Research Journal of Applied Sciences, Engineering and Technology 8(3): 321-330, 2014
DOI:10.19026/rjaset.8.976
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: January 20, 2014 Accepted: January 29, 2014 Published: July 15, 2014

Research Article

A Hybrid Transformation and Filtering Approach for Speech Enhancement with Time Domain Pitch Synchronous Overlap-add

[1]V.R. Balaji and [2]S. Subramanian
[1]Department of ECE, Sri Krishna College of Engineering and Technology, Coimbatore, India
[2]Coimbatore Institute of Engineering and Technology, Coimbatore, India

Abstract: The main goal of Speech enhancement is to enhance the performance of speech communication systems in noisy environments. The problem of enhancing speech which is corrupted by noise is very large, although a lot of techniques have been introduced by the researchers over the past years. This problem is more severe when there is no additional information on the nature of noise degradation is available in which case the enhancement technique must utilize only the specific properties of the speech and noise signals. 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 study, instead of DCT, a hybrid technique called DCTSLT which is the combination of Discrete Cosine Transform (DCT) and Slantlet Transform (SLT) is proposed for continuous 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, the proposed transform is combined with Time Domain Pitch Synchronous Overlap-Add (TD-PSOLA) method. Moreover, in order to improve the performance of noise reduction of the system, a Hybrid Vector Wiener Filter approach (HVWF) is used in this study. Experimental result shows that the proposed system performs well in enhancing the speech as compared with other techniques.

Keywords: Discrete cosine transform, slantlet transform, speech enhancement, time domain pitch synchronous overlap-ad method

INTRODUCTION

In several speech communication systems, recognition of speech signal from a degraded speech signal with back-ground noise is a tedious task chiefly at low SNR values. Speech quality and simplicity may significantly goes down due to the availability of background noise, particularly when the speech signal is used for subsequent processing, such as automatic speech recognition and speech coding. Owing to usage of automatic speech processing systems is increase in real world applications the speech enhancement has become a vital area of research (Rao *et al*., 2011).

In speech signal processing, speech enhancement is one of the most essential process for past few decades. Techniques like the spectral subtraction approach, adaptive noise canceling, the signal subspace approach and the iterative Wiener filter are presented by various authors in Boll (1979) and Berouti *et al*. (1979).

Spectral subtraction is the most basic method for enhancing speech corrupted by preservative noise (Boll, 1979). This technique calculates the spectrum of the dirt free noise-signal by the subtraction of the estimated noise magnitude spectrum from the noisy signal magnitude spectrum whereas keeping the phase spectrum of the noisy signal. The disadvantage of this technique is that, it contains residual noise. Signal subspace technique is proposed in Ephraim and Van Trees (1993). It is used for enhancing a speech signal corrupted by non correlated additive noise or colored noise (Ephraim and Van Trees, 1995).

For speech enhancement, the author proposed several techniques and these techniques are typically operated by calculating the infected noise initially and then removing it from the noisy speech to leave an enhanced speech signal (Scalart and Filho, 1996). These generally uses either a voice action detector to find speech inactive periods and update noise model parameters, or minimum statistics techniques where the noise model takes on minimum power levels found in the input audio signal (Martin, 2001). Yehia *et al*. (1998) the author presented a visual speech features, extracted from a speaker's mouth, to afford the clean audio speech statistics required in Wiener filtering. This method depends on correlation existing between the visual features and the audio signal. This is sustained by the generation process of speech, which is associated to

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

movements of articulators and provides correlation between the resultant audio and the shape of the mouth. Obviously, a spectrally comprehensive audio signal cannot be calculated from the shape of mouth but a spectral envelope calculation can be obtained.

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 (Ephraim and Malah, 1984) was extended to the Fourier transform and became a very widely used approach.

Thus, a hybrid speech enhancement system namely Discrete Cosine Transform (DCT) and Slantlet Transform (SLT) (Jung *et al*., 2006) for Speech Enhancement with Hybrid Vector Wiener Filter approach (HVWF) (Faizal *et al*., 2012) based Time Domain Pitch Synchronous Overlap-Ad (TD-PSOLA) (Michaeli and Eldar, year) Analysis is proposed in this approach.

LITERATURE REVIEW

The author in Nar *et al*. (2013) introduced a speech enhancement technique which can be used in various noise environments. A Wavelet Packet Transform (WPT) and a Best Fitting Regression Line (BFRL) are the techniques proposed by the author to calculate the parameters accurately for the spectral subtraction method based on the time-varying gain function. Note that this method does not make use of statistical information of pause region identified by voice activity detector. The evaluation is done on various environments where the noisy speech are between SNR -5~15 dB, in different noises.

In Abd El-Fattah *et al*. (2008) the author introduced an application of the Wiener filter is an adaptive way in speech enhancement. The adaptive Wiener filter is based on the variation of the filter transfer function from model to model based on the speech signal information that is, mean and variance. The adaptive Wiener filter is specified in time domain sooner than in frequency domain to assist for the varying nature of the speech signal. The presented method is compared to the conventional Wiener filter and spectral subtraction methods; it shows that the proposed method gives better results.

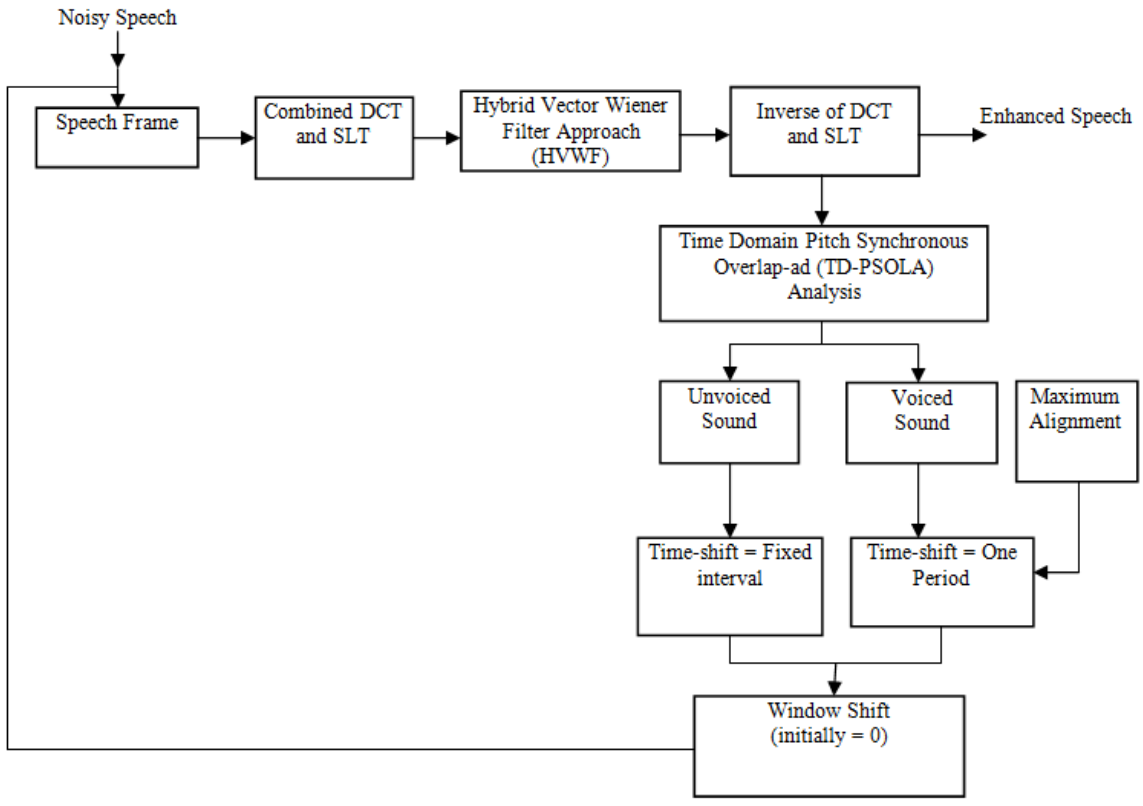

Fig. 1: Block diagram of the proposed method

In (Joon-Hyuk) Warped Discrete Cosine Transforms (WDCT) is presented by the author to improve the corrupted speech by background noise environments. For developing a useful term of the frequency characteristics of the input speech, the variable frequency warping filter is functional to the conventional Discrete Cosine Transform (DCT). The frequency warping control parameter is adjusted according to the analysis of spectral distribution in each frame. For a more accurate analysis of spectral characteristics, the split-band approach in which the global soft decision for speech presence is performed in each band separately is employed. A number of subjective and objective tests show that the WDCT-based enhancement method yields better performance than the conventional DCT-based algorithm.

METHODOLOGY

Hybrid transform and filter based Time Domain Pitch Synchronous Overlap-Ad (TD-PSOLA) analysis: The overall proposed block diagram is shown in the 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 (Balaji and Subramanian, 2014).

Window 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 (Balaji and Subramanian, 2014). 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 (Balaji and Subramanian, 2014). Similar to Fourier transform, DCT and SLT combined together 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. AC-2, etc.

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 and SLT 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. There is also a compromised adoption with trapezoidal window being applied in (Chang, 2005). In this study, rectangular window is used for better performance of the system with Hybrid DCT and SLT.

Hybrid DCT and SLT: In order to enhance the performance, hybrid technique which is the combination of Discrete Cosine Transform (DCT) and Slantlet Transform (SLT) is used in this approach. In DCT domain, representation of signal becomes a dynamic part of research in signal processing. The major benefit of this approach is resulting in aggravating blocking artifact.

DCT is a transform of base that takes real valued functions and transforms them with respect to an orthonormal cosine basis. For the real-valued speech data sequence $x[n], n = 0,1, \dots, N-1$ the DCT is defined in Shuwang *et al*. (2009) as:

$$Y[k] = a_k \sum_{n=0}^{N-1} x\,[n] \cos\left[\frac{\pi k(2n+1)}{2N}\right], k = 0,1, \dots, N-1 \quad (1)$$

The DCT (Balaji and Subramanian, 2014) is defined as:

$$x[n] = \sum_{k=0}^{N-1} a_k\, Y[k] \cos\left[\frac{\pi k(2n+1)}{2N}\right] \quad (2)$$

where, the normalization constant $a_o = \sqrt{1/N}$ for k = 0 and $a_k = \sqrt{2/N}$ for k = 1, 2,…,N-1.

The fast algorithms for calculation of DCT are based on the FFT or they are based on the direct factorization of the DCT matrix is explained in Kumar and Mutto (2009).

For an MXN speech f (x, y) its two-dimensional discrete cosine transform is defined in Shuwang *et al*. (2009):

$$C(u,v) = a(u)a(v) \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) x \cos\frac{(2x+1)u\pi}{2M} \cos\frac{(2y+1)v\pi}{2N} \quad (3)$$

where, $u = 0,1,2, \dots . M-1;\ v = 0,1,2, \dots . N-1.$
a (u) and a (v) are, respectively defined:

$$\begin{cases} a(u) = \frac{\sqrt{1}}{m} \\ \frac{\sqrt{2}}{m}, u = 0,1,2,\ldots,M-1 \end{cases}, u = 0$$

$$\begin{cases} a(v) = \frac{\sqrt{1}}{N}, v = 0 \\ \frac{\sqrt{2}}{m}, v = 0,1,2,\ldots,N-1 \end{cases}, u = 0$$

Its two-dimensional Inverse Discrete Cosine Transform (IDCT) is given as:

$$f(x) = \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} a(u)a(v)C(u,v) \times \cos \frac{(2x+1)u\pi}{2M} \cos \frac{(2y+1)v\pi}{2N} \quad (4)$$

Slantlet Transform (SLT): SLT is the same as DWT but gives better time localization due to lesser support of component filters is explained in Kumar and Mutto (2009). DWT typically put into practice in form of an iterated bank with tree structure, but SLT draws its inspiration from an equivalent form of parallel structure with parallel branches (Maitra *et al*., 2008).

The data is initially applied to two-level filter structures $H_0(z), H_1(z), H_2(z)$ and $H_3(z)$ as shown in Fig. 2.

The output is down sampled by a factor of 4 which are the transform coefficients then thresholded with appropriate parameter. The Inverse Slantlet Transform (ISLT) is performing to reconstruct the original data on these thresholded. The filter coefficient used here is SLT filter bank.

As mentioned above, DCT and SLT have shown its considerable capability in securing the speech data.

The speech is seperated into SLT sub-band like LL1, LH1, HL1 and HH1 and selects LL1 as sub-band embedding in 1D-SLT. The LL1 sub-band again separate into four sub-bands and then selected LH2 in 2D-SLT. Henceforward, the speech will be divided into 63 coefficients. By the low frequency coefficient DCT, secret bit is entered into selected coefficient. Next step is the inversion by DCT and finally inverse by SLT.

Hybrid vector wiener filtering: Wiener filter has been demonstrated to be the optimal filter for the real transform in Mean Square Error (MSE) logic. While implementing, it is entirely based on the estimation of the a priori SNR. Priori SNR can be measured by several ways among which the decision-directed approach (Kim and Su, 1991) is broadly used.

At present, the problem of recovering the high resolution description $x_{HR}[n]$ from the low resolution sequences $c_k[n], k = 1,\ldots,K$, based on the continuous-space model (1).

Statistical model: The main objective is to formulate an algorithm which produces an estimate $\hat{x}_{HR}[n]$ minimizing the MSE:

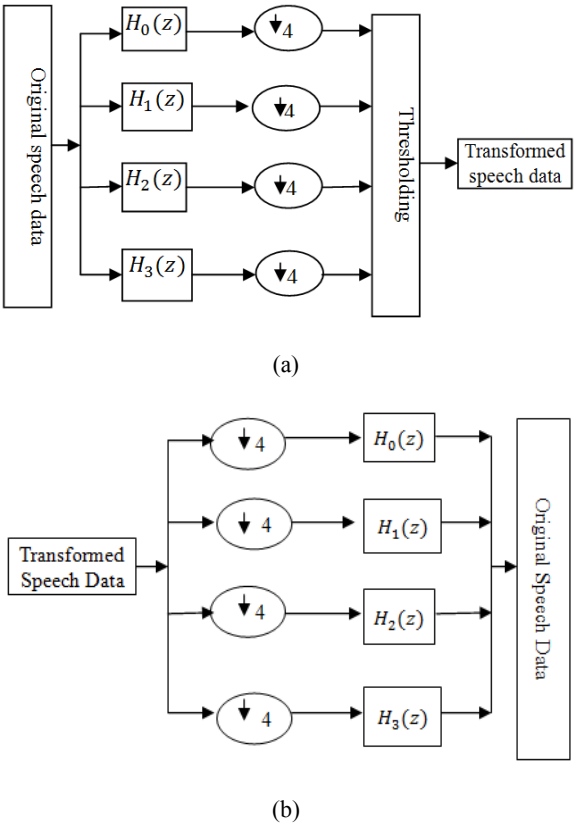

Fig. 2: (a) Two-level SLT based data transformation, (b) two-level SLT based reconstruction scheme (Chang, 2005)

$$\epsilon_x^2[n] = E[(\hat{x}_{HR}[n] - x_{HR}[n])^2 \,|x(t)] \quad (5)$$

for every pixel $n \in \mathbb{Z}^d$,
where, the expectation is over realizations of the noise sequences $u_k[n], k = 1\ldots,K$ in (5). Alas, the MSE is based on the fundamental view because $x_{HR}[n]$ is a function of $x(t)$, as in (5). Minimizing $\epsilon_x^2[n]$ uniformly over all signals x (t) is not possible. As a result, assessment between different SR approaches is not a well defined problem as one method may be better than another for some signals $x(t)$ and worse for others. To prevail over this problem, a signal $x(t)$ is proposed based on the realization of a random process with known statistics. This permits the replacement of the signal dependent MSE $\epsilon_x^2[n]$ by its expectation $\epsilon^2[n] = E[\epsilon_x^2[n]]$ over all feasible signal realizations, resulting in the condition as:

$$\epsilon^2[n] = E[(\hat{x}_{HR}[n] - x_{HR}[n])^2] \quad (6)$$

In the SR literature, different statistical priors have been proposed for describing the distinctive behavior of speech. These comprises of Gaussian and Huber random Markov fields (Hardie *et al*., 1997). Though, these investigation modeled the statistics of the

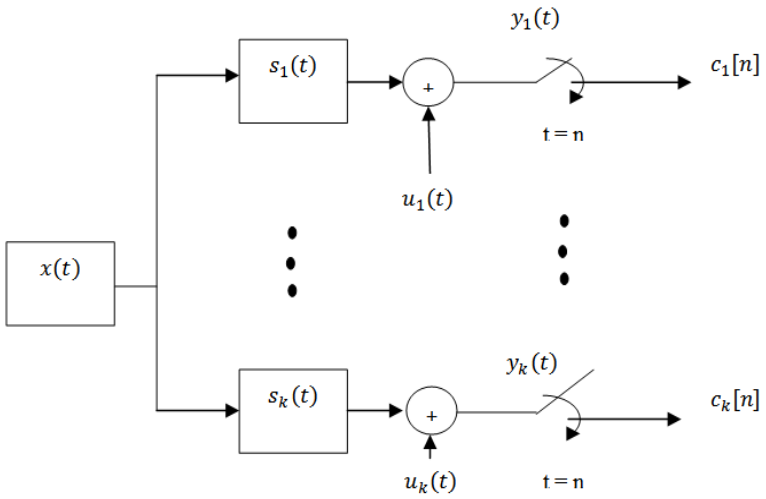

Fig. 3: Multichannel sampling scheme

preferred discrete-space signal $x_{HR}[n]$ rather than that of the continuous-space signal $x(t)$. It has been confirmed that when the Power-Spectral Density (PSD) of $x(t)$ is preferred to have a polynomial decay in frequency, the resulting recovery is better to the Shannon interpolation, which relies on the conventional band limited model. By these findings, the stationary assumption is done and expands the Bayesian single speech recovery techniques to the multi-frame SR scenario.

Optimal linear recovery: Our goal is to linearly approximate $x_{HR}[n]$ given the measurements $c_1[n],\ldots,c_K[n]$. Owing to the linearity of the convolution operation in (7), the Linear Minimum MSE (LMMSE) estimate of $x_{HR}[n]$ is given by:

$$\hat{x}_{HR}[n] = [(w * \hat{x})(t)]_{t=\frac{n}{\Delta}} \quad (7)$$

where, $\hat{x}(t)$ is the LMMSE estimate of $x(t)$ given the measurements. Hence, to achieve a closed form expression $\hat{x}_{HR}[n]$, initially compute the continuous-space LMMSE recovery $\hat{x}(t)$ and then sample it on the high-resolution grid. Before examining the problem in brief, would like to note down that the purely translational SR setting treated in this study can be viewed as a special case of multichannel sampling, as schematically shown in Fig. 3.

In this setting, a signal x(t) passes through K filters, which in case correspond to $s_k(t) = s(t - t_k)$, $k = 1,\ldots,K$, contaminated by continuous-space noise processes $u_1(t),\ldots,u_K(t)$ and then sampled on the grid $\{t = n : n \in \mathbb{Z}^d\}$ to give up the experiential measurements $c_1[n],\ldots,c_K[n]$. Therefore:

$$yk(t) = (x * s_k)(t) + uk(t), k = 1, \ldots, K \quad (8)$$

The main aim is to calculate a continuous-space signal $x(t)$ based on equidistant samples of a set of continuous-space processes $y_1(t),\ldots,yK(t)$, which are statistically associated to x (t), such that the MSE $E[(x(t) - \hat{x}(t)2]$ is reduced for every $t \in \mathbb{R}^d$.

Pitch synchronization: In order to implement the DFrCT based Adaptive Kalman Filter Combined with Perceptual Weighting Filter algorithm; the pitch period should be extracted first. There are many ways to estimate the pitch periodicity of a speech signal. Hence, initially, noise reduction filtering is carried out and the Wiener filtered speech $\hat{\xi}_{m,k}$ can be given by:

$$\hat{S}_{m,k} = \frac{\hat{\xi}_{m,k}}{\hat{\xi}_{m,k+1}} Y_{m,k} \quad (9)$$

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

$$\hat{\xi}_{m,k} = \alpha \frac{|\hat{S}_{m-1,k}|^2}{\lambda_N} + (1-\alpha) max\left[\frac{|Y_{m,k}|^2}{\lambda_N} - 1,0\right] \quad (10)$$

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 is equivalent to the expectation of the power magnitude of the noise signal, $E[|N_{m,k}|^2]$. The noise variance is unspecified to be identified as noise signal is a stationary random process and can be measured during the silence period.

Then, the enhanced speech obtained after inverse DFrCT, $\hat{s}(n)$ is utilized for pitch detection to attain more precise evaluation.

A variety of algorithms are presented in the past to detect the pitch period. The time domain autocorrelation method (Rabiner *et al*., 1976) is a

general approach for solving this problem, as it is efficient for certain noise corruption conditions. It is selected for takeout the pitch period to be used for the time-shift variation. Initially, in observed signal frame DC is detached. Then, clipping process is carried out for a better pitch period estimation. The clipping level is obtained as a fixed percentage of the minimum of the maximum absolute values. Based on this clipping level, the speech signal is processed by a three-level center clipper and the correlation function is computed over a range spanning the expected range of pitch periods. The autocorrelation function of the resulting signal $\hat{s}(n)$ can be defined as:

$$R(n) = \sum_{m=0}^{N-m-1} \hat{s}(m)\hat{s}(n+m) \quad (11)$$

Since fundamental frequency in spoken English language is range bound between 80 to 500 Hz, a distinct peak in this range implies the presence of voiced signals. A distinct peak is defined to be greater than 0.5 times of R (0). If there is no distinct peak is originated, it is possible to be a silence frame or unvoiced frame. For voiced frames, the pitch period is separated and worn as the analysis window shift. It should be observed that, the window length needs to be twice as long as the highest pitch periods of the experiential speech signal.

The last enhanced speech is resulted by overlap add process. In fact, this process is a little different from the original process due to the adaptive window shifting. A suitable solution creates a weighting function which records all the windows frame by frame and measures the net weighting function. The weighting function can be considered from the present and the prior frames and hence can be preceded in real time. Then, the enhanced speech has to be regularized by the weighting function.

Pitch synchronization with maximum alignment: The pitch synchronous analysis can be enhanced by a maximum alignment method which means that, the speech analysis window commencing the short-range of amplitude of the speech signals and the time shift is equivalent to a pitch period. As mentioned previously, the coefficients are the representation of a signal through sum of sinusoids with various frequencies and amplitudes. From the mathematical point of view, DFrCT are organized by different basis functions together with certain boundary conditions. There are four standard types of DFrCT based on their basis functions. It can be expressed as follows:

$$S(k) = \alpha(k)\sum_{n=0}^{N-1} s(n)cos\left[\frac{\pi(2n+1)k}{2N}\right] \quad (12)$$
$$k = 0,1,\dots,N-1$$

where,

$$\alpha(0) = \sqrt{\frac{1}{N}} \text{ and } \alpha(k) = \sqrt{\frac{2}{N}}$$

$$k = 0,1,\dots,N-1$$

TD-PSOLA (Time Domain Pitch Synchronous Overlap-Add): The time domain pitch synchronous overlap-add method is for disintegration of the signal to synchronized overlapping frames with pitch period. After alteration of the speech signal, reliability and correctness of the pitch marks should be conserved (Abdelkader and Adnan, 2010). For this pitch detection in speech is performed to produce pitch marks throughout overlapping windowed speech.

The Input signal $s\,[n]$ and $s_a[n]$ centered at t_a time is defined as:

$$s_a[n] = s[t_{a+n}] \quad (13)$$

where, t_a is an analysis marks.

A $s_a[n]$ is a short-time version by multiplying it by a window $w_a[n]$ is defined as $z_a[n]$:

$$z_a[n] = w_a[n] \times s_a[n] \quad (14)$$

The length of the window is twice of the local pitch period. To synthesize speech at various pitch periods, the Short Time signals (ST) are just overlapped and added with preferred spacing.

The synthesized speech is defined as follows:

$$z[n] = \sum_{a=-\infty}^{\infty} z_n[n - t_a] \quad (15)$$

where, t_a is time marks.

A better selection for the time marks ta is to match with the instantaneous of closing of the vocal folds which represents the periodicity of speech. For unvoiced speech, these marks could be randomly located. This calculation from speech waveforms is extremely a serious issue. In speech analysis, a series of pitch-marks is given after filtering the speech signal. Voiced/unvoiced choice is depends on the zero-crossing and the short time energy for each segment between two repeated pitch marks. A voicement coefficient (v/uv) can be measure in turn to quantize the periodicity of the signal (Cheveigne and Ahara, 1998). To decide on pitch marks between local extreme of the speech signal, an amount of mark candidates specified by means of all negative and positive peaks.

The Overlap-Add (OLA) synthesis is functioned based on the superposition-addition of basic signals in the new-fangled positions. These positions are firmed by the height and the length of the synthesis signal. To raise the pitch, the pitch-synchronous individual frames are extort and given to Hanning window. Then output frame stimulated close together and added up, while output frame moved further apart to decrease the pitch. Increasing the pitch will result in a shorter signal, consequently to keep constant duration duplicate frames required to be added. A fast re-sampling method is used to shift the frame specifically, where it will present in

the new signal by means of the pitch mark and the synthesis mark of a given frame.

EXPERIMENTAL RESULTS

To evaluate the proposed approach a numerous different segments of speeches are chosen arbitrarily from the TIMIT database. They are re-sampled at 8 kHz and ruined by three preservative noise types like white noise, fan noise and car noise. The 50% of the speech segments are classified as voiced speech.

The proposed hybrid technique called a combination of discrete cosine transform and Slantlet Transform based HVWF technique is evaluated using two objective measures, segmental SNR (SegSNR) measure and Perceptual Evaluation of Speech Quality (PESQ) measure. Since SegSNR is better interrelated with Mean Opinion Score (MOS) than SNR (Balaji and Subramanian, 2014) and is effortless to implement and it has been widely used to meet the criteria of the enhanced speech. The implementation 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 formula is (Ding and Soon, 2009):

$$\text{SegSNR} = \frac{10}{|Y|}\sum_{l\in Y}\log\frac{\sum_{k=0}^{N/2}|X(k,l)|^2}{\sum_{k=0}^{N/2}|D(k,l)|^2} \quad (16)$$

where, Y denotes the set of frames that contain speech and | Y | its cardinality.

PESQ is defined in ITU-T recommendation P.862 and is also studied in Muralishankar *et al*. (2004) is an objective measurement tool that forecast the outcome of subjective listening tests. It uses a sensory model to evaluate the original, unprocessed signals with the improved signals. In Rabiner *et al*. (1976) the SegSNR has an improved estimate in terms of noise reduction, as the PESQ is more accurate in terms of speech distortion prediction. It is also more reliable and highly associated with MOS as compared to further traditional objective measures. In majority situations, PESQ is the effective objective indicator for overall quality of improved speech. Before evaluating the proposed method, the effects of window functions should be presented. Iterative Wiener filter with fixed time-shift analysis of 8 ms is used. Two different window functions, rectangular window and Hann window are used to shorten the input signal.

The window length is fixed to 32 msec. SegSNR and PESQ results are shown in Fig. 4 and 5, respectively. From these two figures, it is clear that rectangular window is better for DCTSLT than the DFrCT and DCT based noise reduction algorithms. For all the noise types taken for consideration, rectangular window is observed to provide better Segmental SNR.

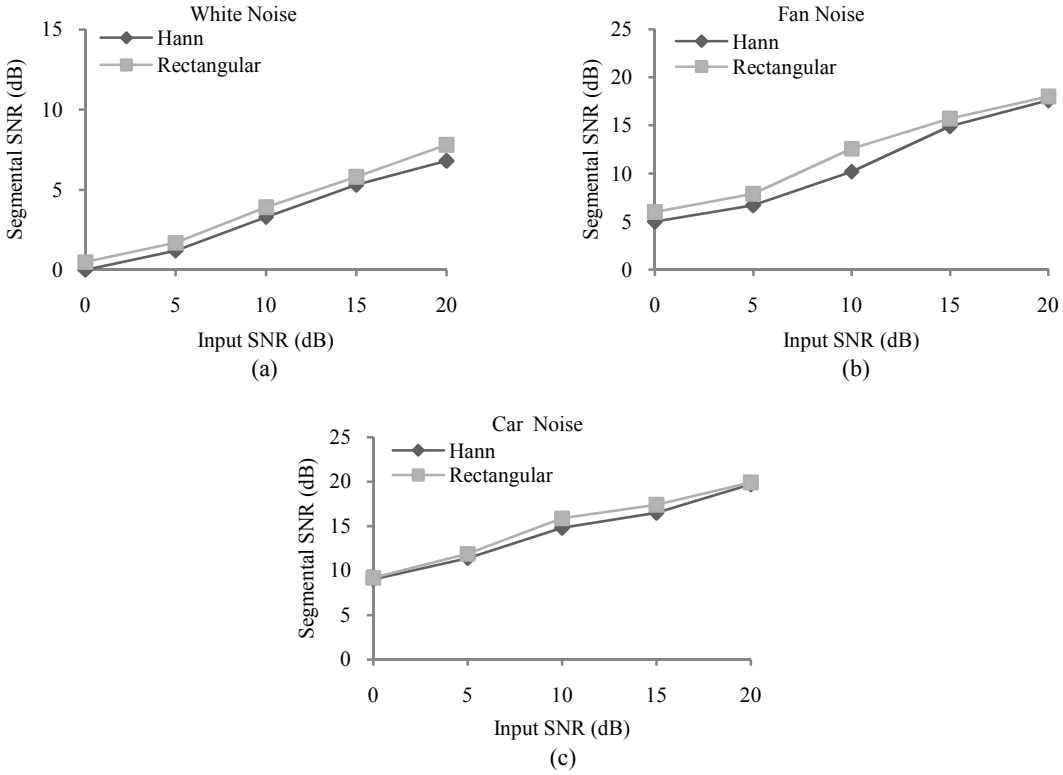

Fig. 4: Segmental SNR results of noisy speech, hybrid vector wiener filter speech with rectangular window and Hann window, (a) white noise, (b) fan noise, (c) car noise

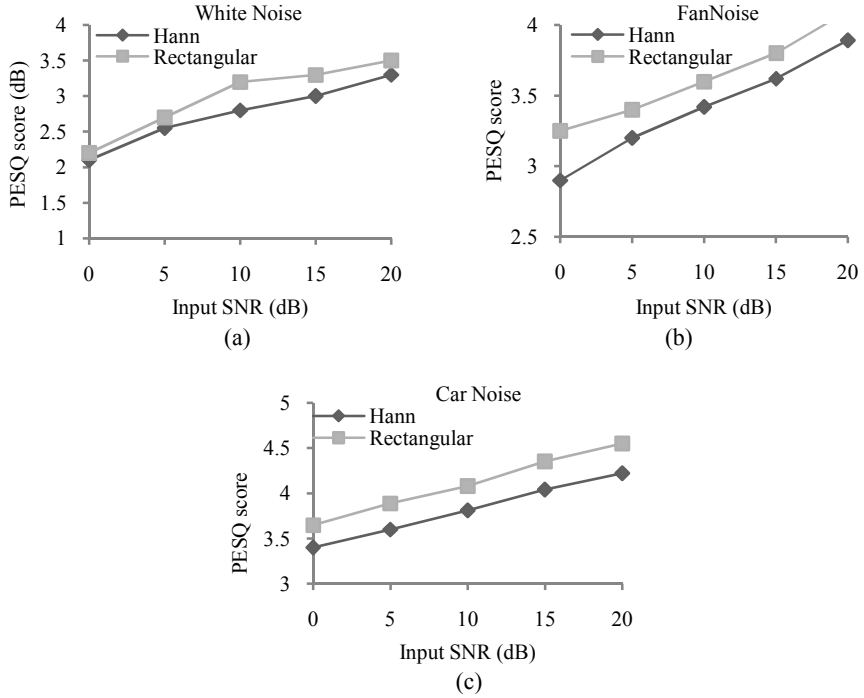

Fig. 5: PESQ score results of noisy speech, hybrid vector wiener filtered speech with rectangular window and Hann window, (a) white noise, (b) fan noise, (c) car noise

Table 1: Comparison of Δ SEGSNR results

		Δ SEGSNR					
Noise type	SNR (dB)	WFHO	DCT based PSWF	ATSA	MDCT based IWFPS	DFrCT based on adaptive kalman filter combined with perceptual weighting filter	DCTSLT based hybrid vector wiener fitter
White	0	5.23	5.24	5.48	5.62	5.89	6.18
	5	4.52	4.53	4.86	4.99	5.12	5.68
	10	3.48	3.53	3.95	4.12	4.46	4.92
	15	2.52	2.56	3.09	3.28	3.51	3.82
Fan	0	8.84	8.97	9.26	9.51	9.74	10.04
	5	8.76	8.94	9.29	9.56	9.79	10.16
	10	8.27	8.52	8.84	9.05	9.21	9.82
	15	7.42	7.71	8.01	8.23	8.52	8.86
Car	0	12.11	12.21	12.72	12.94	13.06	13.54
	5	11.70	11.81	12.34	12.55	12.76	12.96
	10	10.85	11.03	11.48	11.69	11.89	12.24
	15	9.65	9.81	10.22	10.54	10.72	10.96

Table 2: Comparison of Δ PESQ results

		Δ PESQ ($\times 10^{-1}$)					
Noise type	SNR (dB)	WFHO	DCT based PSWF	ATSA	MDCT based IWFPS	DFrCT based on adaptive kalman filter combined with perceptual weighting filter	DCTSLT based hybrid vector wiener fitter
White	0	6.13	6.17	6.27	6.39	6.61	6.85
	5	6.78	6.90	7.00	7.12	7.45	7.62
	10	6.98	7.04	7.21	7.34	7.78	7.92
	15	6.82	6.87	6.99	7.13	7.46	7.64
Fan	0	6.40	6.48	6.70	6.91	7.28	7.53
	5	5.78	5.81	5.96	6.13	6.58	6.81
	10	4.71	4.72	4.88	5.06	5.29	5.58
	15	3.52	3.58	3.73	3.96	4.26	4.54
Car	0	4.53	4.58	4.76	4.95	5.29	5.62
	5	3.40	3.47	3.61	3.87	4.06	4.28
	10	2.17	2.23	2.41	2.63	2.98	3.14
	15	1.14	1.20	1.27	1.39	1.64	1.86

To exhibit the advantages of each component of the proposed DCTSLT based Hybrid vector wiener fitter 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. The fourth approach is Hybrid Vector Weiner Filter.

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, respectively. For white noise, the proposed DCTSLT based Hybrid vector wiener filter provides efficient Δ SEGSNR for all the SNR input values taken for consideration. Similarly for the other noise types, the proposed DCTSLT based Hybrid vector wiener fitter approach outperforms the other approaches taken for comparison.

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 DCTSLT based Hybrid vector wiener Filter approach provides better Δ PESQ results when compared with MDCT, WFHO, DCT based PSWF and ATSA, DFrCT based on Adaptive Kalman Filter Combined with Perceptual Weighting Filter.

CONCLUSION

This research study mainly focuses on developing a well-organized speech enhancement technique. In this study a new transform called DCTSLT is introduced which may outcome in better performance. The autocorrelation function is worn for finding the pitch period which is in turn used as the amount of shift for the analysis window. For further improvement of the system a hybrid vector wiener filter is used which is based on covered distinctiveness of human hearing system which produces good quality enhanced speech when compared with other existing methods. Two objective measures, segmental SNR and PESQ are used in this system to evaluate the proposed system.

REFERENCES

Abd El-Fattah, M.A., M.I. Dessouky, S.M. Diab and F.E. Abd El-Samie, 2008. Adaptive wiener filtering approach for speech enhancement. Prog. Electromagn. Res., 4: 167-184.

Abdelkader, C. and C. Adnan, 2010. Implementation of the Arabic speech synthesis with TD-PSOLA modifier. Int. J. Signal Syst. Control Eng. Appl., 3(4): 77-80.

Balaji, V.R. and S. Subramanian, 2014. A novel speech enhancement approach based on modified DCT and improved pitch synchronous analysis. Am. J. Appl. Sci., 11(1): 24-37.

Berouti, M., R. Schwartz and J. Makhoul, 1979. Enhancement of speech corrupted by acoustic noise. Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '79), pp: 208-211.

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE T. Acoust. Speech, 27(2): 113-120.

Chang, J.H., 2005. Warped discrete cosine transform-based noisy speech enhancement. IEEE T. Circuits-II, 52(9): 535-539.

Cheveigne, A. and H. Ahara, 1998. A comparative evaluation of Fo estimation algorithm. Proceedings of the Euro Speech Conference (ESC'98). Norvege, pp: 453-467.

Ding, H. and I.Y. Soon, 2009. An adaptive time-shift analysis for DCT based speech enhancement. Proceeding of 7th International Conference on Information, Communications and Signal Processing (ICICSP, 2009), pp: 1-4.

Ephraim, Y. and D. Malah, 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE T. Acoust. Speech, 32(6): 1109-1121.

Ephraim, Y. and H.L. Van Trees, 1993. A signal subspace approach for speech enhancement. IEEE T. Speech Audi. P., 3(4): 251-266 .

Ephraim, Y. and H.L. Van Trees, 1995. A spectrally based signal subspace approaches for speech enhancement. Proceeding of International Conference on Acoustics, Speech and Signal Processing (ICASSP, 1995), pp: 804-807.

Faizal, M.A., H.B. Rahmalan, E.H. Rachmawanto and C.A. Sari, 2012. Impact analysis for securing image data using hybrid SLT and DCT. Int. J. Future Comput. Commun., 1(3): 308-311.

Hardie, R.C., K.J. Barnard and E.E. Armstrong, 1997. Joint MAP registration and high-resolution image estimation using a sequence of undersampled images. IEEE T. Image Process., 6(12): 1621-1633.

Jung, S.I., Y.G. Kwon and S.I. Yang, 2006. Speech enhancement by wavelet packet transform with best fitting regression line in various noise environments. Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP, 2006), pp: 01.

Kim, S.P. and W.Y. Su, 1991. Recursive high-resolution reconstruction of blurred multiframe images. Proceeding of International Conference on Acoustics, Speech and Signal Processing (ICASSP, 1991), pp: 2977-2980.

Kumar, S. and S.K. Mutto, 2009. Distortion data hiding based on slantlet transform. Proceedings of the 2009 International Conference on Multimedia Information Networking and Security (MINES '09), pp: 48-52.

Maitra, M., A. Chatterjee and F. Matsuno, 2008. A novel scheme for feature extraction and classification of magnetic resonance brain images based on slantlet transform and support vector machine. Proceedings of the SICE Annual Confference, pp: 1130-1134.

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE T. Speech Audi. P., 9(5): 504-512.

Michaeli, T. and Y.C. Eldar, year. A hybrid vector wiener filter approach to translational super-resolution. IEEE T. Image Process.,

Muralishankar, R., A.G. Ramakrishnan and P. Prathibha, 2004. Modification of pitch sing DCT in the source domain. Speech Commun., 42(2): 143-154.

Nar, V.V., A.N. Cheeran and S. Banerjee, 2013. Verification of TD-PSOLA for implementing voice modification. Res. Appl., 3(3): 461-465.

Rabiner, L., M. Cheng, A. Rosenberg and C. McGonegal, 1976. A comparative performance study of several pitch detection algorithms. IEEE T. Acoust Speech, 24(5): 399-418.

Rao, V.R., R. Murthy and K.S. Rao, 2011. Speech enhancement using cross-correlation compensated multi-band wiener filter combined with harmonic regeneration. J. Signal Inform. Process., 2: 117-124.

Scalart, P and J.V. Filho, 1996. Speech enhancement based on a priori signal to noise estimation. Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP, 1996), pp: 629-632.

Shuwang, C., A. Tao and H. Litao, 2009. Discrete cosine transform image compression based on genetic algotithm. Proceeding of International Conference on Information Engineering and Computer Science (ICIECS, 2009), pp: 1-3.

Yehia, H., P. Rubin and E. Vatikiotis-Bateson, 1998. Quantitative association of vocal-tract and facial behaviour. Speech Commun., 26(1): 23-43.