ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Silence Removal from Audio Signal Using Framing and Windowing Method and Analyze Various Parameter

J. Meribah Jasmine[1], S. Sandhya [2], Dr. K. Ravichandran [3], Dr. D. Balasubramaniam [4]

Dept. of Nuclear Physics, University of Madras, Chennai, India, [1 2 3]

GKM College of Engg & Technology., Chennai, India[4]

ABSTRACT: Speech signal processing is one of important domain in digital signal processing because variety of noise signals can affect the speech signal so the end user cannot hear the original signal. Noises include quite, ISI, aliasing white noises, etc… This paper presents removal of noise from original speech signal and analyzes the various parameters. This proposed approach is done by framing and windowing method. Initially the sound signal is preprocessed and notices the quite noise (i.e) silent noise and envelope noise removes these noises by setting the maximum envelope level. The parameters include spectrogram, frequency spectrum, magnitude spectrum, thresholding, power spectral density for the noise removed signal. For the implementation of proposed work we use the MATLAB R2011b version software.

KEYWORDS: End point detection, silence removal, Framing, Windowing.

I. INTRODUCTION

Pre-Processing of Speech Signal is very crucial in the applications where silence or background noise is completely undesirable. Applications like Speech and Speaker Recognition needs efficient feature extraction techniques from speech signal where most of the voiced part contains Speech or Speaker specific attributes. Endpoint Detection as well as silence removal is well known techniques adopted for many years for this and also for dimensionality reduction in speech that facilitates the system to be computationally more efficient. This type of classification of speech into voiced or silence/unvoiced sounds finds other applications mainly in Fundamental Frequency Estimation, Formant Extraction or Syllable Marking, Stop Consonant Identification and End Point Detection for isolated utterances. There are several ways of classifying (labelling) events in speech. It is accepted convention to use a three-state representation in which states are

(i) Silence (S), where no speech is produced;
(ii) Unvoiced (U), in which the vocal cords are not vibrating, so the resulting speech waveform is a periodic or random in nature
(iii) Voiced (V), in which the vocal chords are tensed and therefore vibrate periodically when air flows from the lungs, so the resulting waveform is quasi-periodic.

It should be clear that the segmentation of the waveform into well defined regions of silence, unvoiced, signals is not exact; it is often difficult to distinguish a weak, unvoiced sound (like /f/ or /th/) from silence, or weak voiced sound (like /v/ or /m/) from unvoiced sounds or even silence. However, it is usually not critical to segment the signal to a precision much lesser than several milliseconds; hence, small errors in boundary locations usually have no consequence for most applications. Since for most of the practical cases the unvoiced part has low energy content and thus silence (background noise) and unvoiced part is classified together as silence/unvoiced and is distinguished from voiced part.

II. RELATED WORK

In [1] author used Short Time Energy and Statistical method for composite silence removal technique. The performance of the proposed algorithm is compared with the Short Time Energy (STE) algorithm and the statistical

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

method with varying Signal to Noise Ratio (SNR). A comparison between the speaker identification rate including and excluding the silence removal technique shows around 20% increase in identification rate. In [2] author used Time Frequency Transforms for signal decomposition in setting frame theory and provides new insights in application to data analysis. He investigates some interaction of these tools, both theoretically and numerical experiments in order to characterize the signal and their frame transforms. He also used a concepts of persistent homology as an important new subfield for computational topology also formulations of time frequency analysis in frame theory. In [3] author used automatically segment the speech signal into silence, voiced and unvoiced regions which are very beneficial in increasing the accuracy and performance of recognition systems. Proposed algorithm is based on three important characteristics of speech signal namely Zero Crossing Rate, Short Time Energy and Fundamental Frequency.

III. PROPOSED ALGORITHM

End point detection and Silence removal

The captured audio signal may contain silence at different positions such as beginning of signal, in between the words of a sentence, end of signal…. etc. If silent frames are included, modelling resources are spent on parts of the signal which do not contribute to the identification. The silence present must be removed before further processing. There are several ways for doing this: most popular are Short Time Energy and Zeros Crossing Rate. But they have their own limitation regarding setting thresholds as an ad hocbasis.

It assumes that background noise present in the utterances is Gaussian in nature. Usually first 200msec or more of a speech recording corresponds to silence or background noise; because the speaker takes some time to read when recording starts.

Step 1: Calculate the mean (μ) and standard deviation (σ) of the first 200ms samples of the given utterance. The background noise is characterized by this μ and σ.

Step 2: Go from 1st sample to the last sample of the speech recording. In each sample, check whether one-dimensional Mahalanob is distance functions i.e. $| x-\mu |/ \sigma$ greater than 3 or not. If Mahalanob is distance function is greater than 3, the sample is to be treated as voiced sample otherwise it is an unvoiced/silence. The threshold reject the samples up to 99.7% as per given by $P [|x-\mu| \leq 3\sigma] = 0.997$ in a Gaussian distribution thus accepting only the voiced samples.

Step 3: Mark the voiced sample as 1 and unvoiced sample as 0. Divide the whole speech signal into 10 ms non-overlapping windows. Represent the complete speech by only zeros and ones.

Step 4: Consider there are M number of zeros and N number of ones in a window. If $M \geq N$ then convert each of ones to zeros and vice versa. This method adopted here keeping in mind that a speech production system consisting of vocal cord, tongue, vocal tract etc. cannot change abruptly in a short period of time window taken here as 10ms.

Step 5: Collect the voiced part only according to the labelled "1" samples from the windowed array and dump it in a new array. Retrieve the voiced part of the original speech signal from labelled 1 sample.

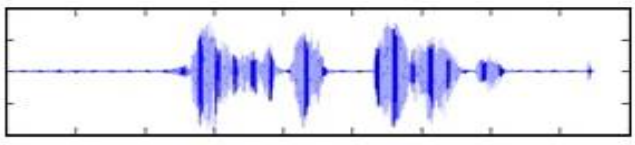


Fig.: Input signal to End-point detection system

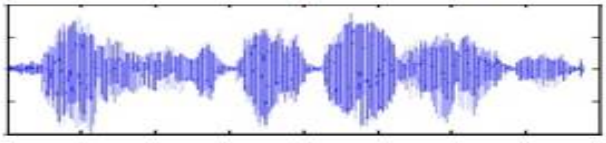


Fig.: Output signal from End point Detection System

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

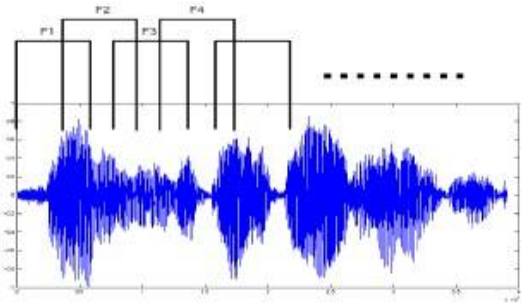
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

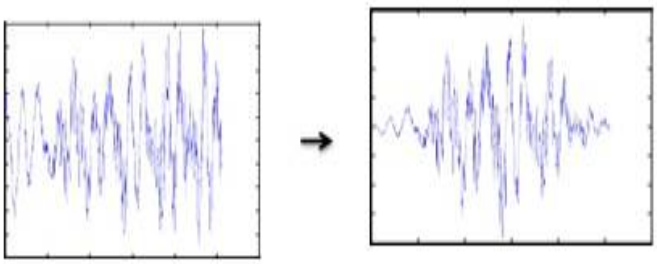
Vol. 4, Issue 4, April 2016

Framing and windowing

Speech is anon-stationary signal, meaning that its statistical properties are not constant across time. Instead, we want to extract spectral features from a small window of speech that characterizes a particular sub phone and for which we can make the (rough) assumption that the signal is stationary (i.e. its statistical properties are constant within this region).We used frame block of 23.22ms with 50% overlapping i.e., 512 samples per frame.



The rectangular window (i.e., no window) can cause problems, when we do Fourier analysis; it abruptly cuts of the signal at its boundaries. A good window function has a narrow main lobe and low side lobe levels in their transfer functions, which shrinks the values of the signal toward zero at the window boundaries, avoiding discontinuities.



PARAMETERS

Parameters that are included

- Spectrogram
- Frequency spectrum
- Magnitude spectrum
- Thresholding
- Power Spectral Density (PSD)

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

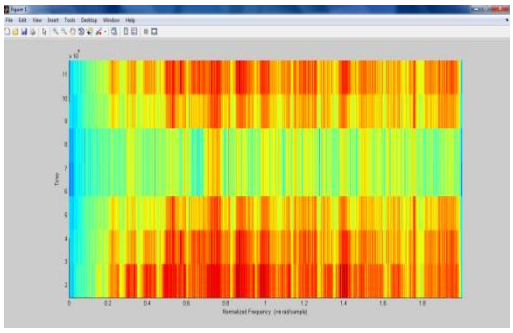
Spectrogram

A graphic representation of a spectrum of frequencies in a sound wave is called a spectrogram. There are many formats of representing a spectrogram. For the implementation of this algorithm, we shall use a three dimensional graph.

- The horizontal axis (X axis) represents time,
- The vertical axis (Y axis) is the frequency and
- The third dimension (Z axis) indicates the amplitude of a particular frequency at a particular time.

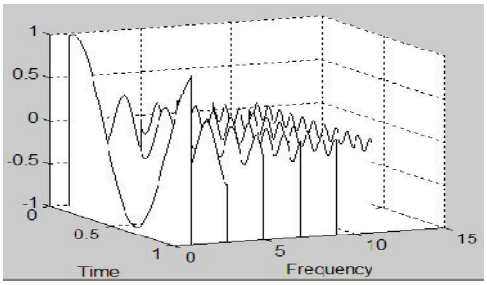
This dimension is represented by the intensity or color of each point in the image. This figure is given below illustrates a typical spectrogram created using this format

.



Frequency Spectrum

When harmonic components of a signal are known, the signal can be presented in a different way that highlights its frequency content rather than its time domain content. Introducing the third axis of frequency perpendicular to the amplitude-time plane the harmonic components can be plotted in the plane that corresponds to their frequencies. For example, a random signal similar to the one in figure below can be presented in such a way that both time and frequency details are visible.



The graph in Figure can be rotated in such a way that the time axis is perpendicular to the observer. This frequency domain view when the time axis is no longer visible and only positive values of magnitude of each sine are plotted is called the magnitude spectrum.

Magnitude spectrum

A sinusoid's frequency content may be graphed in the frequency domain as shown in the figure.1: Spectral magnitude representation of a unit-amplitude sinusoid at frequency Hz (Phase is not shown). An example of a particular sinusoid graphed in Figure is given below.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

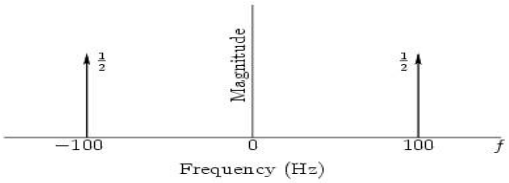
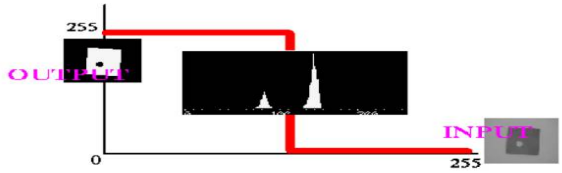


Figure 1: Spectral magnitude representation of a unit-amplitude sinusoid at frequency 100 Hz such as $\cos(200\pi t)$ or $\sin(200\pi t)$. (Phase is not shown.)

Thresholding



Common Names: Threshold, Density slicing

In many vision applications, it is useful to be able to separate out the regions of the audio signal corresponding to original voice in which we are interested, from the regions of the image that correspond to background. Thresholding often provides an easy and convenient way to perform this segmentation on the audio signal.

The time frequency coefficients are grouped in blocks before being attenuated. Block thresholding regulates the estimate and does not create isolated coefficients responsible for our input signal. Making the windowing narrower makes the threshold more sensitive to changes, makes the window wider makes it less sensitive and is a useful parameter.

Power Spectral Estimation

Perhaps one of the more important application areas of digital signal processing (DSP) is the power spectral estimation of periodic and random signals. Speech recognition problems use spectrum analysis as a preliminary measurement to perform speech bandwidth reduction and further acoustic processing. Sonar systems use sophisticated spectrum analysis to locate submarines and surface vessels. Spectral measurements in radar are used to obtain target location and velocity information. Since the estimation of power spectra is statistically based and covers a variety of digital signal processing concepts,

Parametric Methods for Power Spectral Density Estimation

As discussed earlier, we would like to estimate the power spectral density (PSD) of the signal y(t), which is obtained by filtering white noise e(t) of power σ2 through the rational stable and causal filter with the transfer function H(ω) = B(ω)/A(ω), where

$$A(\omega) = 1 + a_1 e^{-i\omega} + \ldots + a_n e^{-in\omega}$$

$$B(\omega) = 1 + b_1 e^{-i\omega} + \ldots + b_m e^{-im\omega}.$$

In the time domain, the above filtering can be represented as

We further divide this problem into three categories based on the values of m and n:

• If both m and n are non-zero, then the signal is said to be Auto-regressive moving average (ARMA) and is denoted by ARMA (n, m).

• If m = 0, then the signal is an auto-regressive (AR) signal and is denoted by AR (n); and finally,

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

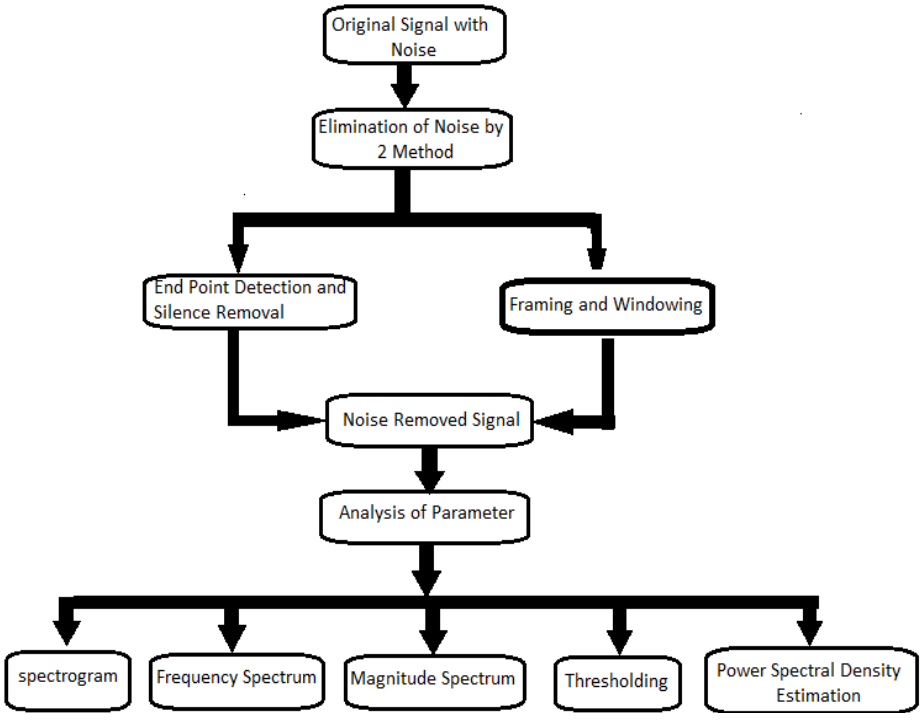
(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

• If n = 0, the signal is a moving average (MA) signal and is denoted by MA (m).
In the time domain, the above filtering can be represented as

$$y(t) + \sum_{i=1}^{n} a_i y(t-i) = \sum_{j=0}^{m} b_j e(t-j), \quad (b_0 = 1).$$

IV. WORKFLOW



V. SIMULATION RESULTS

The simulation studies involve the silence removal from audio signal and analyse various parameters using MATLAB. Figure 1 shows the output of the noise removed signal. In this fig the 1st part shows the noise from the input signal. The noise is indicated by red color. The 2nd part shows the noise and silence removed signal. Figure 2 shows the Spectrogram of the input signal and three dimensional view of the input signal can be viewed from this output in which the x axis represents the time, y axis represents the frequency and z axis represents the amplitude the input signal. Figure 3 shows the frequency domain view when the time axis is no longer visible and only positive values of magnitude of each sine are plotted, is called the frequency spectrum**.** Figure 4 shows the output of magnitude spectrum. The magnitude spectrum represents the unit-amplitude sinusoid at frequency. Figure 5 shows the thresholding of the input signal which separate out the regions of the audio signal corresponding to original voice from the regions of the audio signal. Figure 6 shows the Power Spectral Density of the input signal. This Power Spectral Density describes how power of a signal or time series is distributed over frequency.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

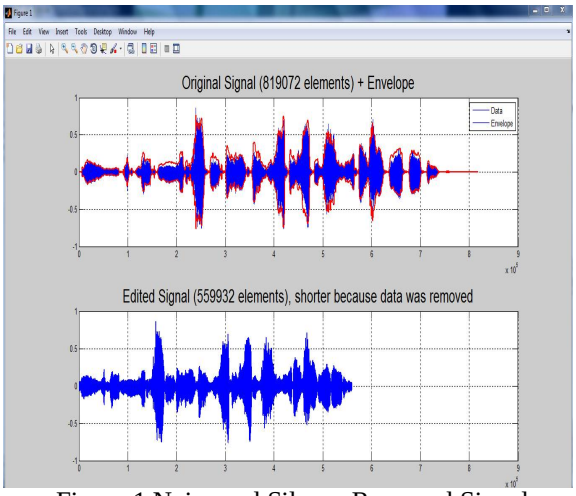


Figure 1 Noise and Silence Removed Signal

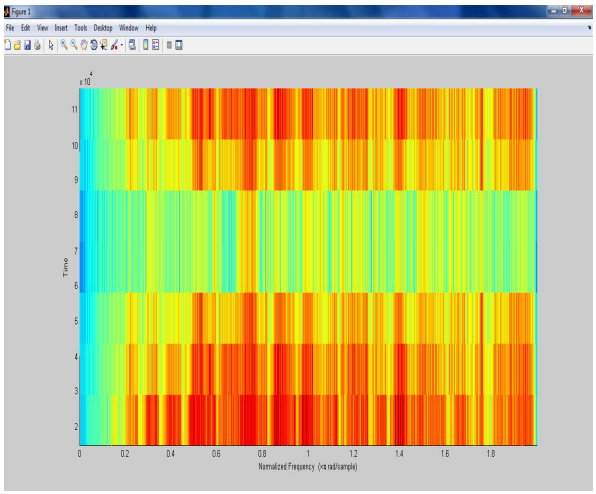


Figure 2 Spectrogram of the signal

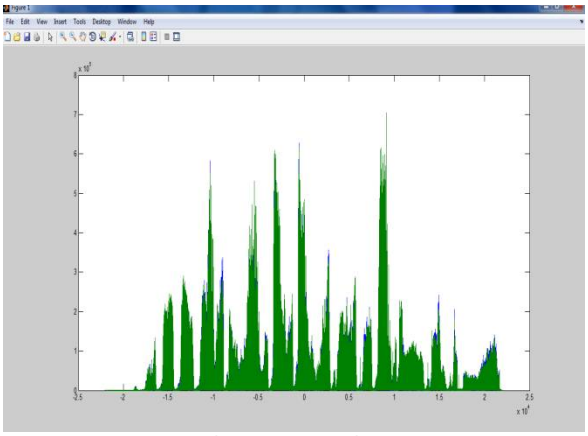


Figure 3 Frequency Spectrum

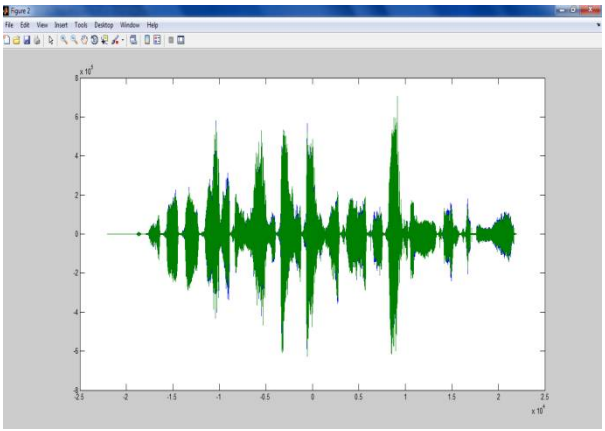


Figure 4 Magnitude Spectrum

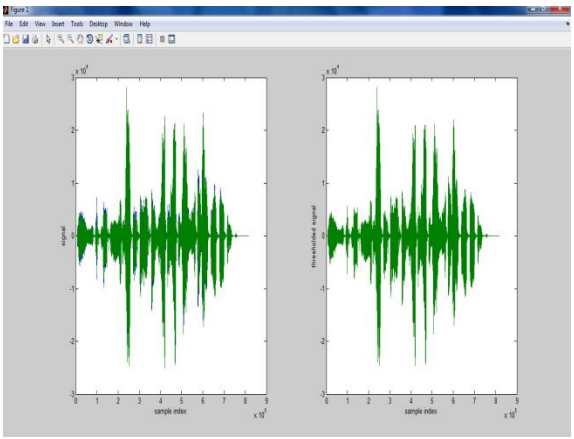


Figure 5 Thresholding

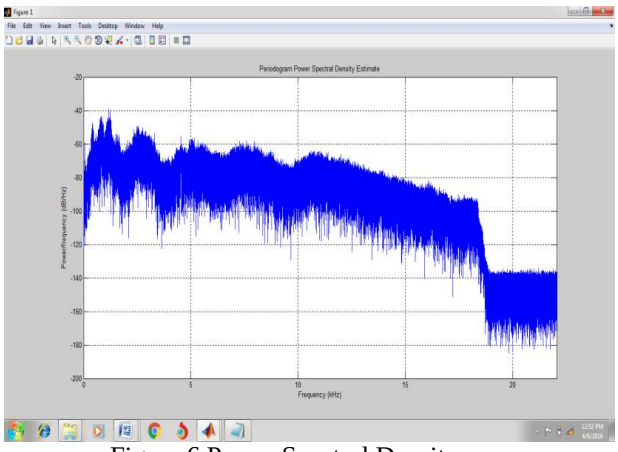


Figure 6 Power Spectral Density

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

VI. CONCLUSION AND FUTURE WORK

In this project, Silence and noise of an input signal is removed by using MATLAB software. And it gives the resultant output without any noise and silence. I also added another five parameters which gives additional techniques of the signal. This process can also be implemented by using LMS loop adaptive filter algorithm for getting better accuracy.

REFERENCES

1. Tushar Ranjan Sahoo, Sabyasachi Patra, "Silence Removal and Endpoint detection of speech signal for text independent Speaker identification", IJIGSP Vol.6, No.6, May 2014
2. Mijail Guillemard, Gitta Kutyniok, Holger Boche, Friedrich Philipp,"Signal Analysis with Frame Theory and Persistent Homology", June 2013
3. Poonam Sharma and Abha Rajpoot, "Automatic Identification of Silence, Unvoiced and Voiced Chunks in Speech", Journal of Computer Science & Information Technology (CS & IT) 3(5), 87-96, 2013
4. J. P. Campbell, Jr., "Speaker Recognition: A Tutorial", Proceedings of The IEEE, Vol.85, No.9, pp.1437-1462, Sept.1997.
5. Geneva, "Endpoint Detection by Using Filled Pauses", Eurospeech 2003 , pp. 1237-1240.
6. S. E. Bou-Ghazale and K. Assaleh, "A robust endpoint detection of speech for noisy environments with application to automatic speech recognition", in Proc. ICASSP2002, vol. 4, 2002, pp. 3808–3811.
7. A. Martin, D. Charlet, and L. Mauuary, "Robust speech / non-speech detection using LDA applied to MFCC", in Proc. ICASSP2001}, vol. 1, 2001, pp. 237–240.
8. K. Ishizaka and J.L Flanagan, "Synthesis of voiced Sounds from a Two-Mass Model of the Vocal Chords," Bell System Technical J., 50(6): 1233-1268, July-Aug., 1972.
9. Atal, B.; Rabiner, L., "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition" Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on , Volume: 24 , Issue:3 , Jun 1976, Pages: 201 - 212.
10. D. G. Childers, M. Hand, J. M. Larar, " Silent and Voiced/Unvoied/ Mixed Excitation(Four-Way), Classification of Speech", IEEE Transaction on ASSP, Vol-37, No-11, pp. 1771-74, Nov 1989.
11. Richard. O. Duda, Peter E. Hart, David G. Strok,"Pattern Classification", A Wiley-interscience publication, John Wiley & Sons, Inc, Second Edition, 2001.
12. Sarma, V.; Venugopal, D., "Studies on pattern recognition approach to voiced-unvoiced-silence classification", Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'78. Volume: 3, Apr 1978, Pages: 1 - 4.
13. L. R. Rabiner and R.W. Schafer, "Digital Processing of Speech Signals", First Edition, Chapter-4, Pearson Education, Prentice-Hall.
14. http://cslu.ece.ogi.edu/nsel/data/SpEAR_technical.html.
15. J. L. Flanagan, "Speech Analysis, Synthesis, and Perception" 2nd ed., Springer-Verlag, New York, 1972.
16. L.R. Rabiner and B. H. Juang, "Fundamentals of speech recognition," 1st Indian Reprint, Pearson Education.
17. Adrian E. Villanueva- Luna, Alberto Jaramillo-Nuñez, Daniel Sanchez-Lucero, Carlos M. Ortiz-Lima, J. Gabriel Aguilar-Soto, Aaron Flores-Gil and Manuel May-Alarcon (2011).
18. De-Noising Audio Signals Using MATLAB Wavelets Toolbox, Engineering Education and Research Using MATLAB, Dr. Ali Assi (Ed.), ISBN: 978-953-307-656-0, InTech, Available from: http://www.intechopen.com/books/engineering-education-and-researchusing-matlab/de-noising-audio-signals-using-matlab-wavelets-toolbox
19. Vinay K. Ingle and John G. Proakis, 2012-2013, *Digital Signal Processing Using MATLAB®, Third Edition,* pp 2-5
20. Li Deng Douglas O'Shaughnessy, SPEECH PROCESSING A Dynamic and Optimization-Oriented Approach, pp 5-8
21. Milind Bhattacharya1, Sweekar Bandkar2, Amit Badala3, MUSIC ANALYZER AND PLAGIARISM, Volume: 03 Special Issue: 05 | May-2014 | NCEITCS-2014, Available @ http://www.ijret.org

BIOGRAPHY

1. **Miss J.Meribah Jasmine** M.Sc, University of Madras
2. **Miss. S. Sandhya** (Guest faculty), University of Madras
3. **Dr. K. Ravichandran,** (Associate Professor), University of Madras
4. **Dr. D. Balasubramaniam**, HOD of ECE Dept, GKM College of Engineering and Technology