Transform Based Speech Enhancement Using DCT Based MMSE Filter,& Its Comparison With DFT Filter

Muhammad Safder Shafi[1], Mansoor Khan[1],
Naveed Abdul Sattar[1], Muhammad Rizwan[1], Affan Aziz Baba[1], Rizwan Ghani[1]

[1]COMSATS Institute of Information Technology (CIIT),
Islamabad, PAKISTAN
SafderShafi@yahoo.com

***Abstract*—In this paper we have illustrated the properties of Discrete Cosine Transform (DCT) as compared to the standard Discrete Fourier Transform (DFT) in case of noise removal from the speech. The results shows that DCT has better energy compaction and less computations as compared to DFT. The implementation and results of the DCT based minimum mean square error filter are described. The proposed algorithm is implemented for the reduction of residual noise by using probability of speech absence technique. The proposed techniques use adaptive schemes which will track the probability of speech absence in a noisy speech. It estimates the received spectral amplitude by a binary classification i.e. either the speech is present or absent state.**

Keywords-DCT; DFT; MMSE;residual noise

I. INTRODUCTION

The usage of speech processing systems for voice communication and recognition tasks are now very common due to the increasing power and the falling cost of digital signal processors. There are many examples in product form of speech processing systems that are using voice communication. The common example is the cellular radio telephony system. There are many other examples like use of hands-free input systems. The hands-free can be used for voice activated security systems in which recognition is involved. Voice dialing through hands free is also an example of voice recognition. The more natural way to communicate with the system is speaking rather than typing commands or giving input by either means. These systems will be efficient if they are faster in speech processing and accurate in speech recognition.

The presence of different noises like:

- Background noise
- Channel noise
- Quantization noise

degrade the system performance significantly like speech coders and voice recognition systems so we have to take some pre-processing step in these systems by incorporating speech enhancement to eliminate noise [1, 2, 3].

Different types of noise require different noise models and their own unique set of solutions. The scope of speech enhancement explored in our paper is focused on the suppression of *background noise*. (It has been an active research for more than 30 years)

Noise is present in many situations of daily life and it can never be avoided. Microphones will capture both undesired noise and desired speech. But the goal is to reconstruct the original speech signal. The process of a filtering should be done in order to filter a signal in order to remove noise. So we can define processing of filtering as follow:

"The process of extracting the information carrying signal X(n) from the observed signal Y(n) , where Y(n) = X(n) + N(n) and N(n) is a noise process , is called filtering. "

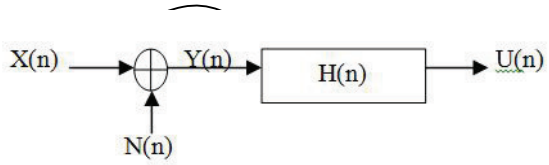

Figure 1: Typical Filtering Process

Different algorithms both in time and frequency domain are used to remove the noise embedded in the noisy speech signal. In our paper we discuss only frequency domain (DCT & DFT), as it is easier to separate the speech energy and noise energy in transform domain.

Furthermore, most of the algorithms, in order to remove noise from the noise corrupted speech signal, only modifies the spectral amplitudes of the noise

corrupted speech signal and leaves the noise corrupted phase information. Hence one of the main advantage of using the transformation is that the problem of not correcting for the phase will result in less severe consequences.

For the case of DCT considered in our paper the best estimate of the phase of the speech component is the phase of the corrupted signal itself. In DCT the coefficients are real.

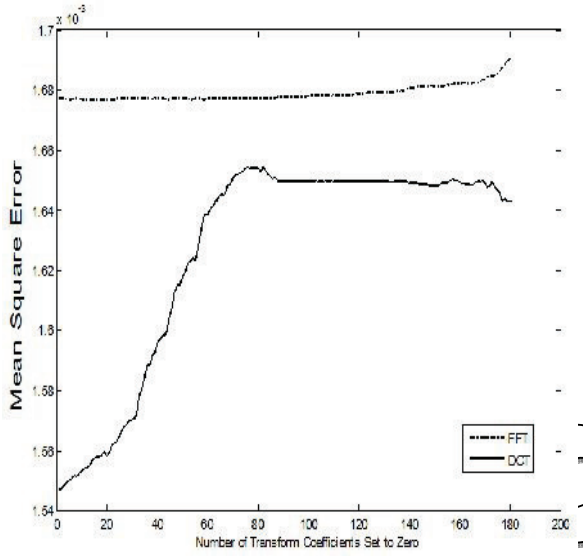

Figure 2: Comparison Of Energy Compaction

There is no phase component. Amplitude estimator is obtained on the assumption that amplitudes of both the noise and clean speech signal modeled by zero mean Gaussian distributed random variables in the frequency domain [4].

I. COMPARISON OF DCT WITH DFT

A. DCT Versus DFT

DCT compact the speech signal well as compared to DFT and because of its good energy compaction property, it is widely used in image compression applications, noise reduction etc. The reduction of noise can be obtained quite easily if the energy of the signal is compacted into few coefficients but if noise is white noise. Discrete Cosine Transform as compared to Discrete Fourier Transform provides higher compaction of energy that is get from the previous work on speech coding [5].

In this experiment show the energy compaction of the DCT from DFT has been illustrated using the actual speech signal. In this experiment we first take a sample clean speech and then it is divided into frames with 50% overlap, and the transform is performed.
After transformed into transform domain we take coefficients and sorted them according to their magnitudes and the n coefficients with lowest energy are set to zero. The clean speech is then reconstructed by using the weighted overlap and adds technique [6]. The mean square error is then computed. Then we plot mean square error and the coefficients n for both the DFT and DCT in Figure 2.

It can be clearly from the figure 2, that DCT provides better compaction and lesser MMSE as compared to DFT.

DFT only attempts to correct the noisy amplitude but not the phase component. The effects of phase on the signal discussed in [7], if the signal is out of phase by $\pi/8$ on which DFT is applied then the signal is distorted or rough. Therefore it is not possible to leave the phase, by doing this basically we damage the speech signal. But in case of DCT the coefficients are real. So there is no phase component which means DCT have an advantage over the DFT over the problem of the phase.

II. MINIMUM MEAN SQUARE ERROR FILTER FOR DCT

The following derivation uses the same notation as that was used in section 1. The transformed signals of the clean speech, noisy speech and the noise be denoted by $X(k), Y(k)$,and $N(k)$, respectively, where k denotes the position of the coefficients in the transform domain.$\hat{X}(k)$ represents the estimated amplitude. The assumption used here is that we use multiplicative filter to obtain the enhanced coefficients, the relationship between the noisy coefficients and enhanced coefficients can be expressed as follows:

$$\hat{X}(k) = W(k)Y(k) \qquad (1)$$

Here we will find an expression for W(k). W(k) Minimizes the mean square error between X(k) and $\hat{X}(k)$.

$$D(k) = E\left[\left(\hat{X}(k) - X(k)\right)^2\right] \qquad (2)$$

$$= E\left[\left(W(k)Y(k) - X(k)\right)^2\right] \qquad (3)$$

$$= E[W(k)^2Y(k)^2 - 2W(k)Y(k)X(k) + X(k)^2] \qquad (4)$$

where E[.] denotes the expectation operator. By differentiating Equation (3) with respect to W(K) and equating to zero, we have:

$$\frac{d(D(k))}{d(W(k))} = 2W(k)E[Y(k)^2] - 2E[X(k)Y(k)] = 0 \qquad (5)$$

This implies that

$$W(k) = \frac{E[X(k)Y(k)]}{E[Y(k)^2]} \qquad (6)$$

$$= \frac{E[X(k)(X(k)+N(k))]}{E[(X(k)+N(k))^2]} \quad (7)$$

$$= \frac{E[X(k)^2]+E[X(k)N(k)]}{E[X(k)^2]+2E[X(k)N(k)]+E[N(k)^2]} \quad (8)$$

Since X(k) and N(k) can be modelled as zero mean random variables which are independent of each other, E[X(k)N(k)]=0. Hence W(k) can be expressed as:

$$W(k) = \frac{E[X(k)^2]}{E[X(k)^2]+E[N(k)^2]} \quad (9)$$

Denoting $E[X(k)^2]$ by $\lambda_x(k)$ and $E[N(k)^2]$ by $\lambda_n(k)$, $\hat{X}(k)$. can be represented as

$$\hat{X}(k) = \frac{\lambda_x(k)}{\lambda_x(k)+\lambda_n(k)} Y(k) \quad (10)$$

Or

$$\hat{X}(k) = \frac{\xi(k)}{\xi(k)+1} Y(k) \quad (11)$$

Where

$$\xi(k) = \frac{\lambda_x(k)}{\lambda_n(k)} \quad (12)$$

$\xi(k)$ is known as the priori SNR by some authors. The above derivation shows that the MMSE amplitude estimator is the Wiener filter for the real transform case.

Methods for estimating the $\lambda_n(k)$ are discussed in detail in [8, 9]. So we assume value of $\lambda_n(k)$ is known in our paper. For estimating the $\lambda_x(k)$ we use the approach known as Decision Directed Estimation developed by Ephraim and Malah [10, 9] in our paper. The estimate $\widehat{\lambda_x}$ for $\lambda_x$is given by the following formula.

$$\widehat{\lambda_x}(k) = \alpha\widehat{\lambda_x}(k)_p + (1-\alpha)\max\{Y(k)^2 - \lambda_n(k), 0\} \quad (13)$$

Where $\max\{\}$ is the maximum function used to ensure that a non-negative value is obtained as an estimate.$\widehat{\lambda_x}(k)_p$show the estimated value of previous frame. $\alpha$is the constant, we adjust it in order to achieve better results, its significance is same as Nyquest theorem for sampling.

α = 0.8	Musical tone in the residual noise.
α = 0.98	Better results achieved.
α = 1	Severe distortions heard in Speech signal.

Table 1: Constant Settings For Better Results

I. RESULTS& DISCUSSIONS

We acquired the speech signal and noise from the real time environment via microphone in our paper. The duration of the speech signal and noise are 9 seconds and sampling frequency 8000 samples/second. The proposed speech enhancement algorithm tested on the clean speech data with bandwidth 4000 Hz and corrupted by the fan noise. Then corrupted speech data are divided into frames of 256 samples with an overlap of 192 samples with the neighboring frame. Then before enhanced to each frame, Hanning window is performed on each frame individually and pass through the DFT/DCT transform stage. The magnitude and phase of complex coefficients are then separated. The magnitude of the noisy speech is then filtered while the phase is left untouched. After the filtering is applied the filtered magnitudes are then combined to the phase and inverse transform is applied. The final enhanced speech is reconstructed from the enhanced frames using the weighted overlap and adds technique [11]. The block diagram of the proposed algorithm is shown in figure 3.

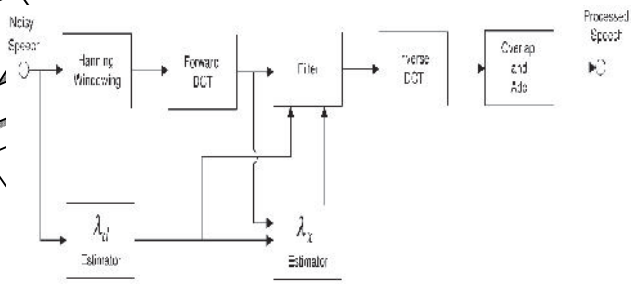

Figure 3: Block Diagram Of The Noise Reduction Filter [4]

DCT based speech enhancement filter describes in section III is implemented and its results are compared with DFT based speech enhancement filter. The outcome of designed filters on noisy speech is noticeably depicted in Figures 4 to 7. The clean speech and noise signal are depicted in the Figure 4 and 5.

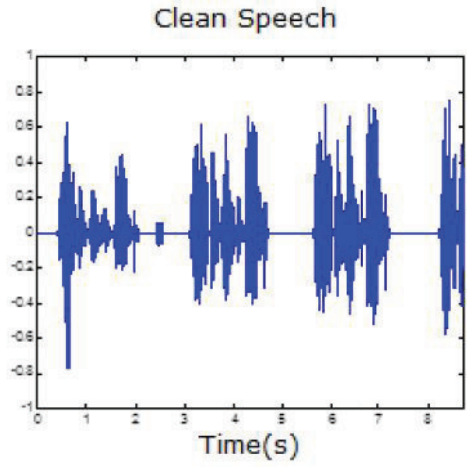

Figure 4: The desired speech signal

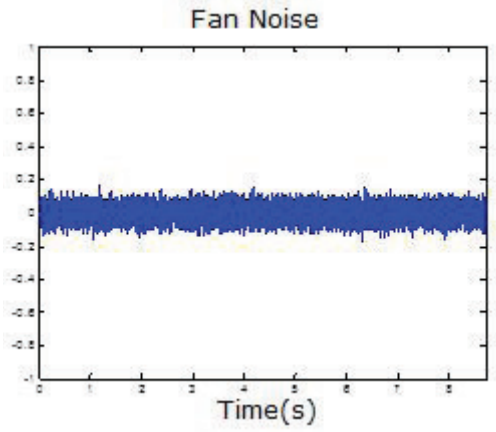

Figure 5: The Noise Signal

Figure 6 and 7 shows the effect of DFT and DCT filters on the noisy speech (clean speech + noise) and it can be seen that significant amount of noise has been removed.

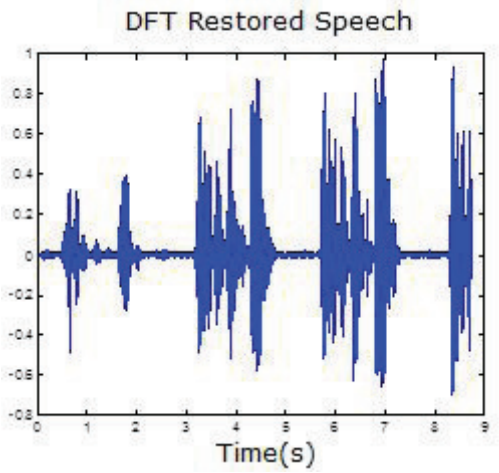

Figure 6: Signal obtained with DFT Filter

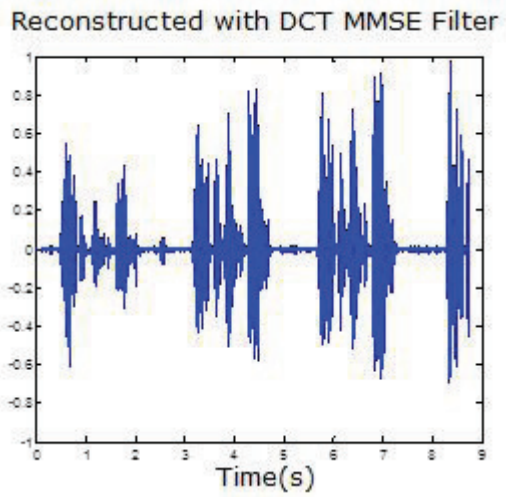

Figure 7: The signal obtained with DCT Filter

For clear understanding of noise reduction in transform domain, spectrograms for clean speech, noise and effects of different filters on the noisy speech are depicted in the Figure 8 to 11.

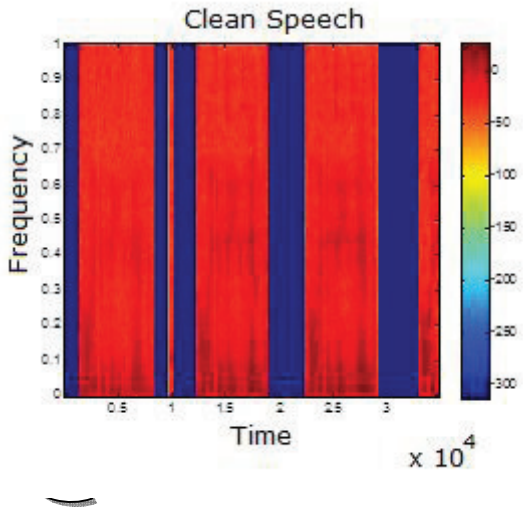

Figure 8: Spectrogram of desired speech signal

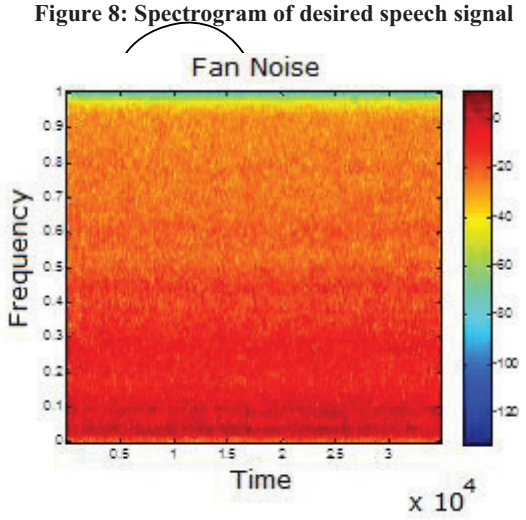

Figure 9: Spectrogram of noise signal

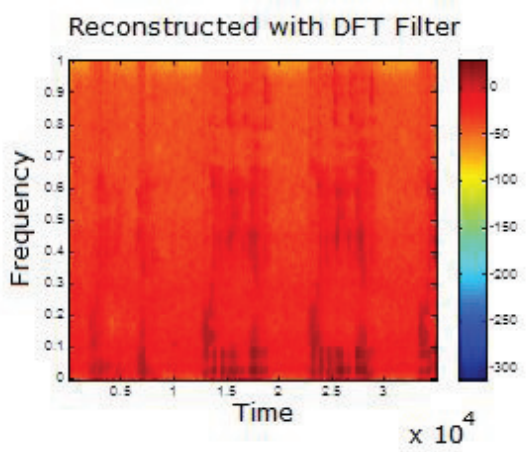

Figure 10: Spectrogram of signal obtained with DFT Filter

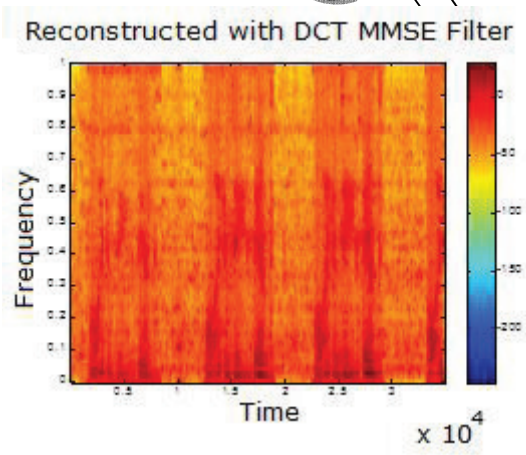

Figure 11: Spectrogram of signal obtained with DCT Filter

From the comparison between the clean speech and enhanced speech it is concluded that more than 85% of the noise is removed by using DCT filter and output signal is much more intelligible than its counterpart DFT filter output. Thus in the proposed schema DCT based MMSE filters outperforms DFT filters when used for speech enhancement in a real-time environment.

I. CONCLUSION

This paper presents the real-time transform based speech enhancement scheme proposed in [4] using MATLAB. The speech signal is corrupted by various noises. Table 2 gives the average powers of corrupted and processed speech signals in transform domain for various types of noises and shows speech enhancement by filtering out noise in transform domain based on DCT filters which is better than its alternative DFT due to better energy compaction property of DCT.

The improvement of overall schema is measured in terms Minimum Mean Square Error (MMSE) after obtaining Wiener filter coefficients $W(k)$ as described in section III that minimizes the mean square error between clean speech and processed speech. Table 3 compares Minimum Mean Square Error (MMSE) in transform domain between DCT coefficients of clean speech signal $X(k)$ and processed speech signal $\hat{X}(k)$(equation 2 of section III) for various noises.

Noise	Average Power (Corrupted Signal) (C) $\frac{1}{N}\sum_{i=1}^{N} Y_i^2(k)$	Average Power (After Filtering/Processing) (P) $\frac{1}{N}\sum_{i=1}^{N} \hat{X}_i^2(k)$	Amount of Filtered Noise F=C-P
Fan	5.6664e-004	2.5229e-005	5.41411e-004
F16 Plane	3.1e-003	6.1762e-005	3.03824e-003
Babble	3.3e-003	4.6923e-004	2.83077e-003

Table 2: Average Powers

Noise	MMSE
Fan	6.4146e-005
F16 Plane	9.5310e-005
Babble	5.0687e-004

Table 3: MMSE (Minimum Mean Square Error)

The results presented in this paper can serve as a benchmark for further research in speech signal processing techniques, in which proposed algorithms can be combined in various ways to produce better results for the filtered speech signal.

References

[1] M. R. Sambur and N. S. Jayant, "LPC analysis/synthesis from speech inputs containing white noise or additive white noise," *IEEE Trans. Acoust., Speech , Signal Processing*, vol. 24, pp. 484-494, 1976.

[2] B. H. Juang, "Recent Developments in speech recognition under adverse conditions," in *Proc. Int. Conf. Spoken Language Process,* (Kobe, Japan), pp. 1113-116, November 1990.

[3] J. H. L Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 795-805, 1991.

[4] Y. Soon, S. N. Koh, and C. K. Yeo, "Noisy speech enhancement using discrete cosine transform," Speech Communication, vol. 24, pp. 249-257, 1998.

[5] R. Zelinski and P. Noll, "Adaptive Transform Coding of Speech Signals," *IEEE Trans. On Acoust., Speech and Signal Processing*, vol. 25, pp. 89-95, 1979.

[6] R. E. Crochiere, "A weighted overlap-add method of short-time Fourier analysis / Synthesis," *IEEE Trans. , Speech, Signal Processing,* vol. 28, pp. 99-102, 1980.

[7] P. Vary, "Noise Suppression by Spectral Magnitude Estimation – Mechanism and Theoretical Limits," *Signal Processing*, vol. 8, pp. 287-400, 1985.

[8] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 27, pp. 113-120, 1979.

[9] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 137-145, 1980.

[10] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech Signal Processing*, Vol. 32, pp. 1109-1121, 1984.

[11] R. E. Crochiere, "A weighted overlap-add method of short-time Fourier analysis / Synthesis," *IEEE Trans. , Speech, Signal Processing,* vol. 28, pp. 99-102, 1980.