An Advanced Speech Enhancement Approach with Improved Pitch Synchronous Analysis

V.R. Balaji\textsuperscript{1} and Dr. S. Subramanian\textsuperscript{2}

\textsuperscript{1}Assistant Professor/Department of ECE, Sri Krishna College of Engineering and Technology, Coimbatore. E-mail: balajivrresearch@gmail.com
\textsuperscript{2}Advisor, Coimbatore Institute of Engineering and Technology, Coimbatore. E-mail: drsraju49@gmail.com

Abstract: Speech enhancement has become an essential issue within the field of speech and signal processing, because of the necessity to enhance the performance of voice communication systems in noisy environment. There has been a number of research works being carried out in speech processing but still there is always room for improvement. The main aim is to enhance the apparent quality of the speech and to improve the intelligibility. Signal representation and enhancement in cosine transformation is observed to provide significant results. Discrete Cosine Transformation has been widely used for speech enhancement. In this research work, instead of DCT, Advanced DCT (ADCT) Transform which simultaneous offers energy compaction along with critical sampling and flexible window switching. In order to deal with the issue of frame to frame deviations of the Cosine Transformations, ADCT is integrated with Pitch Synchronous Analysis (PSA). Moreover, in order to improve the noise minimization performance of the system, Wiener filtering approach is used in this approach. Thus, a novel ADCT based speech enhancement using improved iterative filtering algorithm integrated with PSA is used in this approach.

Keywords: Wiener Filtering, Advanced Discrete Cosine Transform, Pitch Synchronous Analysis, Perceptual Evaluation of Speech Quality.

1. Introduction

Speech enhancement is the technique which enhances the quality of speech signals which are corrupted by adverse noise and channel distortion. Speech enhancement has been used in a number of applications in recent years [1]. The main aim of speech enhancement is to enhance the quality and clarity of the speech signal. A number of techniques have been developed for providing better clarity speech signals which comprises of the techniques such as spectral subtraction [2], Wiener filtering [3] and Ephraim Malah filtering [4].

For the past two decades, speech enhancement has become one of the most active researches in the field of signal process but still there are no standard techniques for both speech and noise [5]. Transform domain filters are widely used in the speech enhancement process. These filters compute the transform coefficients initially followed by the enhancement process. Finally, the inverse transform must be applied to attain the ultimate desired speech. A number of speech enhancement algorithms largely function in the transform domain as the speech energy is not present in all the transform coefficients and it is thus easier to filter off the noise particularly for the noise-only coefficients. Different transforms may require different analysis methods. For single-channel speech enhancement, a number of transform-based algorithms have been investigated in the past. Among these, DFT-based algorithms are the most active. Moreover, spectral subtraction algorithm [4] was extended to the Fourier transform by Boll [5] and became a very widely used approach.

This paper will focus on DCT during the frame-based analysis along with an improved noise reduction filter. In traditional DCT-based speech enhancement algorithms, the transform is carried out by a short-term cosine transform which is almost the same as Short-Term Fourier Transform (STFT) except that DCT is used rather than DFT. In such algorithms, the observed speech is partitioned into fixed overlapping frames ranging from 50\% to 75\% and then processed by DCT.

Moreover, a noise suppression filter is applied on the DCT coefficients. One of the key differences is that the DCT coefficients are real, while the DFT coefficients are complex and it consists of a magnitude and phase representation. Without a phase representation DCT coefficients’ magnitudes obtained by a standard window-shift illustrate much higher variation compared to those of DFT for a strictly stationary signal. This will have influence negatively on the inter-frame approaches such as the decision directed approach [7] for the assessment of \textit{a priori} SNR.

ADCT is widely used in audio processing, where the overlapping minimizes artifacts from the block boundaries [8]. Hence, in this research work uses ADCT in order to improve the speech quality. Hence, pitch synchronous analysis is an efficient key which also helps in offering better performance [9]. This would improve the overall performance of DCT-based speech enhancement algorithms especially those using inter-frame techniques. This system also incorporates the pitch synchronous processing which will be improved by a Maximum Alignment technique.
Thus, an advanced speech enhancement system namely Advanced Discrete Cosine Transform Speech Enhancement (ADCT based WFPS) is proposed in this approach.

2. Literature Survey

A number of DFT based techniques only concentrates to filter the spectral magnitude while leaving the noise corrupted phase information intact, as it has been reported that the best estimate of the phase is the corrupted one itself [7]. DCT can attain a higher upper bound than DFT since no such action generally results in an upper bound on the maximum possible improvement in Signal-To-Noise Ratio (SNR) [6]. DFT only creates about half the independent spectral constituents as the other half are complex conjugates, while DCT creates fully independent spectral components. Depending on these benefits, it is also proven that DCT is a suitable choice to the discrete Fourier transform (DFT) for speech enhancement [10].

Pitch synchronous analysis has been earlier used in various speech signal processing systems such as speech analysis/synthesis system [11] [12], prosody modification system [13] and speech recognition system [14]. The fundamental scheme of pitch synchronous processing is to initially partition the speech signal into pitch periods for the voiced sounds and into pseudo pitch periods for unvoiced sounds. A number of different processes can then be applied on the resulting pitch synchronous segments for various functions.

Pitch synchronous Overlap Adds (PSOLA) technique is applied in the time domain and it makes the algorithm to be competent to control the value of the synthesized pitch and the duration of the synthesized signal [13]. PSOLA technique can also be used in other domains such as frequency domain [15]. Fourier transform is applied on the pitch synchronous sections and the resulting spectra are approximated by a pattern of zeros and poles to attain the pitch synchronous depiction for examining the voiced sounds [16]. Evangelista [17] also uses this pitch synchronous representation and utilizes Wavelet transform on it to attain a new depiction of pseudo-periodic signal through a regularized oscillatory component and fluctuations. This depiction provides a number of scales for examining the fluctuations which is superior to Fourier representation with only one scale.

Pitch synchronous speech segments are transferred to linear prediction residual on which the DCT is applied for resampling the residual signal by truncation or zero padding [18]. DCT is applied as an application tool as DCT is efficient at energy compaction. The energy loss with the DCT-based linear prediction technique is lesser than that with the direct linear prediction technique and this algorithm is thus superior to the original fundamental algorithm.

Most of the existing research work demonstrates that the pitch synchronous processing assists in minimizing the discontinuities connected with windowing and it focuses on a key point, which is the pitch period [19]. Pitch synchronous processing has been extensively applied in speech processing but is being rarely been used for the purpose of speech enhancement [20].

Line Spectrum Pair (LSP) was first introduced by Itakura [36] [37] as an alternative kind of LPC spectral representation.

3. ADCT and WFPS Pitch Synchronous Based Speech Enhancement

The structure of this proposed speech enhancement system is shown in Figure 1. The initial speech frame is filtered by a noise reduction technique, and then a voiced/unvoiced decision is made. If it contains voiced signal, the time-shift will be changed to one pitch period. Otherwise, the time-shift will fall back to the original fixed value. In this way, the analysis window shift adapts to the underlying speech properties and it is no longer fixed [21].

In order to improve the performance, Advanced Discrete Cosine Transform is used in this approach. Signal representation in ADCT domain has become an active area of research in signal processing. ADCT is being effectively used in superior quality audio coding due to its unique characteristic features. The main advantage of ADCT is its energy compaction capability. Moreover, it also attains critical sampling, a minimization of block effect and flexible window switching [8].

In certain applications such as streaming audio to handheld devices, it is very essential to have quick implementations and optimized codec structures. In many circumstances, it is also efficient to carry out ADCT domain audio processing such as error concealment, which lessens the deprivation of subjective audio quality.

The above said characteristic features of ADCT motivated the application of ADCT in this research work.

The direct and inverse ADCT are defined as [23, 24]:

\[
\alpha_r = \sum_{k=0}^{2N-1} \tilde{a}_k \cos \left[ \frac{\pi}{N} \left( k + (N + 1)/2 \right) (r + 1/2) \right] \\
\tilde{a}_k = \frac{2}{N} \sum_{r=0}^{N-1} \alpha_r \cos \left[ \frac{\pi}{N} \left( k + (N + 1)/2 \right) (r + 1/2) \right] \\
k = 0, \ldots, 2N - 1
\]  

(1)

Where \( \tilde{a}_k = h_k a_k \) is the windowed input signal, \( a_k \) is the input signal of 2N samples, \( h_k \) is a window function. We assume an identical analysis-synthesis time window. Certain limitations of perfect reconstruction are [25, 26]:

\[
\alpha_r = \sum_{k=0}^{2N-1} \tilde{a}_k \cos \left[ \frac{\pi}{N} \left( k + (N + 1)/2 \right) (r + 1/2) \right] \\
\tilde{a}_k = \frac{2}{N} \sum_{r=0}^{N-1} \alpha_r \cos \left[ \frac{\pi}{N} \left( k + (N + 1)/2 \right) (r + 1/2) \right] \\
k = 0, \ldots, 2N - 1
\]
A sine window is widely used in audio coding because it offers good stop-band attenuation, gives good attenuation of the block edge effect and allows perfect reconstruction. Other optimized windows can be applied as well [25]. The sine window is defined as:

\[ h_k = \sin[\pi (k + 1/2)/2N] \quad k = 0, ..., 2N - 1 \]  

(3)

\( \hat{a}_k \) in (1) are the IADCT coefficients of \( \alpha_r \). It contains time domain aliasing

\[ \hat{a}_k = \left\{ \begin{array}{ll} \hat{a}_k - \hat{a}_{N-k}, & k = 0, \ldots, N - 1 \\ \hat{a}_k + \hat{a}_{3N-k}, & k = N, \ldots, 2N - 1 \end{array} \right. \]  

(4)

Rectangular window does have some advantages. It has a narrower main-lobe which is able to resolve comparable strength signals. Besides, one advantage of using the DCT as compared to DFT is that there is no discontinuity problem caused by rectangular window at the endpoints, since DCT is based on an even symmetrical extension during the transform of a finite signal.

Therefore, the selection of the window is based on a tradeoff between spectral resolution and leakage effects. In the literature of DCT-based speech enhancement algorithms, the Hann window is very popular [6]. There is also a compromised adoption with trapezoidal window being applied in [22]. In this paper, rectangular window is used for better performance of the system with Advanced Discrete Cosine Transform.

Wiener Filtering

Wiener filter has been proven to be the optimal filter for the real transform in mean square error (MSE) sense. During the implementation, it fully depends on the estimation of the \textit{a priori} SNR. The \textit{a priori} SNR can be computed by many ways among which the decision-directed approach [7] is widely used. Let the noisy speech, clean speech and noise signal be denoted as \( y, s \) and \( n \), respectively, and their ADCT representations are \( Y_{mk}, S_{mk} \) and \( N_{mk} \) where is the time frame index and is the frequency index.

This paper proposes a smoothed noise update technique that uses the estimated signal spectrum for subsequent signal estimation. It leads to a more efficient result than the soft-decision based noise estimate found in literature. Further, the WF performance is improved using codebook constraints in the LAR domain instead of LSP domain.

The Wiener filter is a popular statistical approach based on the assumption that signal and the noise are stationary linear stochastic processes with known spectral characteristics that has been used for noise reduction in speech signal. Assuming that the clean speech, \( s(t) \), degraded by an additive noise, \( w(t) \). The noisy speech, \( x(t) \) is defined as [27]

\[ x(t) = s(t) + w(t) \]

Wiener filter is an optimal filter that minimize the Mean Squared Error (MSE). In case of Eq.(10), the filter can be defined as

\[ S(\omega) = H(\omega)X(\omega) \]

Where \( \omega \) is the frequency index and \( S(\omega), X(\omega) \) and \( H(\omega) \) are the discrete Fourier transform of clean speech, noisy speech and that of the Wiener filter respectively. The MSE is defined as follows. The error is defined as:

\[ E(\omega) = S(\omega) - \delta(\omega) \]

\[ = S(\omega) - H(\omega)X(\omega) \]
The Mean Squared Error of Eq.(10) is defined as:

\[ E[|E(\omega)|^2] = E[|S(\omega) - H(\omega) \cdot X(\omega)|^2] \]

Where \( E[.] \) stands for expectation operator. To minimize the MSE, the Wiener filter can be estimated

\[
\frac{\delta E[|E(\omega)|^2]}{\delta H(\omega)} = 2H(\omega)E[|X(\omega)|^2] - 2E[|X(\omega)S^*(\omega)|] = 2H(\omega)P_{XX}(\omega) - 2P_{XS}(\omega) = 0
\]

Where \( P_{XX}(\omega) \), \( P_{XS}(\omega) \) are the power spectra of noisy speech and cross power spectra between noisy speech and clean speech respectively.

If there is no correlation between the speech signal \( s(t) \) and additive noise \( w(t) \), the power spectrum of the noisy speech and the cross power spectrum can be transformed as:


\[ P_{XS}(\omega) = E[|X(\omega)S^*(\omega)|] = E[|S(\omega)|^2] + E[|W(\omega)|^2]S^*(\omega)] = E[|S(\omega)|^2] = P_{ss}(\omega) \]

Consequently, the Wiener filter can be derived as follows:

\[ H(\omega) = \frac{P_{ss}(\omega)}{P_{ss}(\omega) + P_{ww}(\omega)} \]

The SNR is defined by

\[ SNR = \frac{P_{ss}(\omega)}{P_{ww}(\omega)} \]

The definition can be incorporated to the Wiener filter equation as follows:

\[ H(\omega) = \left[ 1 + \frac{1}{SNR} \right]^{-1} \]

### Spectral Subtraction Based Initialization (SSI)

For each frame in sequential MAP calculation, a set of initial values for vector \( a \) denoted as \( a_o \) is assumed based on which the speech vector \( \hat{s} \) is calculated through the Wiener filter. The current estimate \( \hat{s} \) is in turn used to estimate the next estimate of \( a \). This procedure is continued until convergence is achieved. In [28], \( H(\omega) \) is started as unity which is highly suboptimum. Therefore, it results in two possibilities. The first possibility is that the iterations might converge in such a way that the resulting filter is not perceptually the best. The second possibility is that, though they do converge to an optimal filter, number of iterations taken for convergence will be large. Hence, an initialization technique which provides efficient and quicker convergence is needed. A Spectral Subtraction based Initialization (SSI) method is proposed to deal with the above issues. For each and every frame, power spectral subtraction [29] is performed to obtain the enhanced speech estimate. Following LPC analysis, the above estimate gives \( a_o \) which determines \( H(\omega) \). It is obvious that \( H(\omega) \) is better than starting with a unity WF and it results in better convergence properties of CIWF.

### Robust Parameter Domain Search

The significance of WF lies in approximating the optimum filter by means of a codebook of clear speech vectors. Hence, the parameter space utilized to denote these vectors has a considerable bearing on the successive approximations. Line Spectral Frequencies (LSF), Reflection Coefficients (RC) and Log Area Ratios (LAR) have a one-to-one mapping but they also have different clustering attributes due to the non-linear relationships between them. Hence, each has been used with varied success in speech coding and recognition. In this work, a number of different parameter spaces are explored for WF to discover the best performing parameter. The widely used IS distance measure is used for creating LPC codebooks. The Euclidean Distance (ED) is used for LAR and RC codebooks. For LSPs, ED and other two perception based weighted Euclidean distances such as the Mel-Frequency Warping (MFW) based distance which is modeled on the auditory system and the Inverse Harmonic Mean (IHM) based distance are presented. IHM based distance is perceptually appropriate as it weights each LSF in the inverse proportion of its nearness to its neighbors because of the improved possibility of it denoting formants [30].

The estimated \( a \) priori SNR, can be expressed as follows:

\[ \hat{\xi}_{m,k} = \alpha \frac{|\hat{s}_{m-1,k}|^2}{\lambda_N} + (1 - \alpha)\max \left\{ \frac{|Y_{m,k}|^2}{\lambda_N} - 1, 0 \right\} \] (7)

where \( \hat{s}_{m-1,k} \) is the estimated clean speech in the previous frame, \( \max \) is the maximum function and \( \lambda_N \) is the noise variance which equals to the expectation of the power magnitude of the noise signal, \( E[|N_{m,k}|^2] \). The noise variance is assumed to be known since noise signal is a wide-sense stationary random process and can be computed during the silence period. In (1), the parameter is used to set a proportion of contributions from the previous frames to the current estimate. In Fourier transform domain, the value of \( \alpha \) is normally set to 0.98 which is an empirically obtained value and is known to be a good tradeoff between noise reduction and speech distortion. The same value of is also commonly used in DCT speech enhancement schemes [20]. However, this might not be proper for the new situation since DCT coefficients may require a different value of or even an adaptive one.

In DFT domain, there are some work about adapting for better estimation of the \( a \) priori SNR [31]. Thus, it is
workable to propose an adaptive decision-directed approach in DCT domain which leads to an improved version of Wiener filter. The minimum mean square error (MMSE) criterion is used to derive the optimal expression. Recall the decision-directed approach in (1), the a priori SNR can be expressed as

\[ \hat{\xi}_{m,k} = \alpha_{m,k} \hat{\xi}_{m-1,k} + (1 - \alpha_{m,k}) \max(y_{m,k} - 1, 0) \]  

(8)

Where \( \alpha_{m,k} \) is an adaptive version of \( \alpha \), \( \hat{\xi}_{m-1,k} = |\hat{\xi}_{m-1,k}|^2 / \lambda_N \) and \( y_{m,k} = |Y_{m,k}|^2 / \lambda_N \). Then the error between the estimated a priori SNR \( \hat{\xi}_{m,k} \) and the real one \( \xi_{m,k} \) is

\[ E = E \left\{ (\hat{\xi}_{m,k} - \xi_{m,k})^2 \right\} \]  

(9)

If \( E(y_{m,k} - 1) \) is set to \( \hat{\xi}_{m,k} \) which is reasonable, then (9) can be rewritten as

\[ E = \alpha^2_{m,k} (\hat{\xi}_{m-1,k} - \xi_{m,k})^2 - (1 - \alpha_{m,k})^2 \hat{\xi}^2_{m,k} + (1 - \alpha_{m,k})^2 E \left\{ (y_{m,k} - 1)^2 \right\} \]  

(10)

Based on the assumption that DCT coefficient of speech signal \( S_{m,k} \) and noise signal \( N_{m,k} \) can be modeled as zero mean random Gaussian variables which are independent of each other [6], \( E(y_{m,k} - 1)^2 \) can be expressed as

\[ E \left\{ (y_{m,k} - 1)^2 \right\} = 3\xi^2_{m,k} + 4\xi_{m,k} \]  

(11)

is used in (11) based on the assumption that the DCT coefficient of speech signal \( S(m, k) \) has a Gaussian distribution. Incorporating (10) and (11), the error can be finally obtained by

\[ E = \alpha^2_{m,k} (\hat{\xi}_{m-1,k} - \xi_{m,k})^2 + (1 - \alpha_{m,k})^2 (2\xi^2_{m,k} + 4\xi_{m,k}) \]  

(12)

Equating \( \partial E / \partial \alpha_{m,k} \) to zero, the optimal expression of \( \alpha_{m,k} \) can be obtained as

\[ \alpha_{m,k} = \frac{2\xi^2_{m,k} + 4\xi_{m,k}}{(\hat{\xi}_{m-1,k} - \xi_{m,k})^2 + 2\xi^2_{m,k} + 4\xi_{m,k}} \]  

(13)

The approximation used above is to avoid division by zero. As \( \hat{\xi}_{m,k} \) is unknown, (7) cannot be applied directly. An approximate value of \( \alpha_{m,k} \) can be obtained by substituting \( \xi_{m,k} \) with \( \hat{\xi}(m, k) \) which is defined as follows:

\[ \hat{\xi}_{m,k} = (y_{m,k} - 1) \ast H(m) \]  

(14)

Where \( \ast \) is the convolution operator, \( H(m) \) is a low-pass filter, and a Gaussian mask is applied here to realize this smoothing function of \( H(m) \). The reason for applying this low-pass filter is that it is able to reduce the variance among different speech frames which are caused by noise. This annoying effect can be further reduced by a “moving” value of \( \alpha_{m,k} \)

\[ \hat{\alpha}_{m,k} = \beta \hat{\alpha}_{m-1,k} + (1 - \beta)\alpha_{m,k} \]  

(15)

\( \beta \) represents a parameter which is fixed to 0.5 for the experimental evaluations. From the above equation SNR changes slowly, the parameter \( \alpha_{m,k} \) will be a value close to one. If the SNR has sharp changes, the parameter will take a smaller value enabling to change adaptively. Thus, the adaptive controller is in the range of zero to one.

Pitch Synchronization

In order to implement the ADCT based WFPS algorithm, the pitch period should be extracted first. There are many ways to estimate the pitch periodicity of a speech signal. From periodicity in time or from frequently spaced harmonics in frequency domain the pitch can be predicted. A time domain pitch estimator needs a preprocessor to filter and make simpler the signal through data reduction, basic pitch estimator and a post processor to correct errors.

The autocorrelation approach is mainly used in time domain method for calculating pitch period of a speech signal [38]. For a discrete signal \( x(n) \), the autocorrelation function is

\[ R(m) = \lim_{N \to \infty} \frac{1}{2N + 1} \sum_{n=-N}^{N} x(n) \cdot x(n + m - M_0) \]  

(16)

where \( N \) is the length of analyzed sequence and \( M_0 \) is the number of autocorrelation points to be computed. For pitch detection assume \( x(n) \) is periodic sequence, that is \( x(n) = x(n + P) \) for all \( n \), it is shown that the autocorrelation function is also periodic with the same period, \( R(m) = R(m + P) \). On the contrary, the periodicity in the autocorrelation function point out periodicity in the signal. For a non-stationary signal like speech, the longtime autocorrelation is calculated from (16). Generally operate with short speech segments, consisting of finite number of
samples so the autocorrelation based PDAs short-time autocorrelation function is as below.

\[ R(m) = \frac{1}{N} \sum_{n=0}^{N-1-m} x(n) \cdot x(n+m), 0 \leq m \leq M_0 \]  

(17)

The variable \( m \) in (17) is called delay and the pitch is equal to the value of \( m \) which results in the maximum \( R(m) \). In the proposed approach the pitch period is calculated using autocorrelation method.

The final enhanced speech is obtained by overlap add process. Actually, this process is a little 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.

A New Pitch Synchronous OverLap and Add (PSOLA) Approach to enhance the pitch synchronous analysis

A pitch mark location method is modified for signals with varying fundamental frequency. The analysis stage intend to iteratively collecting sound samples from the input signal at equally spaced fundamental frequencies.

\[ F_{0i} = F_{0i}^{\text{min}} + \frac{F_{0i}^{\text{max}} - F_{0i}^{\text{min}}}{M - 1} \quad i = 0, M - 1 \]  

(18)

where \( M \) is the total number of sound samples to extract.

1. Calculate the growth of the fundamental frequency of the input signal. It is hard to detect, this step is performed by calculate approximately the evolution of the most energetic harmonic, \( k_{\text{max}} F_0 \).

2. This component is got from the input signal with a selective time varying passband filter. The central frequency of the filter is noted at every sample to match the local approximation. Preferably, the resulting signal is a single sinusoid modulated in frequency and amplitude according to the \( k_{\text{max}} F_0 \) evolution and remains in phase with the input signal.

3. Pitch marks are placed so for recalling \( k_{\text{max}} = N_c \) these are placed in the input signal at the initial level for every \( N_c^{\text{th}} \) period of the filtered signal obtained in step 2. For each frequency \( F_{0i} \), a single pitch mark is chosen as the one equivalent to the closest fundamental.

4. For each selected pitch mark, a sound sample is extracted from the input signal with an suitable temporal window. The additive noise \( \omega F_0 \) is naturally extracted with

the harmonic part of the signal and requires no additional operations [39].

\[ x_{F_0}[n] = \sum_{k} A_{F_0,k} \sin \left( 2\pi \frac{k F_0}{F_s} n + \phi_{F_0,k} \right) + \omega F_0[n] \]  

(19)

where \( n \) is the discrete time index, \( F_0 = F_c \) is the fundamental frequency, \( A_{F_0,k} \) and \( \phi_{F_0,k} \) are the amplitudes and the initial phases of the harmonics and \( \omega F_0[n] \) is the stochastic component.

4. Experimental Results

For this experimental setup, a hundred different segments of speeches (half females and half males), are randomly chosen from the TIMIT database. They are resampled at 8 kHz and corrupted by three additive noise types including white noise, fan noise and car noise. The total speech duration of all these test speech segments is 313.998 s including the silence period. Approximately 50% of the speech segments are classified as voiced speech.

The proposed ADCT based WFPS technique is evaluated using two objective measures, segmental SNR (SegSNR) measure and perceptual evaluation of speech quality (PESQ) measure. Since SegSNR is better correlated with mean opinion score (MOS) than SNR as indicated by [33] and is easy to implement and it has been widely used to qualify the enhanced speech. The implementation in [34] is adopted here such that each frame with segmental SNR is thresholded by a dB lower bound and a 35 dB higher bound. The segmental SNR is defined by [20]

\[ \text{SegSNR} = \frac{10}{|Y|} \sum_{l=0}^{N/2} \log \frac{\sum_{k=0}^{N/2} |X(k,l)|^2}{\sum_{k=0}^{N/2} |D(k,l)|^2} \]  

(20)

Where \( Y \) represents the set of frames that contain speech and \( |Y| \) its cardinality. PESQ which is described in ITU-T recommendation P.862 and is also published in [35] is an objective measurement tool that predicts the results of subjective listening tests on telephony systems. It uses a sensory model to compare the original, unprocessed signals with the enhanced signals. In [34] it is indicated that the SegSNR is a better evaluation in terms of noise reduction, while the PESQ is more accurate in terms of speech distortion prediction. The latter is also more reliable and highly correlated with MOS as compared to other traditional objective measures. In most situations, PESQ is the best objective indicator for overall quality of enhanced speech. Before evaluating the ADCT based IWFPS system, the effects of window functions should be presented. Iterative Wiener filter with fixed time-shift analysis of 8 ms is utilized. Two different window functions, rectangular window and Hann window are used to truncate the input signal. The window length is fixed to 32 ms. SegSNR and PESQ results are shown in Figs. 2 and 3, respectively. From these two figures, it is clear that rectangular window is better for DCT
based noise reduction algorithms. For all the noise types taken for consideration, rectangular window is observed to provide better Segmental SNR.

![White Noise](image1)

![Fan Noise](image2)

![Car Noise](image3)

Figure 2: Segmentsal SNR results of noisy speech, Wiener filtered speech with rectangular window and Hann Window

To exhibit the advantages of each component of the proposed ADCT based WFPS system, three speech enhancement schemes are compared. The first approach is Wiener filtering with a higher fixed overlap which can be denoted as WFHO. The second one is the pitch-synchronized Wiener filtering named as PSWF. The third approach is the Adaptive Time-Shift Analysis speech (ATSA) approach.

Table 1 shows the comparison of SegSNR results. The comparison is carried out for three noise types such as White noise, Fan noise and Car noise. The Input SNR taken for experimentation are 0, 5, 10 and 15. For white noise, the proposed ADCT based WFPS provides efficient SegSNR for all the SNR input values taken for consideration. Similarly for the other noise types, the proposed ADCT based WFPS approach outperforms the other approaches taken for comparison.

<table>
<thead>
<tr>
<th>Noise Type</th>
<th>SNR (dB)</th>
<th>WFHO</th>
<th>DCT based PSWF</th>
<th>ATSA</th>
<th>ADCT based WFPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>0</td>
<td>5.23</td>
<td>5.24</td>
<td>5.48</td>
<td>5.53</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>4.52</td>
<td>4.53</td>
<td>4.86</td>
<td>4.88</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>3.48</td>
<td>3.53</td>
<td>3.95</td>
<td>4.01</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>2.52</td>
<td>2.56</td>
<td>3.09</td>
<td>3.16</td>
</tr>
<tr>
<td>Fan</td>
<td>0</td>
<td>8.84</td>
<td>8.97</td>
<td>9.26</td>
<td>9.32</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>8.76</td>
<td>8.94</td>
<td>9.29</td>
<td>9.40</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>8.27</td>
<td>8.52</td>
<td>8.84</td>
<td>9.93</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>7.42</td>
<td>7.71</td>
<td>8.01</td>
<td>8.18</td>
</tr>
</tbody>
</table>

TABLE I

COMPARISON OF \(\Delta\) SEGSNR RESULTS
Discrete Cosine Transform is integrated with pitch used as the amount of shift for the analysis window. The autocorrelation function is used for detecting the pitch period which is in turn improved by maximum alignment which results in a much better noise reduction filtering. This technique can be further extended to the DCT basis functions. This proposed approach improves the overall performance, Advanced Discrete Cosine Transform is integrated with pitch synchronous analysis technique. Wiener filtering is used in this approach for better performance. The autocorrelation function is used for detecting the pitch period which is in turn used as the amount of shift for the analysis window. Therefore, a consistent DCT spectrogram is generated for better noise reduction filtering. This technique can be further improved by maximum alignment which results in a much better fit to the DCT basis functions. This proposed approach is called ADCT based WFPS which produces good quality enhanced speech. Two objective measures, segmental SNR and PESQ are utilized to evaluate the proposed system.

5. Conclusion

This research work focuses on developing an efficient speech enhancement technique. DCT based speech enhancement approaches are observed to produce better results. In conventional DCT-based noise reduction algorithms, the observed speech signal is partitioned into fixed overlapping frames and transformed into DCT domain which results in variation of DCT coefficients from one frame to another due to non-ideal analysis window positions. In order to improve the overall performance, Advanced Discrete Cosine Transform is integrated with pitch synchronous analysis technique. Wiener filtering is used in this approach for better performance. The autocorrelation function is used for detecting the pitch period which is in turn used as the amount of shift for the analysis window. Therefore, a consistent DCT spectrogram is generated for better noise reduction filtering. This technique can be further improved by maximum alignment which results in a much better fit to the DCT basis functions. This proposed approach is called ADCT based WFPS which produces good quality enhanced speech. Two objective measures, segmental SNR and PESQ are utilized to evaluate the proposed system.

Table 2 shows the performance comparison of the proposed speech enhancement approach with other approaches such as WFHO, DCT based PSWF and ATSA in terms of PESQ score. It is observed that the proposed ADCT based WFPS approach provides better results. In conventional DCT-based noise reduction approaches such as WFHO, DCT based PSWF and ATSA in terms of PESQ score. It is observed that the proposed ADCT based WFPS approach provides better

TABLE II COMPARISON OF Δ PESQ RESULTS

<table>
<thead>
<tr>
<th>Noise Type</th>
<th>SNR (dB)</th>
<th>Δ PESQ (×10⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WFHO</td>
<td>DCT based PSWF</td>
</tr>
<tr>
<td>White</td>
<td>0</td>
<td>6.13</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>6.78</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>6.98</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>6.82</td>
</tr>
<tr>
<td>Fan</td>
<td>0</td>
<td>6.40</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>5.78</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>4.71</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>3.52</td>
</tr>
<tr>
<td>Car</td>
<td>0</td>
<td>4.53</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>3.40</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>2.17</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>1.14</td>
</tr>
</tbody>
</table>

Table 2 shows the performance comparison of the proposed speech enhancement approach with other approaches such as WFHO, DCT based PSWF and ATSA in terms of PESQ score. It is observed that the proposed ADCT based WFPS approach provides better results.

References


