Multi-Resolution Speech Spectrogram

Rohini R. Mergu
Lecturer
WIT, Solapur

Dr.Shantanu K. Dixit
Professor & Head
WIT, Solapur

ABSTRACT
An important aid in analysis & display of speech is sound spectrogram. It represents time-frequency-intensity display of short time spectrum. The quality of speech can be studied by visual inspection of spectrogram. This is one of the important applications of spectrogram in speech processing especially in speech enhancement. Another application of spectrogram is in isolating voiced and unvoiced regions. But to conclude from visual inspection the clarity of spectrogram is also important. Before plotting the spectrogram the time domain speech signal is converted to frequency domain. The transform domain used plays vital role in resolution of spectrogram. Generally Fast Fourier Transform is used to convert the time domain signal into frequency domain signal. This paper discusses the effect of using different transform for converting the time domain speech signal into frequency domain before plotting spectrogram. . It is observed that resolution of speech spectrogram is transform dependent.

Keywords
Spectrogram, Speech Enhancement, Speech Processing, Speech & Noise, Speech Quality, SNR, Resolution.

1. INTRODUCTION
In many practical situations, speech has to be recorded in the presence of undesirable background noise. As noise often degrades the quality/intelligibility. In many practical situations, speech has to be recorded in the presence of undesirable background noise. As noise often degrades the quality/intelligibility of recorded speech, it is beneficial to carry out noise suppression. In the literature, a variety of speech enhancement methods capable of suppressing noise has been proposed. In speech enhancement the graphical representation of speech is spectrogram plays vital role to examine speech quality.

The quality of speech can be observed quickly using spectrogram. This is one of the important applications of spectrogram in speech enhancement. Another application of spectrogram is in isolating voiced and unvoiced regions. But to conclude from visual inspection the clarity of spectrogram is also important. Before plotting the spectrogram the time domain speech signal is converted to frequency domain. The transform domain used plays vital role in resolution of spectrogram. Generally Fast Fourier Transform is used to convert the time domain signal into frequency domain signal. This paper discusses the effect of using different transform for converting the speech signal into frequency domain before plotting spectrogram.

Zenton Goh, Kah-Chye Tan, and B.T.G.Tan [1] examined the spectrograms of typical clean speech, noisy speech, and enhanced speech. The horizontal axis of the spectrogram denotes time, vertical axis frequency, and the spectral magnitude is shown with gray shade (darker shade indicates larger value). It is observed that a large portion of the spectrogram is practically blank (i.e., unshaded) and the speech energy is concentrated in a few isolated regions. The voiced portion of speech is characterized by dark parallel "stripes" whereas unvoiced portion is characterized by gray patches. Some parallel stripes are horizontal while some are slanting up or down, indicating a change in the pitch of the speech signal. When white Gaussian noise amounting to the clean speech, the blank region of the spectrogram become shaded, and some of the stripes corresponding to voiced speech disappear. With an appropriate spectral subtraction, obtained an enhanced speech with spectrogram and observed a significant reduction of the unwanted short stripes. By observation of spectrogram [1] concluded about speech quality.

S. Gannot, D. Burshtein, and Ehud Weinstein [6] presented a class of Kalman filter-based algorithms with some extensions, modifications, and improvements of previous work. The first algorithm employs the estimate-maximize (EM) method to iteratively estimate the spectral parameters of the speech and noise parameters. The enhanced speech signal is obtained as a byproduct of the parameter estimation algorithm. And used sound spectrogram for comparison of speech quality using Kalman-EM-Iterative (KEMI) algorithm and log spectral amplitude estimator (LSAE) algorithm. R.C.Hendriks, R.Heusdens, and J. Jensen [2] used a deterministic model in combination with the well-known stochastic models for speech enhancement. Thus derived a minimum mean-square error(MMSE) estimator under a combined stochastic–deterministic speech model with speech presence uncertainty and show that for different distributions of the DFT coefficients the combined stochastic–deterministic speech model leads to improved performance and used speech spectrogram for classification of speech component as deterministic or stochastic. Nicholas W.D. Evans, John S.Mason and Matt J. Roach [5] described the application of morphological filtering to speech spectrograms for noise robust automatic speech recognition. Speech regions of the spectrogram are identified based on the proximity of high energy regions to neighboring high energy regions in the three-dimensional space.

H.Ding, I.Y.Soon, S.N.Koh,C.K.Yeo[4] proposed a hybrid Wiener spectrogram filter (HWSF) for effective noise reduction, followed by a multi-blade post-processor which exploits the 2D features of the spectrogram to preserve the speech quality and to further reduce the residual noise. Spectrogram comparisons show that in the proposed scheme, musical noise is significantly reduced. Cyril Plapous, Claude Marro, and Pascal Scalart [8] proposed a method called two-step noise reduction

(TSNR) technique which solves reverberation problem while maintaining the benefits of the decision-directed approach. However, classic short-time noise reduction techniques, including TSNR, introduce harmonic distortion in enhanced speech because of the unreliability of estimators for small signal-to-noise ratios. To overcome this problem, proposed a method called harmonic regeneration noise reduction (HRNR). Nonlinearity is used to regenerate the degraded harmonics of the distorted signal in an efficient way. Spectrogram of noisy speech enhanced by TSNR technique and enhanced by HRNR technique. The spectrograms of clean speech and enhanced by two techniques are compared [5].

2. SPECTRAL ANALYSIS OF SPEECH: SPECTROGRAM

A spectrogram is a time-varying spectral representation that shows how the spectral density of a signal varies with time. In the field of time–frequency signal processing, it is one of the most popular quadratic Time-Frequency distribution that represents a signal in a joint time-frequency domain. Also known as spectral waterfalls, sonograms, voiceprints, or voicegrams, spectrograms are used to identify phonetic sounds, to analyze the cries of animals; they were also used in many other fields including music, sonar/radar, speech processing, seismology, etc. The instrument that generates a spectrogram is called a spectrograph. The most common format is a graph with two geometric dimensions: the horizontal axis represents time, the vertical axis is frequency; a third dimension indicating the amplitude of a particular frequency at a particular time is represented by the intensity or colour of each point in the image.

Spectrograms are usually created in one of two ways: approximated as a filter bank that results from a series of band pass filters (this was the only way before the advent of modern digital signal processing), or calculated from the time signal using the short-time Fourier transform (STFT) [1,2,4,7]. These two methods actually form two different quadratic Time-Frequency Distributions, but are equivalent under some conditions. Creating a spectrogram using the STFT is usually a digital process. Digitally sampled data, in the time domain, is broken up into chunks, which usually overlap, and Fourier transformed to calculate the magnitude of the frequency spectrum for each chunk. Each chunk then corresponds to a vertical line in the image; a measurement of magnitude versus frequency for a specific moment in time. The spectrums or time plots are then "laid side by side" to form the image or a three-dimensional surface [5].

A spectrogram shown in Figure 1 is created from the speech waveform. The spectra computed by the Fourier transform are displayed parallel to the vertical or y-axis. The horizontal axis represents time. As we move right along the x-axis we shift forward in time, traversing one spectrum after another. Spectrograms are normally computed and kept in computer memory as a two-dimensional array of acoustic energy values. For a given spectrogram S, the strength of a given frequency component f at a given time t in the speech signal is represented by the darkness or color of the corresponding point
S (t , f).

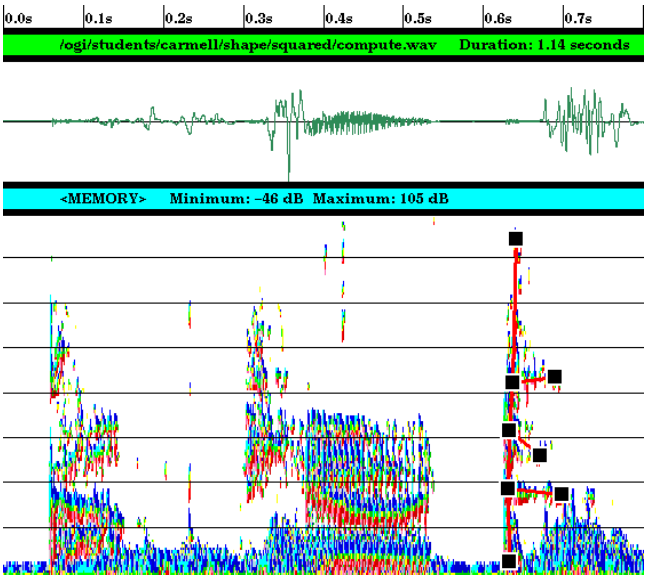

Figure 1 : Speech Spectrogram

The use of colour to highlights the important features of a spectrogram. In the spectrogram shown in Figure 1 the shades of red indicates increasing energy along the frequency axis, blue to mean decreasing energy, and yellow and green to mean an energy maximum. Areas which are white do not have enough energy to be of interest to us

3. COMPUTAION OF SPECTROGRAM

The use of spectrogram in speech enhancement is discussed in this paper.

The additive noise model is described by the following equation,

$$y(t) = x(t) + n(t) \qquad (1)$$

Where, $y(t)$ is the observed noisy speech, $x(t)$ is the clean speech and $n(t)$ is the additive background noise.

The observed speech is then divided into overlapping frames of length of 256 samples in each frame .The amount of overlap is normally either 50% or 75%. In this paper, 75% overlapping is used throughout. The nth frame can be represented by a column vector described by the following equation:

$$\mathbf{f}_L = [y(64L)y(64L+1)y(64L+2)\ldots y(64L+255)]^T. \quad (2)$$

All indices used in this paper starts from zero. A speech block can be obtained by arranging a number of frames together to form a matrix. Suitable numbers of frames are found experimentally to be 8, 16 and 32.In this paper, the number of frames used is 16 throughout. Similarly each block overlaps its neighboring block by 75%. Then the speech block can be represented mathematically as a matrix, of size 256 by 16 as shown in the following equation:

$$\mathbf{b}_n = [\mathbf{f}_{8n} \quad \mathbf{f}_{8n+1} \quad \mathbf{f}_{8n+2} \quad \ldots \quad \mathbf{f}_{8n+15}]. \qquad (3)$$

This signal is windowed using Hamming window. Then the transform can be applied onto the speech block.

3.1 Using DFT

Discrete Fourier Transform (DFT) can be computed efficiently using a fast Fourier transform (FFT) algorithm. The discrete Fourier transform (DFT) is a specific kind of Fourier transform, used in Fourier analysis. It transforms the time domain function into frequency domain representation. FFT algorithms are so commonly employed to compute DFTs that the term FFT is often used to mean DFT in colloquial settings.

DFT can be defined as,

For length N input vector x, the DFT is a length N vector,

$$X(k) = \sum_{j=1}^{N} x(j)\omega_N^{(j-1)(k-1)} \quad (4)$$

$$x(j) = \left(\frac{1}{N}\right)\sum_{k-1}^{N} X(k)\omega_N^{-(j-1)(k-1)} \quad (5)$$

where,

$$\omega_N = e^{(2\pi i/N)} \quad (6)$$

3.2 Using DCT

A Discrete Cosine Transform (DCT) expresses a sequence of finitely many data points in terms of a sum of cosine functions oscillating at different frequencies. It turns out that cosine functions are much more efficient as fewer terms are needed to approximate a typical signal. In particular, a DCT is a Fourier-related transform similar to the discrete Fourier transform (DFT), but using only real numbers. DCTs are equivalent to DFTs of roughly twice the length, operating on real data with even symmetry.

$$y(k) = \omega(k)\sum_{n=1}^{N} x(n)cos\frac{\pi(2n-1)(k-1)}{2N} \quad ,k = 1,\ldots.N \quad (7)$$

$$\omega(k) = \begin{cases} \frac{1}{\sqrt{N}}, & k = 1 \\ \sqrt{\frac{2}{N}}, & 2 \le k \le N \end{cases} \quad (8)$$

4. RESULTS & DISCUSSIONS

Here spectrogram is plotted for different utterances of human speech male & Female. Also for different noise conditions with different SNRs (Signal to Noise Ratio). The speech utterances are obtained from noiseus database. Different speech utterances used in this paper are as shown in Table 1. The spectrograms plotted using 256 point DFT & 256 point DCT are shown in figures 2 to 6.

Table 1: Details of speech utterances

File Name	Gender	Sentence Text
Sp01	Male	The birch canoe slid on the smooth planks.
Sp05	Male	Wipe the grease off his dirty face.
Sp06	Male	Men strive but seldom get rich
Sp19	Female	We talked of the sideshow in the circus.

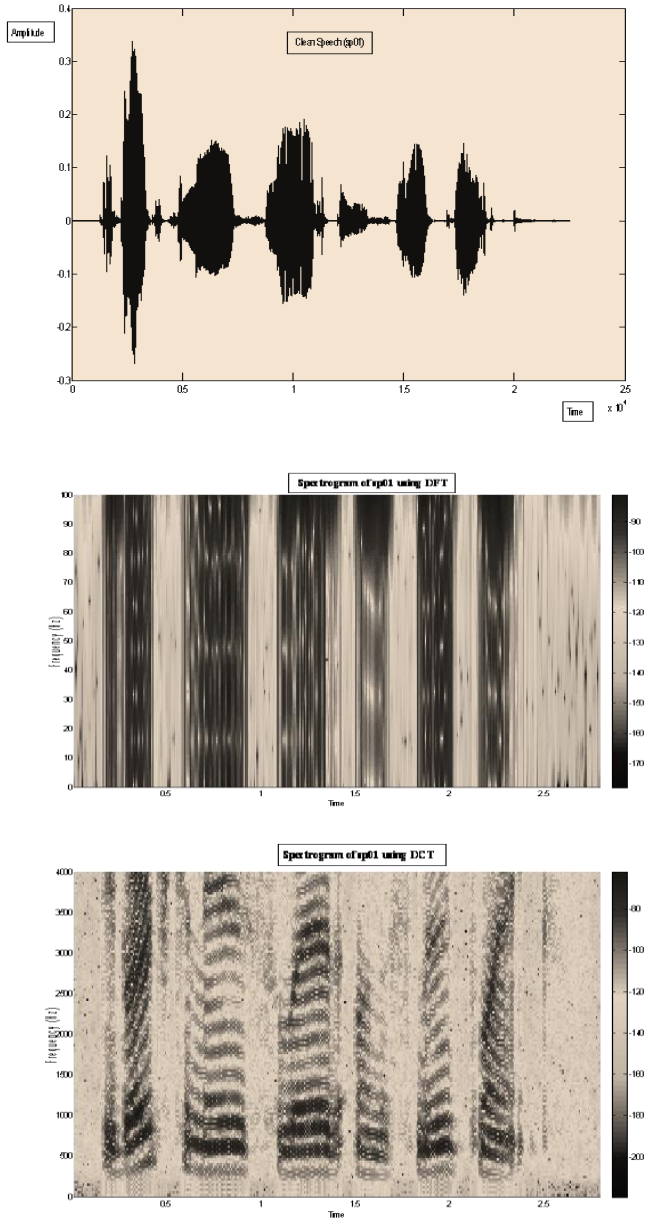

Figure 2: Upper plot - clean speech sp01, Middle plot –spectrogram plotted using DFT, Lower plot – spectrogram plotted using DCT

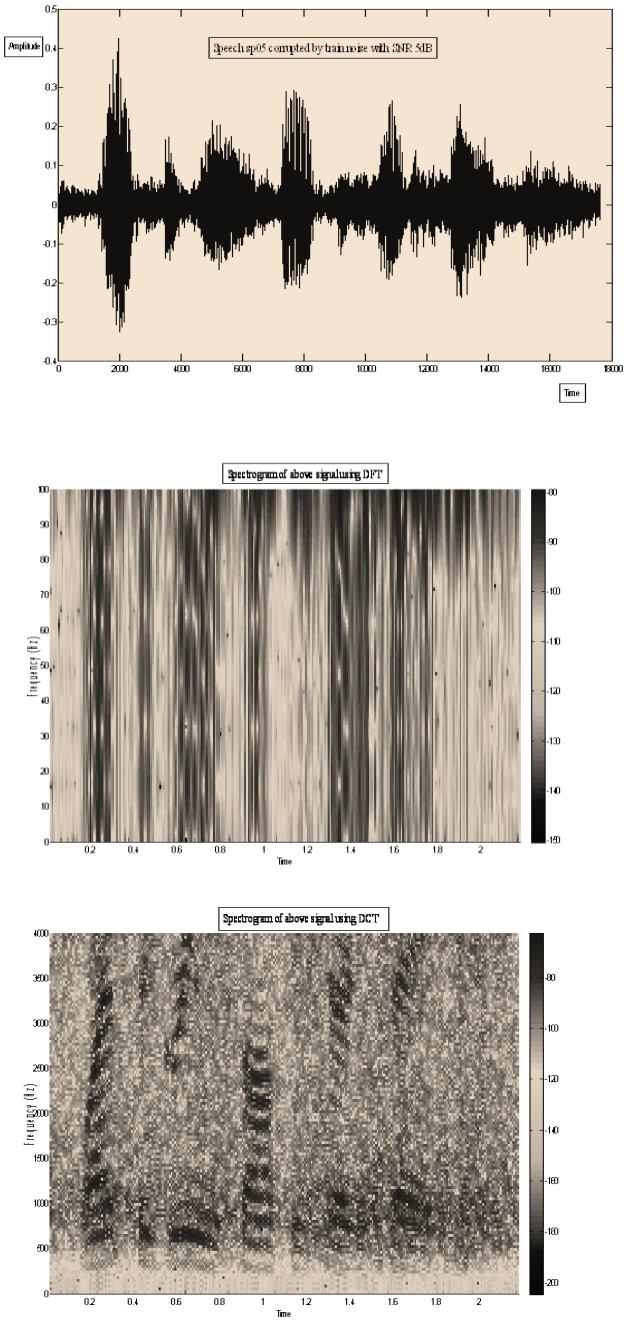

Figure 3: Upper plot –speech sp05 corrupted by train noise SNR = 5dB,
Middle plot –spectrogram plotted using DFT,
Lower plot – spectrogram plotted using DCT

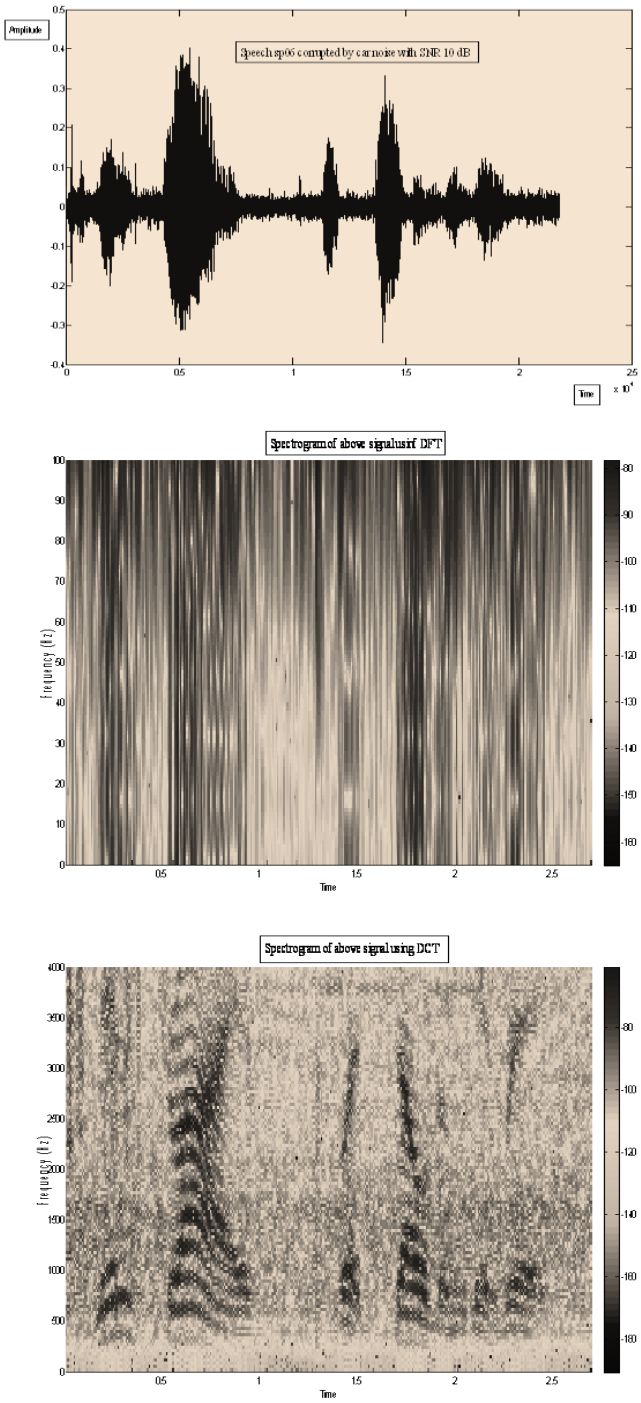

Figure 4- Upper plot –speech signal sp06 corrupted car noise SNR = 10 dB,
Middle plot –spectrogram plotted using DFT,
Lower plot – spectrogram plotted using DCT

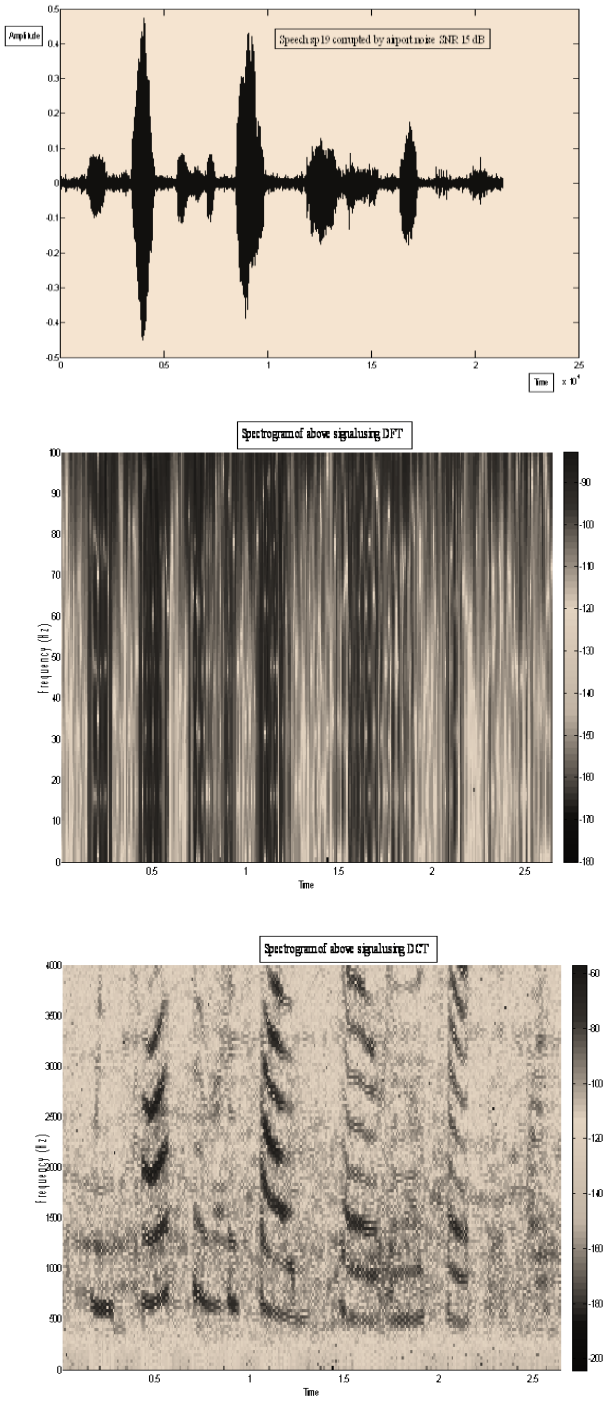

Figure 5 - Upper plot –speech signal sp19 corrupted by airport noise SNR = 15dB,
Middle plot –spectrogram plotted using DFT,
Lower plot – spectrogram plotted using DCT

5. CONCLUSION

From the results shown above we can conclude that the spectrograms plotted using DCT are clearer than the spectrograms plotted using same point DFT. The spectrogram plot using DCT is having higher resolution than that plotted using DFT. From the visual inspection we can see the amount noise available in the speech signal. Thus the quality of input signal can be inspected from spectrogram. From visual inspection of spectrogram plotted using DCT we can say that the noise content is more in signal shown in figure 3 compared to 4 & 5. Also the spectrogram in figure 1 shows that the signal is of free of noise. The voiced and unvoiced regions are very well differentiated and the energy at different time instant in particular frequency bin can be observed very clearly in spectrogram plotted using DCT due to higher resolution. Whereas the in the spectrograms plotted using DFT the energy content, amount of noise and voiced/unvoiced region detection is much difficult. Thus plotting spectrogram using DCT provides higher resolution plot than plotting by the usual method using DFT.

6. REFERENCES

[1] Zenton Goh, Kah-Chye Tan, and B.T.G. Tan," Postprocessing Method for Suppressing Musical Noise Generated by Spectral Subtraction", IEEE trans. on Speech and Audio Processing, vol 6, no.3, pgs. 287-292, May 1998.

[2] Richard C. Hendriks, Richard Heusdens ,and Jesper Jensen," An MMSE Estimator for Speech Enhancement Under A Combined Stochastic–Deterministic Speech Model", IEEE trans on Speech & Audio Processing,Vol.15,No.2,Feb 2007.

[3] Jesper Jensen and John H.L.Hansen, "Speech Enhancement Using a Constrained Iterative Sinusoidal Model", IEEE trans. on Speech and Audio Processing, Vol 9, No.7,pgs. 731-740, Oct 2001.

[4] H. Ding, ,I. Y. Soon, S.N.Koh, C.K. Yeo, "A spectral filtering method based on hybrid wiener filters for speech enhancement", Science Direct, Speech Communication 51(2009) pgs. 259–267

[5] Nicholas W.D. Evans, John S.Mason and Matt J.Roach ,"Noise Compensation using Spectrogram Morphological Filtering", Speech and Image Research Group, Department of Electrical and Electronic Engineering University of Wales Swansea, UK.

[6] Sharon Gannot, David Burshtein, and Ehud Weinstein," Iterative and Sequential KalmanFilter-Based Speech Enhancement Algorithms", IEEE trans on Speech & Audio Processing,Vol.6,No.4,July 1998.

[7] I.Y.Soon, S.N. Koh,"Speech Enhancement Using 2-D Fourier Transform", IEEE trans. on Speech and Audio Processing, Vol 11, No.6,pgs. 717-724, Nov 2003.

[8] Cyril Plapous, Claude Marro, and Pascal Scalart," Improved Signal-to-Noise Ratio Estimation For Speech Enhancement", IEEE trans. on Audio, Speech & Language Processing,Vol.14,No.6,Nov 2006.