American Journal of Applied Sciences 11 (1): 24-37, 2014
ISSN: 1546-9239

doi:10.3844/ajassp.2014.24.37 Published Online 11 (1) 2014 (http://www.thescipub.com/ajas.toc)

A NOVEL SPEECH ENHANCEMENT APPROACH BASED ON MODIFIED DCT AND IMPROVED PITCH SYNCHRONOUS ANALYSIS

[1]Balaji, V.R. and [2]S. Subramanian

[1]Department of ECE, Sri Krishna College of Engineering and Technology, Coimbatore, India
[2]Department of CSE, Coimbatore Institute of Engineering and Technology, Coimbatore, India

Received 2013-06-03, Revised 2013-07-17; Accepted 2013-11-21

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) 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, Improved Iterative Wiener Filtering approach called Constrained Iterative Wiener Filtering (CIWF) 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: Improved Iterative 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 (Paliwal *et al*., 2012). 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 (Raitio *et al*., 2011; Zen *et al*., 2012).

For the past two decades, speech enhancement has become one of the most active researches in the field of signal processing but still there are no standard techniques for both speech and noise (Anusuya and Katti, 2009).

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 much easier to filter off the noise particularly for the noise-only coefficients. Different transforms may require different analysis techniques. 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 (Paliwal *et al*., 2012) was extended to the Fourier transform by

Corresponding Author: Balaji V.R., Department of ECE, Sri Krishna College of Engineering and Technology, Coimbatore, India

(Verteletskaya and Simak, 2011; Anusuya and Katti, 2009) became a very widely used approach.

Recently, it is observed that Discrete Cosine Transform can be effectively used in speech enhancement as it is also a Fourier-related transform which only uses real numbers rather than the complex numbers used in DFT. A number of benefits of DCT have been presented by D'Ambrosio (2011) in enhancing speech as compared to DFT. These features of DCT can be summarized as follows:

- DCT is capable of giving higher energy compaction capability
- It is a real transform without phase information
- It offers a higher resolution for examining the transform coefficients in the same window length

The higher energy compaction ability plays a key role in speech enhancement as if the speech energy can be compressed into lesser coefficients even as the background noise remains white, then the noise can be easily eliminated and there is less chance to distort the speech signal during the noise reduction process.

This study 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 coefficient's 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 for the assessment of α priori SNR.

ADCT is widely used in audio processing, where the overlapping minimizes artifacts from the block boundaries (Ilk and Guler, 2011). Hence, this research work uses ADCT in order to improve the speech quality. So,pitch synchronous analysis is an efficient key which also helps in offering better performance (Morales-Cordovilla *et al*., 2011). 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. An Improved Iterative Wiener Filtering in DCT domain will be introduced which in turn uses the windowing function.

Thus, an advanced speech enhancement system namely Advanced Discrete Cosine Transform Speech Enhancement with Iterative Wiener Filtering based Pitch Synchronous Analysis (ADCT based IWFPS) is proposed in this approach.

2. LITERATURE SURVEY

A number of DFT based techniques concentrates to filter the spectral magnitude only 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 (Ding *et al*., 2011). 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). 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 Szeliski (2010).

Pitch synchronous analysis has been earlier used in various speech signal processing systems such as speech analysis/synthesis system (Morales-Cordovilla *et al*., 2011), prosody modification system (Govind and Prasanna, 2009) and speech recognition system. 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 (Jagla *et al*., 2012). PSOLA technique can also be used in other domains such as frequency domain (Bajibabu *et al*., 2011). 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 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 (Shahnaz *et al*., 2012). DCT is applied as an application tool as it 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. Pitch synchronous processing has been extensively applied in speech processing but is being rarely been used for the purpose of speech enhancement (Ding and Soon, 2009).

3. ADCT AND IWFPS PITCH SYNCHRONOUS BASED SPEECH ENHANCEMENT

The structure of this proposed speech enhancement system is shown in **Fig. 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 (Ding *et al*., 2011).

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 (Kasmani *et al*., 2009).

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 use of ADCT in this research work.

The direct and inverse ADCT are defined as Equation (1):

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

$$r = 0.....N-1$$

$$\tilde{a}\pi = \frac{2}{N}\sum_{r=2}^{N-1} a_r \cos\left[\pi \frac{k+(N+1)/2)(r+1/2)}{N}\right] \quad (1)$$

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. Assume an identical analysis-synthesis time window. Certain limitations of perfect reconstruction are Equation (2):

$$h_k = h_{2N-1-k}\ \ h_k^2 + h_{k=N^2=1} \quad (2)$$

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. The sine window is defined as Equation (3):

$$h_k = \sin[\pi(k+1/2)/2N,\ k = 0...,2N-1 \quad (3)$$

$\hat{a}_k$ in (1) are the IADCT coefficients of a_r. It contains time domain aliasing Equation (4):

$$\hat{a}_k = \begin{cases} \tilde{a}_k - \tilde{a}_{N-1-k,} & k = 0,.....,N-1 \\ \tilde{a}_k + \tilde{a}_{3N-1-k,} & k = N,.....,2N-1 \end{cases} \quad (4)$$

3.1. Windowing Function

In signal processing, if a signal is to be observed over a finite duration, then a window function has to be applied to truncate this signal. The simplest window function is the rectangular window which causes the well-known problem, spectral leakage effect. That is, if there are two sinusoids with similar frequencies, leakage interferes with one buried by the other. If their frequencies are unlike, leakage obstructs when one sinusoid has much weaker amplitude than the other. The main reason is that the rectangular window represented in the frequency domain has strong side-lobes where the first side-lobe is only around 13 dB lower than the main lobe.

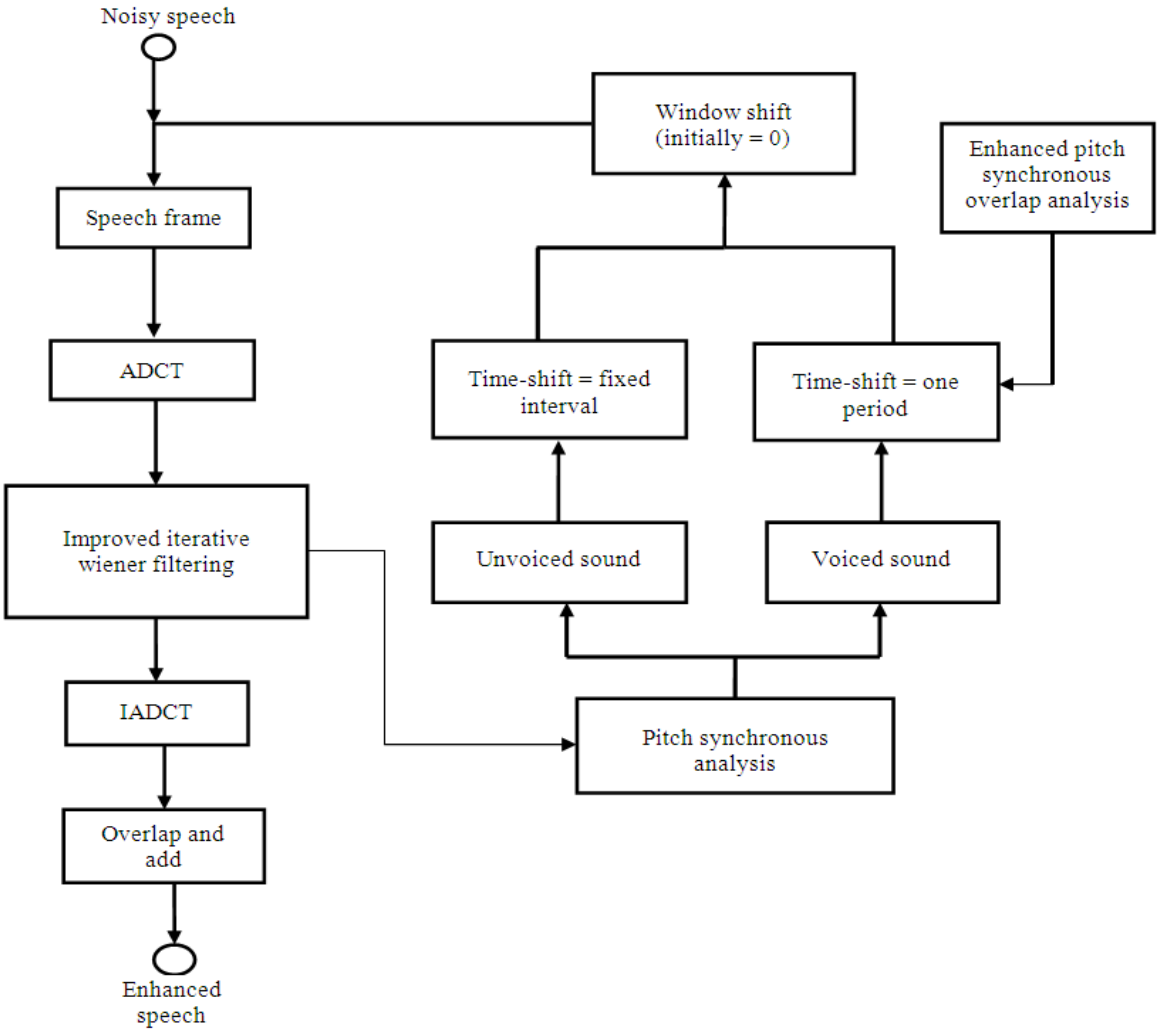

Fig. 1. Block diagram of the ADCT based IWFPS

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. Thus, some other window functions are used instead in many DCT applications. For instance, a sine window is widely used in audio coding because it offers good stop-band attenuation for high quality coding, e.g., MP3 and MPEG-2 ACC. Some other window shapes, such as Kaiser-Bessel derived window are used for Vorbis.

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 (Shekokar and Mali, 2013). In this study, rectangular window is used for better performance of the system with Advanced Discrete Cosine Transform.

3.2. Improved Iterative Wiener Filtering

Wiener filter has been proven to be the optimal filter for the real transform in Mean Square Error (MSE). During the implementation, it fully depends on the estimation of the a priori SNR. The a priori SNR can be computed by many ways among which the decision-directed approach (Paliwal *et al*., 2012) 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 $Y_{m,k}$, S_{mk} and N_{mk}.

This study 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 CIWF performance is improved using codebook constraints in the LAR domain instead of LSP domain.

Figure 2 shows the adaptive CIWF scheme. The noisy signal is x = s+d, where s is the speech signal and d is the noise signal. The speech signal s in IWF is formulated as a response of an all-pole system and the approach is utilized to solve for the MAP estimate of the signal, given x. In scenarios, where background noise psd P_d (ω) is time-varying, the conventional method is to update the noise psd estimate in nonspeech regions. This method has two major limitations: firstly, a speech/non-speech classification is required which in itself is challenging in noisy conditions; secondly, this approach is based on the postulation that adequate non-speech duration is available to update the noise estimate which may not be the case. Furthermore, the noise itself could be changing within a non-speech region. Thus, an inaccurate calculation of $P_d(\omega)$ greatly affects the performance of the wiener filters. A simple and efficient adaptive technique is presented, that tracks the dynamic noise characteristics. As the signal spectrum can be calculated iteratively and the Wiener filter is optimum in calculating the signal, the noise spectrum in each frame can be calculated through signal subtraction. This provides the means of estimating the time-varying spectrum. However, it is assumed that noise is less time-varying than speech and thus, for each frame, the noise estimate is attained by averaging the noise power spectrum of the last L frames as shown below. For each frame m Equation (5 and 6):

$$\hat{P}_d(m;\omega) = \frac{1}{L}\left(\sum_{j=m-1}^{j=m-L}(F(j;\omega).W(m))\right) \quad (5)$$

$$\begin{aligned} &F(j;\omega) = P_x(j;\omega) - \hat{P}_s(j;\omega); \text{if } \hat{P}_x(j;\omega) > \hat{P}_s(j;\omega) \\ &F(j;\omega) = P_x(j;\omega); \text{otherwise} \end{aligned} \quad (6)$$

Parameter is the frequency index, $\hat{P}_d(\omega)$ denotes the noise psd estimate, $\hat{P}_s(\omega)$ represents the speech psd estimates, $P_x(\omega)$ denotes the noisy signal psd estimates of CIWF and W(m) denotes the weighting function.

In order to initiate the consecutive evaluation, the noise psd calculation is attained from assumed initial non-speech duration of 0.2 sec. The speech psd calculation is attained with every iteration of CIWF as shown in the **Fig. 2**. The smoothing parameter L is based on the measure of non-stationarity of the noise. Ideally, the smaller the value of L, the better is the algorithm able to track rapidly varying noise. In addition, the weighting function W(m) chosen as a tapering window takes into account the higher correlation of the nearby frames rather than farther frames. Although, the algorithm makes no assumption regarding the type of noise, it is found to give robust performance for a variety of real world noises (Jingfang, 2011).

3.3. 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 $\hat{S}_1$ is calculated through the Wiener filter. The current estimate $\hat{S}_1$ is in turn used to estimate the next estimate of a. This procedure is continued until convergence is achieved. Santhi and Banu (2011), H(ω) 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 is performed to obtain the enhanced speech estimate. Following LPC analysis, the above estimate gives a_o which determines $H_o(\omega)$. It is obvious that $H_o(\omega)$ is better than starting with a unity WF and it results in better convergence properties of CIWF.

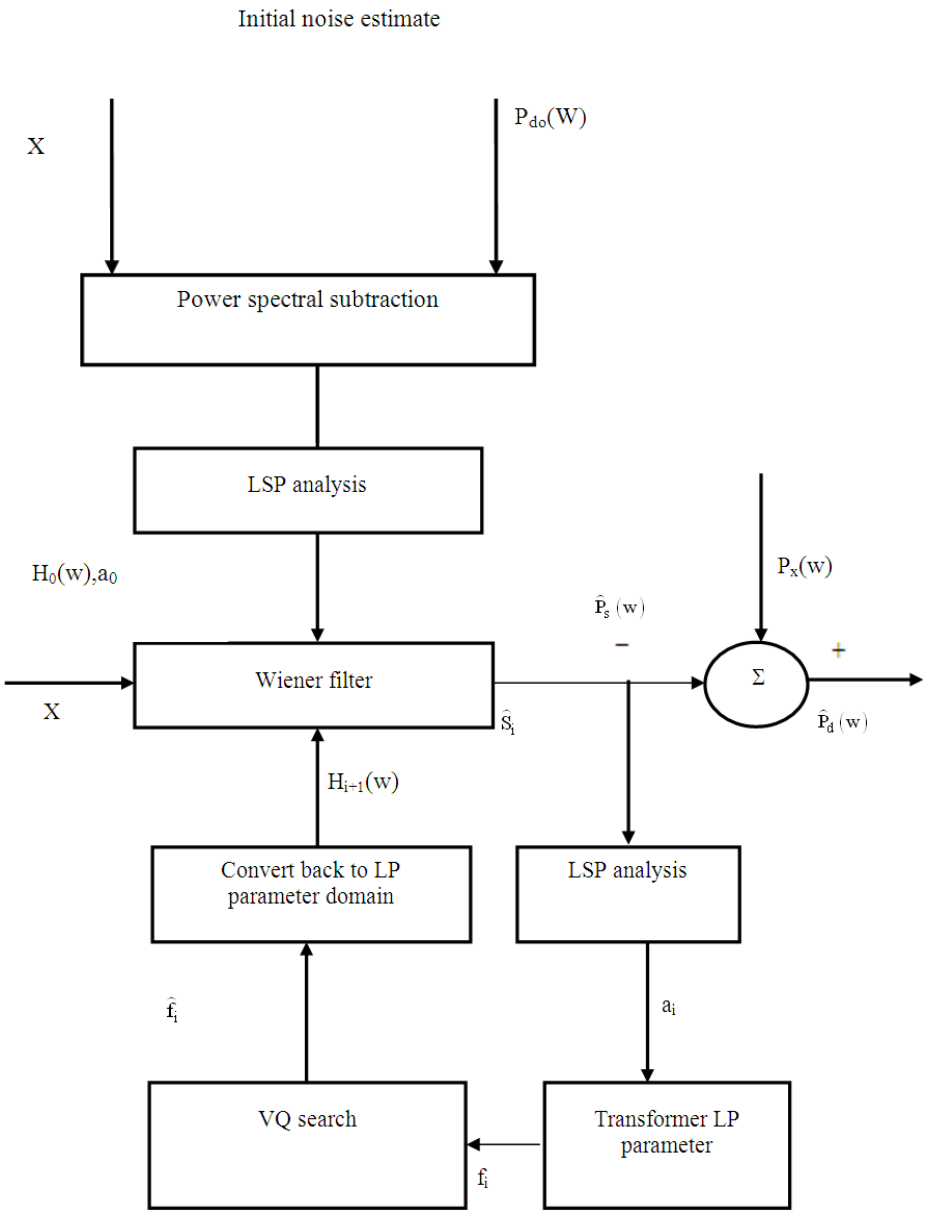

Fig. 2. Adaptive Constrained Iterative Wiener Filtering (CIWF)

3.4. Robust Parameter Domain Search

The significance of CIWF 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 Coeffecients (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 CIWF to discover the best performing parameter. The widely used IS distance measure is used for creating LPC codebooks. The Eucledean Distance (ED) is used for LAR and RC codebooks. For LSPs, ED and other two perception

based weighted Eucledean 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 weighs each LSF in the inverse proportion of its nearness to its neighbors because of the improved possibility of it denoting formants.

The estimated a priori SNR, can be expressed as follows Equation (7):

$$\hat{\zeta}_{m,k} = a\frac{\left|\hat{S}_{m-1k}\right|^2}{\lambda_N} + (1-a)\max\left[\frac{\left|Y_{m,k}\right|^2}{\lambda_N} - 1, 0\right] \quad (7)$$

where, $\hat{S}_{m-1,k}$ is the estimated clean speech in the previous frame, max is the maximum function and λ_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 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 (Ding and Soon, 2009). 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 (Shafi and Khan, 2012). Thus, it is workable to propose an adaptive for 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 Equation (8):

$$\hat{\zeta}_{m,k} = a_{m,k}\hat{\zeta}_{m-1,k} + (1-\alpha_{m,k})\max(\gamma_{m,k} - 1, 0) \quad (8)$$

where, $a_{m,k}$ is an adaptive version of a, $\xi_{m-1,k} = |\hat{S}_{m-1,k}|^2/\lambda N$ and $\lambda_{m,k} = |Y_{m,k}|^2/\lambda_N$. Then the error between the estimated a priori SNR $\xi_{m,k}$ and the real one $\xi_{m,k}$ is Equation (9):

$$J_a = E\left\{\left(\hat{\zeta}_{m,k} - \hat{\zeta}_{m,k}\right)^z\right\} \quad (9)$$

If $E\{(\gamma_{m,k}-1)^z\}$ is set to $\xi_{m,k}$ which is reasonable, then (9) can be rewritten as Equation (10):

$$J_a = \alpha^2_{m,k}\left(\hat{\xi}_{m-1,k} - \hat{\zeta}_{m,k}\right)^z - \left(1-\alpha_{m,k}\right)^z \hat{\zeta}^2_{m,k} + \left(1-\alpha_{m,k}\right)^z E\left\{\left(\gamma_{m,k}-1\right)^z\right\} \quad (10)$$

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, $E\{(\gamma_{m,k}-1)^z\}$ can be expressed as Equation (11):

$$E\left\{\left(S^4_{m,k}\right)\left|\left(\lambda^2_{N,N}\right)\right.\right\} = 3\xi^2_{m,k} \quad (11)$$

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 Equation (12):

$$J_a = a^2_{m,k}\left(\xi_{m-1,k} - \xi_{m,k}\right)^z + \left(1-\alpha_{m,k}\right)^z\left(2\xi^2_{m,k} + 4\xi_{m,k}\right) \quad (12)$$

Equating $\partial J_a/\partial a_{m,k}$ to zero, the optimal expression of can be obtained as Equation (13):

$$\alpha_{m,k} = \frac{2\xi^2_{m,k} + 4\xi_{m,k}}{\left(\xi_{m-1,k} - \xi_{m,k}\right)^2 + 2\xi^2_{m,k} + 4\xi_{m,k}} \cong \frac{1}{1 + \frac{1}{2}\left(\frac{\xi_{m-1,k} - \xi_{m,k}}{1+\xi_{m,k}}\right)^2} \quad (13)$$

The approximation used above is to avoid division by zero. As $\xi_{m,k}$ is unknown, (7) cannot be applied directly. An approximate value of $a_{m,k}$ can be obtained by substituting $\xi_{m,k}$ with $\xi(m,k)$ which is defined as follows Equation (14):

$$\zeta_{m,\pi} = \left(\gamma_{m,k} - 1\right) * H(m) \quad (14)$$

where, * 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 $a_{m,k}$ Equation (15):

$$\hat{a}_{m,k} = \beta\hat{a}_{m-1,k} + (1-\beta)\alpha_{m,k} \quad (15)$$

β represents a parameter which is fixed to 0.5 for the experimental evaluations. From the above equation SNR changes slowly, the parameter $a_{m,k}$ 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.

3.5. Pitch Synchronization

In order to implement the ADCT based IWFPS 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.

For a discrete signal x (n), the autocorrelation function is Equation (16):

$$R(m) = \lim_{N \to \alpha} \frac{1}{2N+1} \sum_{n=-N}^{N} x(n).\ x(n+1), 0 \le m \le M_0 \quad (16)$$

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 long time autocorrelation is calculated from (16). Generally with short speech segments, consisting of finite number of samples the autocorrelation based PDAs short-time autocorrelation function is as below Equation (17):

$$R(m) = \frac{1}{N} \sum_{n=0}^{N-1-m} x(n).x(n+m), 0 \le m \le M_0 \quad (17)$$

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.

3.6. A New 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 Equation (18):

$$F_{0i} = F_0^{min} + i\frac{F_0^{max} - F_0^{min}}{M-1} i = 0, M\text{-}1 \quad (18)$$

where, M is the total number of sound samples to extract:

- Calculate the growth of the fundamental frequency of the input signal. This step is performed by calculating, approximately the evolution of the most energetic harmonic $k_{Amax} = F_o$
- 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 $k_{Amax} = F_o$ approximation. Preferably, the resulting signal is a single sinusoid modulated in frequency and amplitude according to the $k_{Amax} = F_o$ evolution and remains in phase with the input signal
- Pitch marks are placed so for recalling $k_{Amax} = N_c$ these are placed in the input signal at the intial level for every N_c^{th} period of the filtered signal obtained in step 2. For each frequency F_{oi}, a single pitch mark is chosen as the one equivalent to the closest fundamental

- For each selected pitch mark, a sound sample is extracted from the input signal with an suitable temporal window. The additive noise ωF_o is naturally extracted with the harmonic part of the signal and requires no additional operations (Jagla *et al*., 2012) Equation (19):

$$x_{F_0}[n] = \sum_k A_{F0,k} \sin\left(2\pi \frac{kF_0}{F_s} n + \phi_{Fo,k}\right) + \omega F_0[n] \quad (19)$$

where, n is the discrete time index:

- $F_0 = F_c$ is the fundamental frequency
- $A_{Fo,k}$ and $\phi_{Fo,k}$ are the amplitudes and the initial phases of the harmonics
- $\omega F_o[n]$ 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 313.998s including the silence period. Approximately 50% of the speech segments are classified as voiced speech.

The proposed ADCT based IWFPS 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 (Kressner *et al*., 2013) and is easy to implement and it has been widely used to qualify the enhanced speech. The implementation in (Valentini-Botinhao *et al*., 2011) 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 (Ding and Soon, 2009) Equation (20):

$$\text{SegSNR} = \frac{10}{|\gamma|} \sum_{l \in Y} \log \frac{\sum_{k=0}^{N/2} |X(k,l)|^2}{\sum_{k=0}^{N/2} |D(k,l)|^2} \quad (20)$$

where, γ represents the set of frames that contain speech and $|\gamma|$ its cardinality.

PESQ which is described in ITU-T recommendation P.862 and is also published in (Rix *et al*., 2011) 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. Valentini-Botinhao *et al*. (2011) 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 8ms 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 **Fig. 3 and 4**, 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 into consideration, rectangular window is observed to provide better Segmental SNR.

To exhibit the advantages of each component of the proposed ADCT based IWFPS 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 IWFPS provides efficient ΔSEGSNR for all the SNR input values taken for consideration. Similarly for the other noise types, the proposed ADCT based IWFPS approach outperforms the other approaches taken for comparison.

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

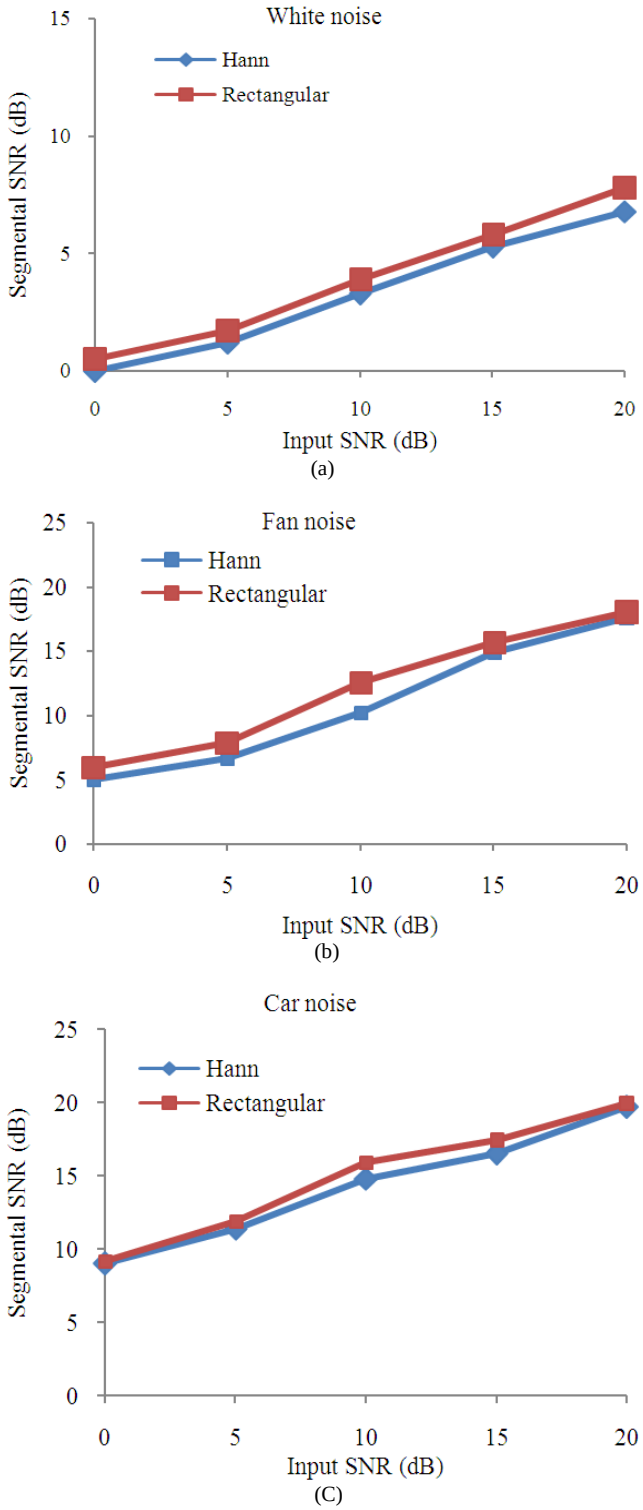

Fig. 3. Segmental SNR results of noisy speech, iterative wiener filtered speech with rectangular window and hann window

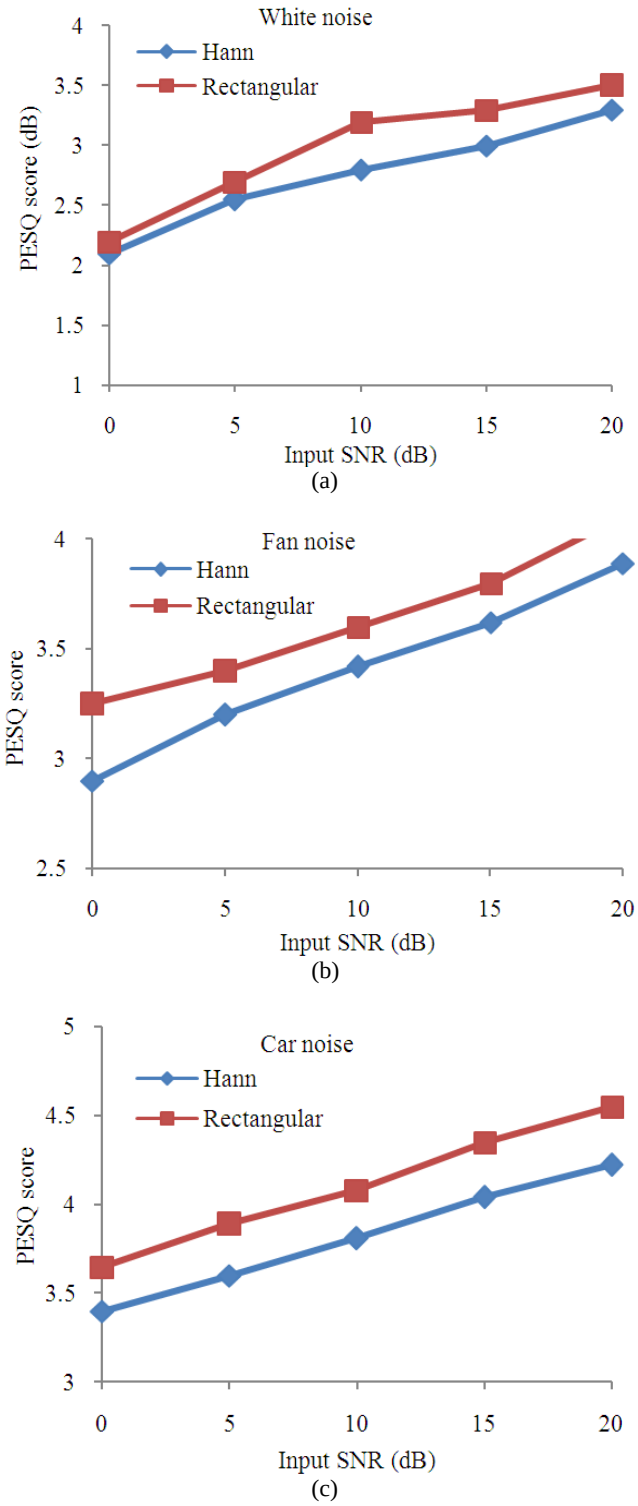

Fig. 4. PESQ score results of noisy speech, iterative wiener filtered speech with rectangular window and hann window

Table 1. Comparison of ΔSEGSNR results

Noise type	SNR (dB)	ΔSEGSNR			
		WFHO	DCT based PSWF	ATSA	ADCT based IWFPS
White	0	5.23	5.24	5.48	5.62
	5	4.52	4.53	4.86	4.99
	10	3.48	3.53	3.95	4.12
	15	2.52	2.56	3.09	3.28
Fan	0	8.84	8.97	9.26	9.51
	5	8.76	8.94	9.29	9.56
	10	8.27	8.52	8.84	9.05
	15	7.42	7.71	8.01	8.23
Car	0	12.11	12.21	12.72	12.94
	5	11.70	11.81	12.34	12.55
	10	10.85	11.03	11.48	11.69
	15	9.65	9.81	10.22	10.54

Table 2. Comparison of Δ PESQ results

Noise type	SNR (dB)	Δ PESQ ($\times 10^{-1}$)			
		WFHO	DCT based PSWF	ATSA	ADCT based IWFPS
White	0	6.13	6.17	6.27	6.39
	5	6.78	6.90	7.00	7.12
	10	6.98	7.04	7.21	7.34
	15	6.82	6.87	6.99	7.13
Fan	0	6.40	6.48	6.70	6.91
	5	5.78	5.81	5.96	6.13
	10	4.71	4.72	4.88	5.06
	15	3.52	3.58	3.73	3.96
Car	0	4.53	4.58	4.76	4.95
	5	3.40	3.47	3.61	3.87
	10	2.17	2.23	2.41	2.63
	15	1.14	1.20	1.27	1.39

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. Iterative Wiener filtering is also used in this approach for better performance. The autocorrelation function is used for detecting the pitch period which in turn,is 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. The proposed approach of ADCT based IWFPS produces good quality enhanced speech. Two objective measures, segmental SNR and PESQ are utilized to evaluate the proposed system. The future work of this research is to use different transformation approaches for evaluating the performance of the system.

6. REFERENCES

Anusuya, M.A. and S.K. Katti, 2009. Speech recognition by machine: A review. Int. J. Comput. Sci. Inform. Sec., 6: 181-205.

Bajibabu, B., R. Srikanth, S.A. Thati, B. Raj and B. Yegnanarayana *et al*., 2011. A comparison of prosody modification using instants of significant excitation and mel-cepstral vocoder. Proceedings of the Centenary Conference on Electrical Engineering, (CEE' 11), Indian Institute of Science, Bangalore, pp: 1-5.

D'Ambrosio, K., 2011. Assessing the benefits of discrete cosine transform compressive sensing for computational electromagnetics. Massachusetts Institute of Technology.

Ding, H. and I.Y. Soon, 2009. An adaptive time-shift analysis for DCT based speech enhancement. Proceedings of the 7th International Conference on Information, Communications and Signal Processing, Dec. 8-10, IEEE Xplore Press, Macau, pp. 1-4. DOI: 10.1109/ICICS.2009.5397500

Ding, H., I.Y. Soon and C.K. Yeo, 2011. A DCT-based speech enhancement system with pitch synchronous analysis. IEEE Trans. Audio, Speech Lang. Process., 19: 2614-2623. DOI: 10.1109/TASL.2011.2156785

Govind, D. and S.R.M. Prasanna, 2009. Expressive speech synthesis using prosodic modification and dynamic time warping. Proceedings of the NCC, Jan. 16-18, IIT Guwahati, pp: 290-293.

Ilk, H.G. and S. Guler, 2011. Signal transformation and interpolation based on modified DCT synthesis. Digital Signal Process., 21: 756-763. DOI: 10.1016/j.dsp.2011.01.008

Jagla, J., J. Maillard and N. Martin, 2012. Sample-based engine noise synthesis using an enhanced pitch-synchronous overlap-and-add method. J. Acoustical Soc. Am., 132: 3098-3108. DOI: 10.1121/1.4754663, PMID: 23145595

Jingfang, W., 2011. Noisy speech in real time iterative wiener filter. Proceedings of the International Conference on Digital Object Identifier Mechatronic Science, Electric Engineering and Computer, IEEE Xplore Press, Jilinm, pp: 2102-2105. DOI: 10.1109/MEC.2011.6025906

Kasmani, S.A., M. Mahfouzi and M. Asfia, 2009. A new pre-processing approach to improve DCT-based watermarkings extraction. Proceedings of the International Association of Computer Science and Information Technology-Spring Conference, Apr. 17-20, IEEE Xplore Press, Singapore, pp: 131-135. DOI: 10.1109/IACSIT-SC.2009.79

Kressner, A.A., D.V. Anderson and C.J. Rozell, 2013. Evaluating the generalization of the Hearing Aid Speech Quality Index (HASQI). IEEE Trans. Audio Speech Lang. Process., 21: 407-415. DOI: 10.1109/TASL.2012.2217132

Morales-Cordovilla, J.A., A.M. Peinado, V. Sanchez and J.A. Gonzalez, 2011. Feature extraction based on pitch-synchronous averaging for robust speech recognition. IEEE Trans. Audio Speech Lang. Process., 19: 640-651. DOI: 10.1109/TASL.2010.2053846

Paliwal, K., B. Schwerin and K. Wojcicki, 2012. Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator. Speech Commun., 54: 282-305. DOI: 10.1016/j.specom.2011.09.003

Raitio, T., A. Suni, J. Yamagishi, H. Pulakka and J. Nurminen *et al*., 2011. HMM-based speech synthesis utilizing glottal inverse filtering. IEEE Trans. Audio Speech Lang. Process., 19: 153-165. DOI: 10.1109/TASL.2010.2045239

Rix, A.W., J.G. Beerends, M.P. Hollier and A.P. Hekstra, 2001. Perceptual evaluation of speech quality (PESQ)-A new method for speech quality assessment of telephone networks and codecs. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, May 7-11, IEEE Xplore Press, Salt Lake City, UT., pp: 749-752. DOI: 10.1109/ICASSP.2001.941023

Santhi, M. and R.S.D.W. Banu, 2011. Enhancing the color Set Partitioning in Hierarchical Tree (SPIHT) algorithm using correlation theory. J. Comput. Sci., 7: 1204-1211. DOI: 10.3844/jcssp.2011.1204.1211

Shafi, M.S. and M. Khan, 2012. Transform based speech enhancement using DCT based MMSE filter and its comparison with DFT filter. J. Space Technol., 1: 47-52.

Shahnaz, C., W.P. Zhu and M.O. Ahmad, 2012. Pitch estimation based on a harmonic sinusoidal autocorrelation model and a time-domain matching scheme. IEEE Trans. Audio Speech Lang. Process., 20: 322-335. DOI: 10.1109/TASL.2011.2161579

Shekokar, S.C. and M.B. Mali, 2013. A brief survey of a DCT-based speech enhancement system. Int. J. Scientif. Eng. Res., 4: 1-3.

Szeliski, R., 2010. Computer Vision: Algorithms and Applications. 1st Edn., Springer, London, ISBN-10: 1848829353, pp: 832.

Valentini-Botinhao, C., J. Yamagishi and S. King, 2011. Evaluation of objective measures for intelligibility prediction of hmm-based synthetic speech in noise. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 22-27, IEEE Xplore Press, Prague, pp: 5112-5115. DOI: 10.1109/ICASSP.2011.5947507

Verteletskaya, E. and B. Simak, 2011. Noise reduction based on modified spectral subtraction method. IAENG Int. J. Comput. Sci.

Zen, H., N. Braunschweiler, S. Buchholz, M.J.F. Gales and K. Knill *et al*., 2012. Statistical parametric speech synthesis based on speaker and language factorization. IEEE Trans. Audio Speech Lang. Process., 20: 1713-1724. DOI: 10.1109/TASL.2012.2187195