Processing, Volume 86, Issue 12, 2006, pp 3549–3558.
6. Mats Blomberg “Adaptation to a speaker's voice in a speech recognition system based on synthetic phoneme references” Speech Communication, Volume 10, Issues 5–6, 1991, pp 453–461.
7. Sadaoki Furui “Recent advances in speaker recognition” Pattern Recognition Letters, Volume 18, Issue 9, September1997, pp 859–872.
8. Mike Talbot “Adapting to the speaker in automatic speech recognition” International Journal of Man-Machine Studies, Volume 27, Issue 4, October 1987, pp 449–457.
9. Sadaoki Furui “Recent advances in speech recognition technology at NTT laboratories” Speech Communication, Volume 11, Issues 2–3, 1992, pp 195–204.
10. Joseph Picone “Duration in context clustering for speech recognition” Speech Communication, Volume 9, Issue 2, April 1990, pp 119–128.
11. Howard C. Nusbaum., David B. Pisoni “Automatic measurement of speech recognition performance: a comparison of six speaker-dependent recognition devices” Computer Speech & Language, Volume 2, Issue 2, 1987, pp 87–108.
12. Renato De Mori., Regis Cardin., Ettore Merlo., Mathew, P., Jean Rouat “A network of actions for automatic speech recognition” Speech Communication, Volume 7, Issue 4, 1988, pp 337–353.
13. John R. Hershey., Steven J. Rennie., Peder A. Olsen., Trausti T. K “Super-human multi-talker speech recognition: A graphical modeling approach” Computer Speech & Language, Volume 24, Issue 1, 2010, pp 45–66.
14. Ron J. Weiss., Daniel P.W. Ellis “Speech separation using speaker-adapted eigenvoice speech models” Computer Speech & Language, Volume 24, Issue 1, 2010, pp 16–29.
15. Kai-Fu Lee., Hsiao-Wuen Hon., Mei-Yuh Hwang., Xuedong Huang “Speech recognition using hidden Markov models: A CMU

perspective" Speech Communication, Volume 9, Issues 5–6, Dec 1990, pp 497–508.

16. Kanaka Durga Returi., Y. Radhika., Vaka Murali Mohan "A Novel Approach for Speaker Recognition By Using Wavelet Analysis and Support Vector Machines" 2nd International Conference on Computer and Communication Technologies - IC3T 2015, during July 24 -26, 2015 at CMR Technical Campus, Hyderabad, Telangana, India (Technically co-sponsored by CSI Hyderabad Section) IC3T 2015, Volume 1, pp 163-174. (ISBN: 978 – 81 – 322 – 2517 – 1).
17. Kanaka Durga Returi., Vaka Murali Mohan and Praveen Kumar, L "A Comparative Study of different Approaches for the Speaker Recognition" 3rd International Conference on Information System Design and Intelligent Applications INDIA 2016, during January 8-9, 2016 at ANIL NEERUKONDA Institute of Technology & Sciences, Visakhapatnam, AP, India Technically co-sponsored by CSI Visakhapatnam Section), INDIA 2016, Volume 1, pp 599-608, (ISBN: 978 – 81 – 322 – 2755 – 7).
18. Kanaka Durga Returi and Y. Radhika "An Artificial Neural Networks Model by Using Wavelet Analysis for Speaker Recognition" India 2015, J.K. Mandal et al. (eds.), Information Systems Design and Intelligent Applications, Advances in Intelligent Systems and Computing 340, DOI 10.1007/978-81-322-2247-7_87, pp 859-874.