# A brief survey of a DCT-Based Speech Enhancement System

S.C.Shekokar, Prof. M. B. Mali

**Abstract-** Discrete Cosine Transform (DCT) has similar performance to the Karhunen-Loeve Transform (KLT) & same properties to the Discrete Fourier Transform (DFT). It is advantageous for speech enhancement as it provides better energy compaction capability. Though there is a perfectly stationary signal, frame to frame variations of DCT coefficients are observed. In pitch synchronous analysis DCT based speech enhancement system is introduced to overcome the problem. The window shift is based on pitch period hence the drawbacks of fixed window shift are reduced by DCT. Afterwards a Wiener filter with an adaptive controller is used to noise. This proposed speech enhancement system gives good quality of speech which can be used in no. of applications.

**Index Terms**–Pitch Synchronisation, Speech enhancement, Speech processing

## 1 INTRODUCTION

Speech enhancement can be performed in the time domain as well as in the frequency domain. Time domain filters include Finite Impulse Response (FIR) and Infinite Impulse Response (IIR) filters, Linear Predictive Coefficients (LPC) and Hidden Markov Model (HMM) etc. In transform domain techniques first transformation is performed on noisy speech before filtering then the inverse transformation take to give original speech. In Transform domain filters are those which calculate the transform coefficients first followed by the enhancement process. At the end, the inverse transform must be applied to obtain the enhanced speech. The main advantage of noise filtering process is relative ease of distinguishing and removing noise from the speech. Many speech enhancement algorithms prefer operating in the transform domain since the speech energy is not present in all the transform coefficients and it is therefore easier to filter off the noise especially for the noise-only coefficients. Different types of transforms may require different type's analysis methods.

DFT-based algorithms are the most active and popular among the number of transform based algorithms which are proposed in the past for single channel speech enhancement. One of the famous spectral subtraction algorithms which was extended to the Fourier transform by Boll has became a very popular method. Fourier domain is another important area of speech enhancement. In this more noise reduction takes place & it also reduced the level of tonal noise as compared to spectral subtraction. The fast Fourier transform is the property of DFT which is used reduces the computation load. Another widely used transform method is KLT, which has been applied to speech enhancement and it is used in one of the subspace speech enhancement algorithms. The main drawback of KLT-based algorithms is the high computational complexity.

DCT provide higher energy compaction as compared to DFT. Its performance is very similar to KLT. There is no fast Fourier transform method possible in KLT. But it has high energy compaction. Therefore DCT is widely used instead of KLT and also it has fast Fourier transform algorithm. Unlike the DFT the DCT coefficient are real and there is no phase component. Therefore, DCT should be a good choice for speech enhancement. Several advantages of DCT have been presented by Soon et al. in [1] in enhancing speech as compared to DFT. These advantages of DCT are as follows:

- It gives a significantly higher energy compaction capability.
- It is a real transform without phase information.
- It provides a higher resolution for analyzing the transform coefficients in the same window length.

In this paper the focus on a zoom-in portion of DCT during the frame based analysis together with an improved noise reduction filter. In DCT based speech enhancement algorithms, the transform is performed by a short-term cosine transform which is almost the same as short term Fourier transform (STFT) except that DCT is used instead of DFT. In such algorithms, there is fixed overlapping frame is present which can be ranging from 50% to 75% and then it is processed by DCT. Afterword, a noise suppression filter is applied on the DCT coefficients. But, this procedure does not take into account because of the differences between DCT and DFT. One of the major differences between DFT and DCT is that the DCT coefficients are real, while the DFT coefficients are complex and comprises a magnitude and phase representation. Without a phase representation, in DCT magnitudes is obtained by a standard window-shift which show much higher variation compared to those of DFT which is used only for a stationary signal. This will impact negatively on the inter-frame techniques such as the decision-directed approach for the estimation of a priori SNR. Therefore, pitch synchronous analysis is a good solution to compensate for this difference between DCT and DFT [4]. It helps to improve the performance of DCT based speech enhancement algorithms especially those using inter-frame techniques. An advanced speech enhancement system, named adaptive-time shift analysis (ATSA) system is proposed. This system also includes the pitch synchronous processing which will be improved by using a maximum alignment technique proposed. An adaptive Wiener filter in DCT domain will be introduced.

## 2 METHODOLOGY

The new speech enhancement technique is introduced an adaptive time-shift analysis speech (ATSA) enhancement system. This proposed technique work on the pitch period of speech. The proposed technique used in no. of applications.

**2.1 Speech Enhancement System with Pitch Synchronous Analysis:** This method of pitch synchronous analysis shifts the analysis window by the pitch period to obtain the speech segments and it will theoretically produce constant DCT coefficients for stationary signals. This is valid for only voiced speech signal, for unvoiced/silence part of speech signal is lags. Despite this, significant improvements can still be obtained as voiced speech is more dominant than unvoiced speech. The structure of this proposed adaptive time-shift analysis speech enhancement system [5] is shown in Figure 1. By using noise reduction technique voiced/unvoiced decision is made from the initial speech frame. If it contains voiced signal, the time-shift will be changed to one pitch period. or else the time-shift will fall back to the original fixed value. In this manner, the analysis window shift adapts to the underlying speech properties and it is no longer fixed.

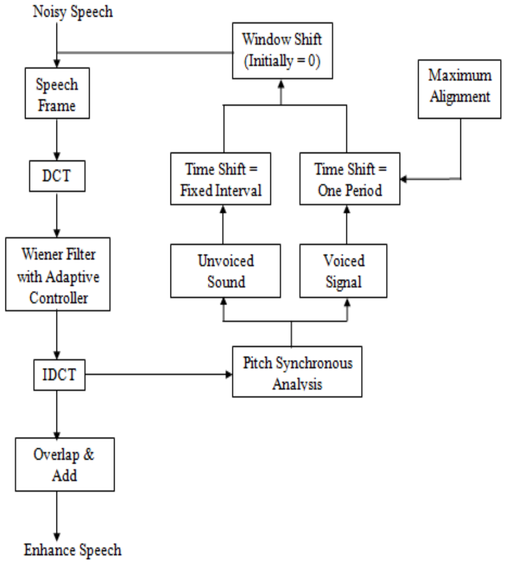

Fig.1. Block diagram of the proposed adaptive time-shift analysis speech (ATSA) enhancement system.

**2.2 Introduction to Windowing Function:** A window function can be applied when a signal is observed for a finite duration for truncating the signal, in signal processing. Rectangular window is the simplest window function which creates the well-known problem, spectral leakage effect. Spectral leakage effect means, if there are two sinusoids waves are present with same frequencies, then leakage are present with both waves interface with each other. If their frequencies are different, then leakage interferes are present of those having less amplitude than other. The main reason of using rectangular window is, it gives strong side-lobes in frequency domain where the first side-lobe is only around 13 dB lower than the main lobe. Similar to Fourier transform, DCT has the same problem with the rectangular window. The rectangular window also has some disadvantages such as discontinuities at the endpoints or maximum scalloping loss for frequency component that is exactly in the middle of two FFT coefficients.

Rectangular window also have some advantages.

1. Due to narrower main-lobe it is able to resolve comparable strength signals.
2. By using rectangular window there is no discontinuity problem at the end point in DCT as compared to DFT. Though DCT is based on an even symmetrical extension during the transform of a finite signal.

Though there are other windows such as Hann window one of the very popular [1] and trapezoidal window [2] use in speech analysis, this paper is focusing on rectangular window though it has some unsatisfactory aspects but some interesting properties.

**2.3 Weiner Filter with Adaptive controller:** The Wiener filter solves the signal estimation problem for stationary signals. The filter is optimal in the sense of the MMSE (Minimum Mean Square Error). It depends on estimation of the a-priori SNR which can be calculated as

Let the y=noisy speech, s=clean speech and n=noise signal, and their respective DCT representations are $Y_{m.k}$, $S_{m.k}$and $N_{m.k}$ where *m* is the time frame index and *k* is the frequency index. Then the a-priori SNR, ξ, can be expressed as follow:

$$\tilde{\xi}_{m,k} = \alpha \frac{\left|\tilde{S}_{m-1,k}\right|^2}{\lambda_N} + (1-\alpha)\max\left[\frac{|Y_{m.k}|^2}{\lambda_N} -1,0\right] \qquad (1)$$

Where $\tilde{S}_{m-1,k}$ = estimated clean speech in the previous frame, max is the maximum function

$\lambda_N$ = noise variance which equals to the expectation of the power magnitude of the noise signal, $E[|Nm,k|^2]$.

In above $\lambda_N$ is assumed to be known as noise signal is a wide sense stationary random process and can be computed during the silence period.

In Equation (1), the parameter *α* is used to set a proportion of contributions from the previous frames to the current estimate. In Fourier transform domain, the value of *α* is normally set to 0.98. The same value of *α* is also commonly used in DCT speech enhancement schemes [1], [4]. DCT coefficients may require a different value of *α* or even an adaptive one.

For estimation of adaptive controller Minimun Mean Square Error (MMSE) criteria is used along with decision directed approach. It leads to improved version of Weiner filter.

Recall the decision-directed approach in Equation (1), the a-priori SNR can be expressed as:

$$\hat{\xi}_{m,k} = \alpha_{m,k}\tilde{\xi}m\text{-}1,k + (1-\alpha_{m,k})\max\left(\gamma_{m,k} - 1,0\right) \qquad (2)$$

where $\alpha_{m,k}$ is an adaptive version of

$\alpha$, $\tilde{\xi}_{m-1,k} = |\ \tilde{S}_{m-1,k}|^2 / \lambda_N$ and $\gamma_{m,k} = |Ym,k|^2 / \lambda_N$

The error between estimated a-priori SNR $\hat{\xi}_{m,k}$and then real one $\xi_{m,k}$ is

$$J_\alpha = E\left\{\left(\hat{\xi}_{m,k} - \xi_{m,k}\right)^2\right\} \qquad (3)$$

**2.4 Pitch Synchronisation:** Ideally all algorithms work well only in clean situations therefore the above noise reduction filtering is performed first. As in this pitch synchronisation pitch period should extracted first.

The Wiener filtered speech $\hat{S}_{m,k}$ can be given by:

$$\hat{S}_{m,k} = \frac{\hat{\xi}_{m,k}}{\hat{\xi}_{m,k}+1} + Y_{m,k} \qquad (4)$$

where the estimated a-priori SNR $\hat{\xi}_{m,k}$ is obtained by using equation (2).

Following this noise reduction filtering, the enhanced speech after inverse DCT, $\hat{s}(n)$, is utilized for pitch detection to obtain a more accurate estimation.

Out various algorithm for detecting the pitch period the time domain autocorrelation method [3] is quite a common for solving this problem, since it is simple and robust for some noise corruption conditions. It is selected in this for extracting the pitch period to be used for the time-shift. The autocorrelation function of the resulting signal $\hat{s}(n)$ can be defined as:

$$R(n) = \sum_{m=0}^{N-m-1} \hat{s}(m)\, \hat{s}(n+m) \qquad (5)$$

Since the fundamental frequency in spoken English language is range bound between 80Hz to 500Hz. The frequency of a peal is defined it voiced or unvoiced. A distinct peak is defined to be greater than 0.5 times of $R(0)$. If no distinct peak found, it means that it is likely to be a silence or unvoiced frame. For voiced frames, the pitch period is extracted and used as the analysis window shift. In this method, the window length needs to be at least twice as long as the longest pitch period of the observed speech signals. The final enhanced speech is obtained by overlap & adds process. It is different from the original process due to the adaptive window shifting. A convenient solution is to produce a weighting function which records all the windows frame by frame and calculates the net weighting function. The weighting function can be calculated from the current and the previous frames and hence can be performed in real time. Thereafter, the enhanced speech has to be normalized by the weighting function.

The pitch synchronous analysis can be further improved by using maximum alignment technique. In this technique the speech analysis window starts from the short-term maximum amplitude of the speech signals and the time shift equals to one period. For more accurate pitch period compared with noisy speech Wiener filtered speech is used along with maximum alignment technique. Several impulses with period equal to the pitch period of the current voiced frame are generated to calculate the cross-correlation with the Wiener filtered speech. Let $\hat{s}(n)$ is the Wiener filtered speech sequence and $im(n)$ is the impulse sequence then the discrete cross-correlation of these two real signals is given by:

$$R_{\hat{s}im}(n) = \sum_{m=-\infty}^{\infty} \hat{s}(m) im(n+m) \qquad (6)$$

Where the impulse sequence $im(n)$ can be expressed as

$$im(n) = \sum_{m=-N}^{N} \delta(n - mT) \qquad (7)$$

where $\delta(n)$ is the delta function which equals to 1 only when $n = 0$, else $\delta(n) = 0$.
$T$ is the pitch period and $N$ could be $\infty$ or a fixed value which indicates the impulse train is infinite or finite respectively. In this paper, the number of impulses are empirically fixed to 5 on $N = 2$. After words, the position of the maximum amplitude of the current speech frame will be identified by tracking the maximum of the cross-correlation values.

## 3 CONCLUSION

There is variation in DCT coefficients from one frame to another in traditional DCT based algorithm where the observed speech signal is divided into fixed overlapping frames and transformed into DCT domain, because of non-ideal analysis window position. In order to reduce this variation in transform domain, a pitch synchronous analysis technique, is proposed. As autocorrelation function is used for detecting the pitch period which is in turn used as the amount of shift for the analysis window, which results into better noise reduction filtering. It can be further improved by maximum alignment which results in a much better fit to the DCT basis functions. Noise reduction will be applied to the observed signal is after truncating it into speech frames, to enhance the speech. An adaptive parameter $\alpha$ is proposed for a better estimate of the a-priori SNR which affects the optimal DCT filter in mean square error sense. The proposed techniques can be combined into a complete system named adaptive time-shift analysis (ATSA) speech enhancement system which produces good quality enhanced speech.

## 4 REFERENCES

[1] I. Y. Soon, S. N. Koh, and C. K. Yeo, “Noisy speech enhancement using discrete cosine transform,” *Speech Communication*, vol. 24, pp. 249–257, 1998.

[2] S. Ou, X. Zhao, and J. Dong, “Combining DCT and Adaptive KLT for Noisy Speech Enhancement,” in *Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on*, 2007, pp. 2857– 2860.

[3] L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal, “A comparative performance study of several pitch detection algorithms,” *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 24, no. 5, pp. 399–418, 1976.

[4] H. Ding and I. Y. Soon, “An adaptive time-shift analysis for DCT based speech enhancement,” in *Proceedings ICICS*, 2009, pp. 1–4

[5] Huijun Ding, Ing Yann Soon and Chai Kiat Yeo, “A DCT-based speech enhancement system with pitch synchronous analysis*”, IEEE Transactions on Audio, Speech and Language Processing* 2011.