Investigation of Transform Dependency in Speech Enhancement

Rohini R. Mergu, Shantanu K. Dixit

Abstract*— Noise is an unwanted signal. One of the most common type of noise is a background noise which is always present. The paper presents speech enhancement scheme for suppression of background noise. The objective of speech enhancement is to improve the perceptual aspect such as quality and intelligibility of the processed speech. The main objective of this paper is to investigate the use of different transforms for speech enhancement. Speech enhancement using wiener filtering approach is proposed and implemented using DFT, DCT and DWT thus showing the feasibility of utilization of the different transforms.

Keywords*— Speech enhancement, DFTF, DCTF, DWTF, wiener filter, transform

I. INTRODUCTION

Speech enhancement system aims to improve the quality of speech for various different applications. In many practical situations, people may be speaking in a noisy environment, such as in a car or in an airport, during which signals are corrupted by various types of background noises and this actually results in listener fatigue and lowers intelligibility. Hence, there is a strong need to develop speech enhancement algorithms. Traditionally, many speech algorithms operate in the frequency domain or some other transform domains. The Speech enhancement system aims to improve the quality of speech for various different applications. Although a lot of work has been proposed to reduce the effect of background noise in single channel speech enhancement, the problem remains a challenging one. The reason is that speech energy is not present in all the transform co-efficients and it is therefore easier to filter off the noise especially for the noise-only co-efficients. Such algorithms normally operate on the spectral amplitude of the degraded speech. If noise is assumed to be additive, stationary and statistically independent of the speech signals, various mathematical tools can be used to derive an optimal statistical operator with respect to mean square error (MSE) for noise reduction. The minimum mean square error (MMSE) filter by Ephraim and Malah [1] is an important milestone. In these estimation type approaches, the transform co-efficients are filtered in each short-time frame and attenuated independently of their intra-frame neighboring co-efficients as well as inter-frame neighboring co-efficients. Spectral subtraction offers a computationally efficient, processor-independent approach to effective digital speech analysis. The method, requiring about the same computation as high-speed convolution, suppresses stationary noise from speech by subtracting the spectral noise bias calculated during non speech activity [2].

Manuscript Received on July, 2013.

Mrs. Rohini R. Mergu, Assistant Professor WIT, Solapur, India.
Dr. Shantanu K. Dixit, Professor and Head E&TC WIT, Solapur, India.

In It is proposed a method for enhancement of speech in the presence of additive noise. The objective is to selectively enhance the high signal-to-noise ratio(SNR) regions in the noisy speech in the temporal and spectral domains, without causing significant distortion in the resulting enhanced speech [3]. [4] Presents a sinusoidal model based algorithm for enhancement of speech degraded by additive broad-band noise. In order to ensure speech-like characteristics observed in clean speech, smoothness constraints are imposed on the model parameters using a spectral envelope surface (SES) smoothing procedure. Use of the two-dimensional (2-D) Fourier transform for speech enhancement presented [5]. Also, include magnitude spectral subtraction, 2-D Wiener filtering as well as a hybrid filter which effectively combines the one-dimensional (1-D) Wiener filter with the 2-D Wiener filter.

This is apparent in many recent works which view speech as a 2D time–frequency signal, especially in the form of a spectrogram. Evans has applied morphological filtering on the spectrogram [6] using opening operator based on erosion and dilation which is borrowed from digital image processing tools, and has obtained improved results. However, this algorithm emphasizes more on 2D processing without exploiting the characteristics of the speech spectrogram, resulting in the attenuation of the speech content together with the noise. The transform domain used plays vital role in clarity of spectrogram. [7] Generally Fast Fourier Transform is used to convert the time domain signal into frequency domain signal before plotting spectrogram. The transform domain used plays vital role in clarity of spectrogram. It is observed that [8] resolution of spectrogram is transform dependent .The advantages of using the Discrete Cosine Transform (DCT) as compared to the standard Discrete Fourier Transform (DFT) for the purpose of removing noise embedded in a speech signal is illustrated[8]. The derivation of the Minimum Mean Square Error (MMSE) filter based on the statistical modelling of the DCT coefficients and derivation of an over-attenuation factor based on the fact that speech energy is not always present in the noisy signal at all times or in all coefficients[9]. A warped discrete cosine transform (WDCT) based approach to enhance the degraded speech under background noise environments is proposed [10]. For developing an effective expression of the frequency characteristics of the input speech, the variable frequency warping filter is applied to the conventional discrete cosine transform (DCT).

II. METHODOLOGY

A. Overview

The detailed block diagram of the proposed work is shown in fig.1. The Noisy speech is the speech signal corrupted by different types of background noises as fan noise, car noise, aeroplane noise, train noise.The

additive noise model is described by the following equation,

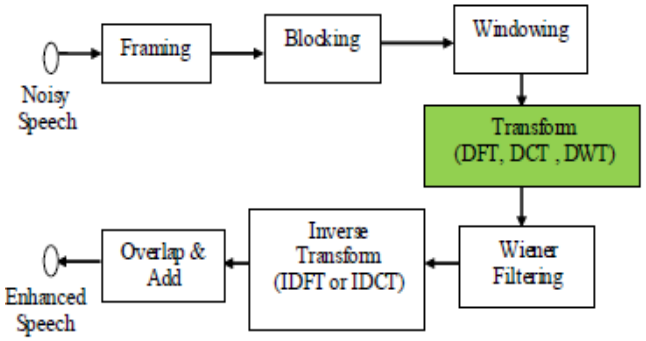


Fig 1 : Block diagram of speech Enhancement System

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

Where,

$y(t)$ is the observed noisy speech,

$x(t)$ is the clean speech

$n(t)$ is the additive background noise.

The noisy speech is divided into overlapping frames of length of 256 samples in each frame and 75% overlapping is used. The n[th] frame, can be represented by a column vector described by the following equation

$$fL = [y(80L)y(80L+1)y(80L+2)\dots y(80L+255)T \quad (2)$$

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

B. Transform

It is easier to remove noise from the noisy speech in frequency domain. Hence, covert time domain speech signal to frequency domain using transform. Generally used and the most popular transform used is Discrete Fourier Transform (DFT). But this paper proposes the use of Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) for converting time domain signal in frequency domain and shows the effect of change of transform on the speech quality and intelligiability.

B.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,

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

B.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.

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

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

B.3) Using DWT :

A Discrete Wavelet Transform (DWT) is computed by successive lowpass and highpass filtering of discrete time domain signal. Commonly used set of DWT is Daubechies wavelets.

One level of the transform -

The DWT of a signal 'x' is calculated by passing it through a series of filters. First the samples are passed through a low pass filter with impulse response 'g'resulting in a convolution of the two:

$$y[n] = (x * g)[n] = \sum_{k=-\infty}^{\infty} x[k]g[n-k]. \quad (9)$$

The signal is also decomposed simultaneously using a high pass filter 'h'. The outputs giving the detail coefficients (from the high-pass filter) and approximation coefficients (from the low-pass). It is important that the two filters are related to each other and they are known as a quadrature mirror filter.

However, since half the frequencies of the signal have now been removed, half the samples can be discarded according to Nyquist's rule. The filter outputs are then subsampled by 2 (Mallat's and the common notation is the opposite, g- high pass and h- low pass)

$$y_{\text{low}}[n] = \sum_{k=-\infty}^{\infty} x[k]g[2n-k] \quad (10)$$

$$y_{\text{high}}[n] = \sum_{k=-\infty}^{\infty} x[k]h[2n-k] \quad (11)$$

This decomposition has halved the time resolution since only half of each filter output characterises the signal. However, each output has half the frequency band of the input so the frequency resolution has been doubled. One level DWT is used in this paper.

C. Wiener Filtering

The wiener filter produces the highest noise attenuation.

$$\hat{x}(t) = g(t) * y(t) \quad (12)$$

Where,
$\hat{x}(t)$ = Estimated signal after Filtering
$g(t)$ = impulse response of Wiener Filter

And the difference between clean speech & estimated signal indicates amount of error. This is given as,

$$e(t) = x(t) - \hat{x}(t) \quad (13)$$

Where,
$e(t)$= error signal

The prime focus in this paper is to reduce error in turn background noise in speech signal. This is done by wiener filtering but the focus in this paper is on transform domain used. Three different well known transforms are used DFT filtered wiener (DFTF), DCT filtered wiener (DCTF), and DWT filtered wiener (DWTF). The results for of processed speech along with clean and noisy speech are shown in figs. 2 to 5.

III.RESULTS & DISCUSSION

The results of the speech corrupted by background noise such as airport noise, babble noise and car noise are considered in this paper. The speech utterances are obtained from noiseus database. Also the results are shown for different (Signal to Noise Ratio) SNRs. The speech utterances “sp01 and sp23” is uttered by “male” speaker and “sp11” is uttered by “female” speaker. The noise reduction for the these sentence in different noise conditions with different SNRs for different transforms are presented below.

From the figs. 2 to 5 we can see that the processed speech shows reduction in noise using DFTF, DCTF and DWTF.

From the results shown in Table I and II we can observe that noise reduction is good in case of DFTF and DCTF compared to DWTF for low SNR values for types of noises considered in this paper. The value of segSNR is large for DFTF and DCTF and less for DWTF for low SNR means 0 dB and 5 dB. But for higher SNRs i.e, 10 dB and 15 dB DWTF gives large value of segSNR than DFTF and DCTF. DFTF is also better than DCTF for high SNRs. The quality of processed speech is good in all the three cases.

From table I DCTF is better for very low SNR of 0 dB and DFTF is better than other for SNR of 5 dB. But DWTF is giving degraded segSNR for SNR of 0and 5 dB and gives high segSNR for 10db and 15 dB.

The overall SNR is good for DFTF and DCTF for all the cases. But DWTF shows less overall SNR but high segSNR for SNR of 10 and 15 dB SNR.

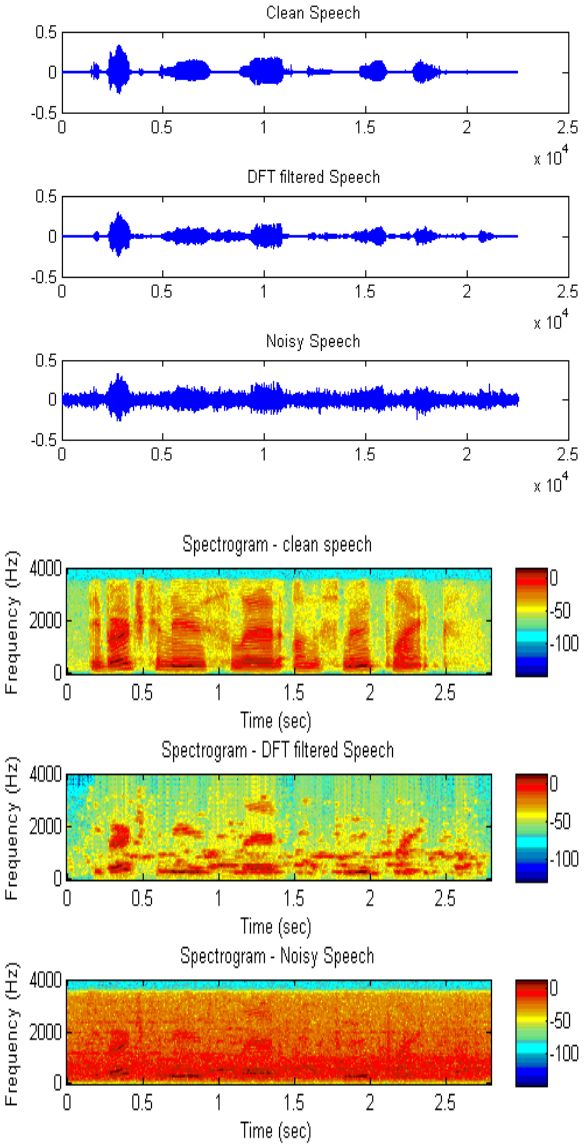


Fig. 2 : Top-Clean speech sp01
Middle- DFTF speech
Bottom- Clean speech sp01 corrupted by airport noise SNR0dB,alongwith the corresponding spectrograms

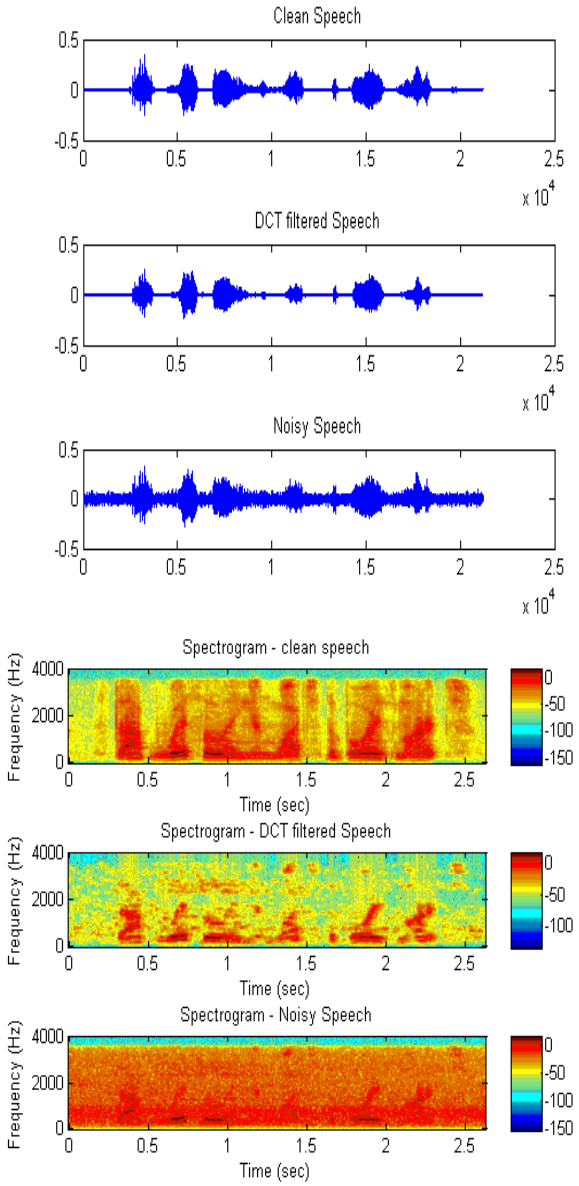
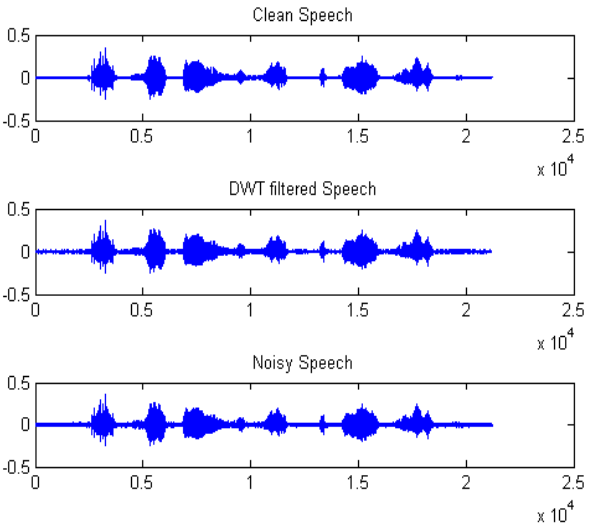


Fig. 3 : Top-Clean speech sp23
Middle- DCTF speech
Bottom- Clean speech sp23 corrupted by car noise SNR 5dB alongwith the corresponding spectrograms



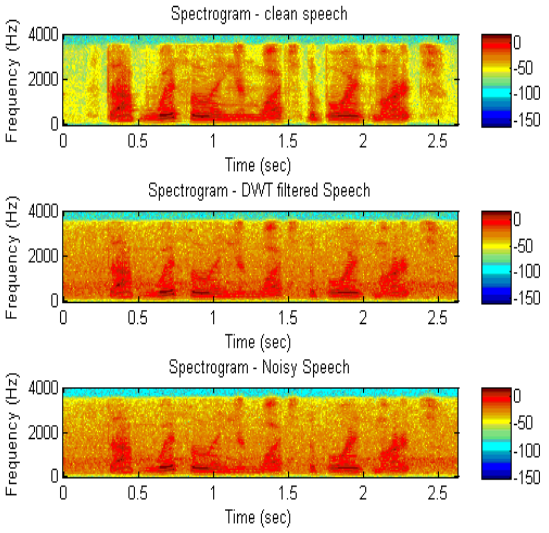


Fig.4 : Top-Clean speech sp23
Middle- DWTF speech
Bottom- Clean speech sp23 corrupted by babble noise SNR 15dB, alongwith the corresponding spectrograms

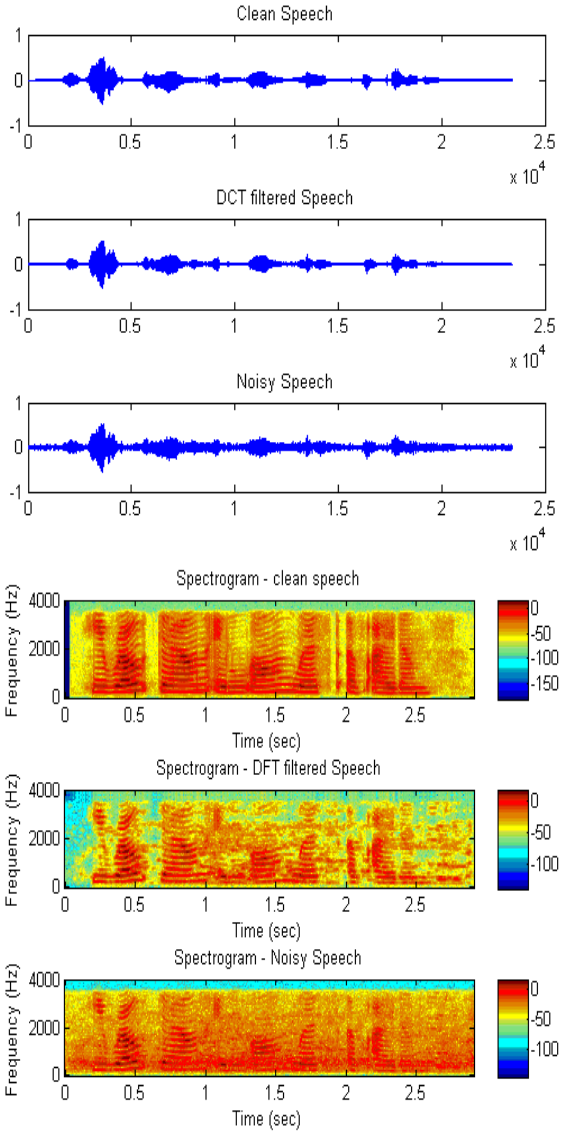


Fig. 5 : Top-Clean speech sp11
Middle- DFTF speech
Bottom- Clean speech sp11 corrupted by airport noise SNR 10dB,

alongwith the corresponding spectrograms

In order to see the effect of use of different transform domain in the proposed system the objective measure used is Segmental Signal to Noise Ratio (segSNR). segSNR is calculated as per[11]. Table I shows the segSNR values for different speech utterances for different noise conditions i.e, airport, babble and car noise for different signal to noise ratio.
Table II shows the overall SNR values for the same database.

TABLE I
segSNR for different speech utterances

Sp01 corrupted by airport noise				
noise SNR	**0 db**	**5 db**	**10 db**	**15 db**
DFTF	2.8576	3.7400	5.6436	7.4413
DCTF	2.4356	3.9566	5.2102	6.2111
DWTF	0.7380	5.0796	7.7466	12.8434
Sp11 corrupted by babble noise				
noise SNR	**0 db**	**5 db**	**10 db**	**15 db**
DFTF	1.6947	3.7198	7.7652	10.4945
DCTF	1.7805	3.2959	7.9166	10.1849
DWTF	-0.1631	2.5013	6.8872	10.5778
Sp23 corrupted by car noise				
noise SNR	**0 db**	**5 db**	**10 db**	**15 db**
DFTF	1.8659	3.3710	5.4066	8.1433
DCTF	1.6501	3.3915	4.8710	7.5561
DWTF	-0.2392	2.8557	6.7041	9.6589

TABLE III
Overall SNR for different speech utterances

Sp01 corrupted by airport noise				
noise SNR	**0 db**	**5 db**	**10 db**	**15 db**
DFTF	3.6870	7.8565	16.4927	25.7060
DCTF	3.7109	7.9936	15.5623	20.7131
DWTF	0.9980	3.1827	9.7251	29.4096
Sp11 corrupted by babble noise				
noise SNR	**0 db**	**5 db**	**10 db**	**15 db**
DFTF	3.8889	8.4003	21.7833	52.3191
DCTF	3.8894	8.3924	23.1455	50.6382
DWTF	1.0327	3.3546	10.2926	31.7167
Sp23 corrupted by car noise				
noise SNR	**0 db**	**5 db**	**10 db**	**15 db**
DFTF	4.3441	9.9840	21.3078	53.6275
DCTF	4.2589	8.9025	19.2750	47.8686
DWTF	1.0020	3.1377	9.9617	30.8854

IV. CONCLUSION & FUTURE SCOPE

From the observations from table I the good quality speech is obtained by DCTF for 0 dB SNR, DFTF for 5 dB SNR and DWTF for 10 dB and 15 dB SNR values. From the observations from table II we can see that overall SNR is high of DFTF than other two in maximum cases. Hence good quality speech can be obtained by DFTF. But DWTF gives good intelligibility at high SNR values. Using DFTF and DCTF quality of speech obtained is good but at the cost of intelligibility. Whereas DWTF gives good intelligibility of processed speech at the cost of quality.The same is shown in the fig. 2 to 5.
The effect of one level DWT is presented in this paper. The effect of higher level of decomposition is under investigation. Thus DFTF and DCTF gives good quality speech and DWTF gives intelligible speech (at high SNRs).

REFERENCES

1. Ephraim Y. and Malah D.," Speech enhancement using a minimum mean-square error log-spectral amplitude estimator", **Volume:** 33 , Issue 2 , pp. 443-445, IEEE trans. On Acoustics, Speech and Signal Processing, 1985
2. Boll S.," Suppression of acoustic noise in speech using spectral subtraction", Volume 27 ,Issue 2 ,pp. 113 - 120 , IEEE trans. On Acoustics, Speech and Signal Processing, 1979
3. B.Yegnanarayana ,Carlos Avendano ,Hynek Hermansky, P.Satyanarayana, Murthy, "Speech enhancement using linear prediction residual", Elsevier, Speech Communication 28, 25-42, 1999
4. Jensen J. , Hansen J.H.L., "Speech enhancement using a constrained iterative sinusoidal model**",** IEEE trans. On speech and Audio Processing, Vol 9, Issue 7, pp. 731 – 740, 2001
5. I.Y. Soon and S.N. Koh, "Speech enhancement using 2-D Fourier transform", IEEE trans. On speech and Audio Processing, vol 11 , Issue: 6,pp. 717 - 724 , 2003
6. Evans, JS Mason, MJ Roach ," Noise Compensation using SpectrogramMorphological Filtering", Proceedings 4th IASTED, 2002
7. Mrs. R. R. Mergu , Dr. S.K. Dixit , "A new paradigm for Plotting Spectrogram ", Journal of Information Systems & Communication, vol-3, Issue-1,pp.158-161, Feb 2012
8. Mrs. R. R. Mergu , Dr. S.K. Dixit , "Multi-Resolution Speech Spectrogram", International Journal of Computer Applications , vol 15, No. 4, Pgs. 28-32, Feb 2011
9. I.Y. Soon, S.N. Koh, C.K. Yeo,Noisy speech enhancement using Discrete Fourier transform", Elsevier, Speech Communication, Vol-24, pp.249-257, 1998
10. Joon Hyuk Chang, "Warped Discrete Cosine Transform Based Noisy Speech Enhancement", IEEE Trans on circuits & Systems-II Expressbriefs,Vol.52,No.9, pp.535-539, September2005
11. Yi Hu and Philipos C. Loizou,, "Evaluation of Objective Quality measures", IEEE trans. On speech and Audio Processing, vol 16, Issue 1, 2008